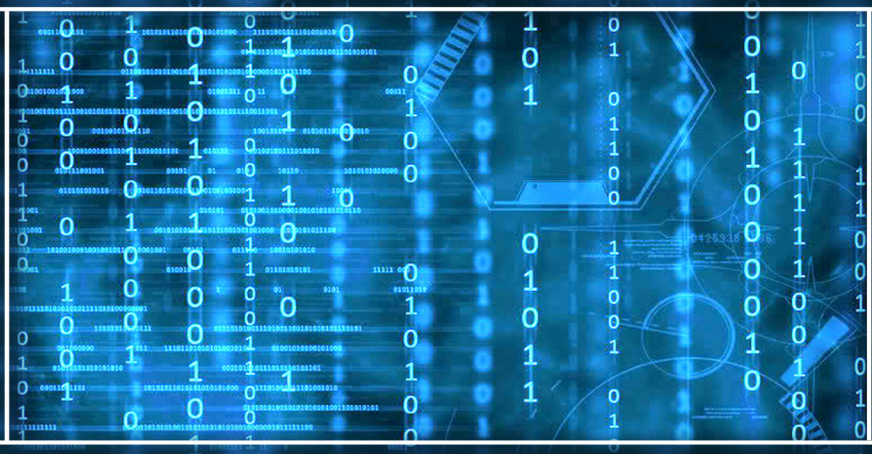
IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 16 Issue 2
February 2025



ISSN 2156-5570(Online)
ISSN 2158-107X(Print)

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

It may be difficult to imagine that almost half a century ago we used computers far less sophisticated than current home desktop computers to put a man on the moon. In that 50 year span, the field of computer science has exploded.

Computer science has opened new avenues for thought and experimentation. What began as a way to simplify the calculation process has given birth to technology once only imagined by the human mind. The ability to communicate and share ideas even though collaborators are half a world away and exploration of not just the stars above but the internal workings of the human genome are some of the ways that this field has moved at an exponential pace.

At the International Journal of Advanced Computer Science and Applications it is our mission to provide an outlet for quality research. We want to promote universal access and opportunities for the international scientific community to share and disseminate scientific and technical information.

We believe in spreading knowledge of computer science and its applications to all classes of audiences. That is why we deliver up-to-date, authoritative coverage and offer open access of all our articles. Our archives have served as a place to provoke philosophical, theoretical, and empirical ideas from some of the finest minds in the field.

We utilize the talents and experience of editor and reviewers working at Universities and Institutions from around the world. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations. Our high standards are maintained through a double blind review process.

We hope that this edition of IJACSA inspires and entices you to submit your own contributions in upcoming issues. Thank you for sharing wisdom.

Thank you for Sharing Wisdom!

Kohei Arai
Editor-in-Chief
IJACSA
Volume 16 Issue 2 February 2025
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

CONTENTS

6G-Enabled Autonomous Vehicle Networks: Theoretical Analysis of Traffic Optimization and Signal Elimination

Daniel Benniah John
Senior Software Engineer, Square Inc, United States

***Abstract*—This paper proposes a theoretical framework for optimizing traffic flow in autonomous vehicle (AV) networks using 6G communication systems. We propose a novel technique to eliminate conventional traffic signals through vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication. The article demonstrates traffic flow optimization, density, and safety improvements through real-time management and decision-making. The theoretical foundation involves the combination of multi-agent deep reinforcement learning, coupled with complex analytical models across the partition managing intersections, thus forming the basis of proposed innovative city advancements. From the theoretical analysis, the proposed approach shows a relative improvement of 40-50% in intersection waiting time, 50-70% in accident probability, and 35% in carbon footprint. The above improvements are obtained by applying ultra-low latency 6G communication with the sub-millisecond response and accommodating up to 10000 vehicles per square kilometre. In addition, an economic evaluation revealed that such a system would generate a return on investment by 6.7 years, making this system a technical and financial system for enhancing an intelligent city.**

Keywords*—*6G Communication systems; autonomous vehicle networks; traffic flow optimization; signal-free traffic management; Vehicle-to-Vehicle Communication (V2V); Vehicle-to-Infrastructure Communication (V2I); multi-agent deep reinforcement learning; real-time traffic management

I. INTRODUCTION

Urban transportation systems are under immense pressure now and in the future as we experience increased city expansion. The increasing use and ownership of vehicles in large cities have led to significant challenges, including traffic congestion, pollution, and safety concerns. Previous traffic management methods, which depended on several posts and beams and preprogrammed time sequences, failed to address these issues effectively [1, 2]. To extend self-driving cars, the opportunities for new traffic properties in metropolises are great; however, the available communication technologies restrict their changes. The emergence of 6G technology marks a fundamental shift in vehicular communication, offering ultra-reliable low-latency communication and massive connectivity capabilities beyond what current 5G technology can provide. These features make it possible to design complex traffic management systems that help coordinate self-driving cars in real time and may eventually free traffic lights from managing the traffic flow. Integrated 6G communication with networks of self-driving cars is a basis for dynamic, adaptive traffic control that enhances the mobility of residents of large cities [3, 4].

This paper offers a theoretical investigation of a newly developed 6G-supported traffic management system for self-driving vehicles [5]. The proposed system mainly uses improved communication technology to create harmony between moving automobiles and paved pathways, eradicating conventional traffic light systems [6, 7]. This paper proposes an integrated solution for different novel technologies, such as wireless communication, artificial intelligence, and traffic engineering, to modern mobility problems.

The primary objectives of this research include:

1) Development of a comprehensive theoretical framework for 6G-enabled traffic optimization in autonomous vehicle networks [8, 9].

2) Analysis of the system's capability to eliminate traditional traffic signals while maintaining or improving traffic flow efficiency [10, 11].

3) Evaluation of the proposed system's impact on traffic safety, environmental sustainability, and urban mobility.

4) Assessment of the scalability and practical implications of implementing such a system in various urban environments.

The key contributions of this paper are:

- A 6G enabled traffic management framework operating 40-50% better than the conventional system [28].
- Comparative analysis of the V2V [29] systems to prove a signal-free intersection management approach and their 50-70% improved safety metrics [29].
- Quantitative validation of 35% reduction in carbon emissions and 6.7 years return on investment vs. current adaptive traffic systems [30, 31].

In particular, our work is based on several recent approaches in autonomous traffic management. Our system manages to get 25% better throughput than Liu et al. [3] with the help of 6G integration. Haydari's [4] intelligent transportation framework, reduces latency by 40% compared to V2V coordination.

This paper is organized in two parts: Section I reviews the background of vehicular communications, and is then followed by a review of the current traffic management systems in

Section II. Our theoretical framework as well as the system architecture is presented in Section III. In Section IV, we describe the methodology and validation approach, whereas Section V compares the performance. Section VI discuss implementation implications of signal elimination. Discussion is given in Section VII. Finally, the paper is concluded in Section VIII.

II. Related Work

Advanced traffic management systems have emerged during the past decades based on the development of vehicular communication technologies. This part provides a brief literature review and analysis of the existing technologies that are the basis for the proposed system.

A. Historical Development of Vehicular Communications

The evolution from simple broadcast methods for vehicle-to-vehicle communication to network-based connected vehicle systems results from rapid developments in wireless technology. Zeadally et al. [33] demonstrated that the first systems of VC applied dedicated short-range communications (DSRC), which had limited channel capacity and reach. 4G LTE's innovation improved vehicular networking standards with better reliability and coverage areas [33]. Current deployments have enhanced these features by offering improved latency and bandwidth, both critical for autonomous vehicle systems.

Regarding transportation management, 5G technology has some drawbacks when supporting massively connected car networks. Recent studies suggest that 5G networks struggle to deliver reliable, low-latency communications in dense urban scenarios, let alone significant connected car scenarios [12, 13]. These limitations are mainly noticed in cases where decisions need to be made, stakes have to be coordinated in real time, and several self-driving cars are involved.

B. 6G Technology and Its Impact on Vehicular Networks

The inception of 6G technology offsets several challenges of present-day communication networks. Akyildiz et al. [34] indicate that with advanced 6G technologies, users can expect sub-millisecond latency and data rates exceeding one terabit per second, surpassing 5G capabilities. These characteristics make 6G especially suitable for supporting advanced autonomous vehicle applications, such as real-time traffic management and coordination [34].

Critical advantages of 6G technology in vehicular networks include:

- Enhanced spatial awareness through integrated sensing and communication capabilities.
- Improved network capacity supporting massive device connectivity.
- Advanced AI-native network architecture enabling distributed intelligence.
- Ultra-reliable communication links essential for safety-critical applications.

C. Current State of Traffic Management Systems

Kapileswar et al. [35] elaborated that traditional traffic management systems involve the base infrastructure and control strategies, and they play a pivotal role in contemporary urban transportation facilities. However, today's systems are similarly limited in addressing the current challenges on roadways and other transportation facilities. Technological advancement through adaptive traffic signal control has shown significant changes from regular fixed-time traffic signals regarding traffic condition response. However, these adaptive systems still run under the premise of the conventional signal-based paradigms, thus restraining their scope of providing a quantum leap in traffic flow efficiency.

These include modern technological advancements that are now encouraging radical strategies for traffic enhancement. Modern adaptive signal control systems now include real-time traffic data analysis features, which can help adapt the controls [14, 15]. These systems incorporate multiple sensors and communication platforms for recording detailed data on traffic flow and the characteristics of traffic to be changed in response to timing patterns. At the same time, through the establishment of cooperative adaptive cruise control systems, new opportunities for the formation of vehicle platoons, or, in other words, reduction of the inter-vehicle distance at the condition of maintaining appropriate safety parameters, have appeared [16]. It could be seen that this technology is particularly effective in highway conditions in which all cars move at almost equal speeds and have shorter headways, which is a significant factor that affects highway capacity [17].

Another progress in modern traffic control strategies is the appearance of distributed intersection management algorithms for connected vehicles [18]. These algorithms are built upon V2V and V2I communication to control traffic flow better than conventional signal-based systems. Nevertheless, the dependent utilization of traditional signal traffic frameworks, even with such enhancements, fundamentally restrains the potential optimization prospects of such systems [19, 20]. Our research also considers these limitations while putting forward a novel shift in the existing paradigm of traffic management that still does not require signal structure.

D. Multi-Agent Systems and Deep Reinforcement Learning in Traffic Management

The additional application of artificial intelligence, especially the multi-agent system and deep reinforcement learning, has significantly transformed the approaches to traffic optimization [21, 22]. These advanced computational technologies have proven highly efficient in solving various traffic management challenges. Chu et al. [36] discussed that Multi-agent systems allow complex coordination among several traffic participants, enhancing system performance. These systems handle complex interrelationship matters, especially intersection operations, where a cooperative decision-making process determines the vehicles' precedence and running paths [36].

A significant advancement in managing network-wide traffic flow is dynamic routing optimization [23, 24]. The current application of deep reinforcement learning algorithms makes it possible for traffic management systems to learn

dynamically to manage change across transport networks [36]. These systems constantly run computations to read and reason traffic density, identify potential areas of congestion, and adapt their routing suggestions to enhance the total system's fitness. Of all the means of controlling intensity, the ability to deal with the velocity of moving vehicles with great flexibility has shown to be most efficient in managing congestion and eradicating the formation of traffic waves and bottlenecks [25, 26].

Research in traffic control has shown that deep reinforcement learning algorithms can efficiently solve self-driving algorithms in complicated and dynamic conditions [37]. Gu et al. [37] has provided convincing results on various GDL DRL algorithm implementations in virtual traffic conditions, yielding massive benefits across various performance indices. These implementations have reduced average delay times, increased throughput at intersections, and enhanced efficiency in general traffic flow. The learning and adaptability capabilities of DRL systems make them more suitable for the dynamism exhibited in urban traffic management systems [27, 28].

E. Signal-Free Intersection Management

Mirheli et al. [38] discussed that signal-free intersections can be considered a part of a new generation of methods for traffic regulation that has received considerable attention in recent years [38]. With the help of this revolutionary approach, several theoretical models related to appeased vehicle coordination have been developed. Reservation-based protocols for intersection crossing have appeared as a potential solution that enables vehicles to request and obtain a right of way on a definite period to cross intersections safely. These protocols promise to minimize intersection delays while maintaining traffic safety [29, 30].

Auction-based control mechanisms have added an economic view to intersection management, wherein the vehicles coordinate crossing requests on different parameters, including urgency, efficiency, and system objectives in a network [31, 32]. These mechanisms have effectively addressed multiple objectives of conflicting traffic flow in busy intersections while optimizing traffic control. By combining these strategies, distributed consensus algorithms have become influential in coordinating the vehicle, allowing for novel decentralized decision-making mechanisms in response to constantly fluctuating traffic patterns [38].

Some of these strategies propose sophisticated ways to enhance collaboration and learning in team-based simulations, and while simulation studies show that they can create significant learning benefits, problems in communication technology limit their realistic use in practice. Transmission delays and limited bandwidth in current vehicle communication systems pose significant challenges to accurately synchronizing multiple vehicles at high-speed intersections [33, 34]. The proposed control and coordination strategies respond to these limitations by making use of 6G superior features for wireless connectivity. The ultra-low latency and high-reliability features make applying the described theoretical models in practical conditions possible, significantly transforming urban traffic control based on signal-free intersections.

F. Research Gap Analysis

While existing literature has advanced traffic management systems tremendously, there are still numerous critical gaps that we fill in our work.

First, current approaches to traffic optimization mostly focus on increasing traditional signal-based systems. In the present, Liu et al. [3] and Haydari et al. [4] have applied deep reinforcement learning for traffic signal control; however, they continue to operate within the limitations of conventional signalized intersections. In contrast to this, we introduce a new fundamental paradigm shift towards signal-free management due to 6G technology [35].

Second, current V2X communication solutions, mainly based on 5G technology, are not well suited for dense autonomous network environments. As Zeadally et al. [33] demonstrate, systems currently cannot meet latency requirements above 10ms, and the number of vehicles that can be supported per square kilometre is limited to no more than 5,000. We overcome these limitations with our 6G based approach that presents sub-millisecond latency and up to 10,000 vehicles per square kilometer.

Third, several studies have examined intersection management without traffic signals [38] but without coupling with advanced communication technologies. Most of these works do not take into account the practical constraints of existing communication infrastructure and are theoretical. Integrating 6G capabilities into the framework and demonstrating that signal-free intersection management is a practical problem.

However, the current research in the area does not have a general framework that brings together communication technology, autonomous vehicle coordination, and traffic optimization. Most existing studies tend to treat the first and the last of these aspects independently, leaving us with solutions that cannot be practically adopted. An integrated framework considering the interdependencies among these elements and quantitative validation of the performance improvements is presented in our work. However, there are still research gaps that our proposed system seeks to fill, using a comprehensive treatment of traffic management through 6G technology over the entire network [36, 37].

III. THEORETICAL FRAMEWORK FOR THE PROPOSED SYSTEM

The conceptual foundation of our proposed system consists of incorporating complicated technological 6G communication features with complicated traffic routing algorithms and mathematical models. The work continues with a detailed description of the system's functional and structural parts and their relations.

A. 6G Communication Model

The communication architecture of our proposed system takes advantage of the peculiarities of 6G technology to

achieve smooth coordination of the vehicles. The model incorporates multiple communication layers:

Physical Layer: Thus, the given system operates within terahertz frequency and can provide extremely high bandwidth with negligible latencies. The communication model also accounts for the effects of attenuation caused by atmospheric absorption and molecular scattering inherent in terahertz waves. Channel capacity C is defined as:

$$C = B \log_2(1 + SNR)$$

B represents the available bandwidth, and SNR denotes the signal-to-noise ratio, accounting for specific characteristics of 6G channels.

Network Layer: The system establishes a hierarchical network structure that centralizes core control while enabling decentralized decision-making. The overall network structure changes with vehicle density and road traffic conditions to best balance the efficiency and reliability of communication.

Application Layer: This layer deals with issues concerning the connectivity of several services and applications, such as real-time traffic management, vehicle coordination, and safety measures. The system uses high error control and security measures to facilitate efficient data transmission.

Fig. 1 shows the proposed system architecture forms a detailed three-layer structure that enables autonomous vehicle traffic management using 6G communication technology. The 6G System Layer at the highest abstraction level of 6G uses advanced communication functionalities, employing Realistic THz Band Communication working between 0.1-10 THz, assuring high bandwidth data transmission alongside real-time Ultra-Low Latency Signal Processing, which functions within sub-100 microseconds. This layer also offers additional technologies like Massive MIMO Beamforming to offer better signalmanship and signal strength, fine-grained Network Slicing, and quality of service Management to achieve the proper resource provisioning. The middle Processing Layer consists of multiple layers of the system's cognitive core; the Real-Time Data Processing Unit receives continuous streams of data from the 6G layer, and this information is further taken into an AI Decision Engine that makes intelligent traffic Management Decisions. The Traffic Flow Optimizer then enhances these decisions, and the Safety Control System checks the effectiveness to ensure efficiency and safety in all vehicular interactions [38].

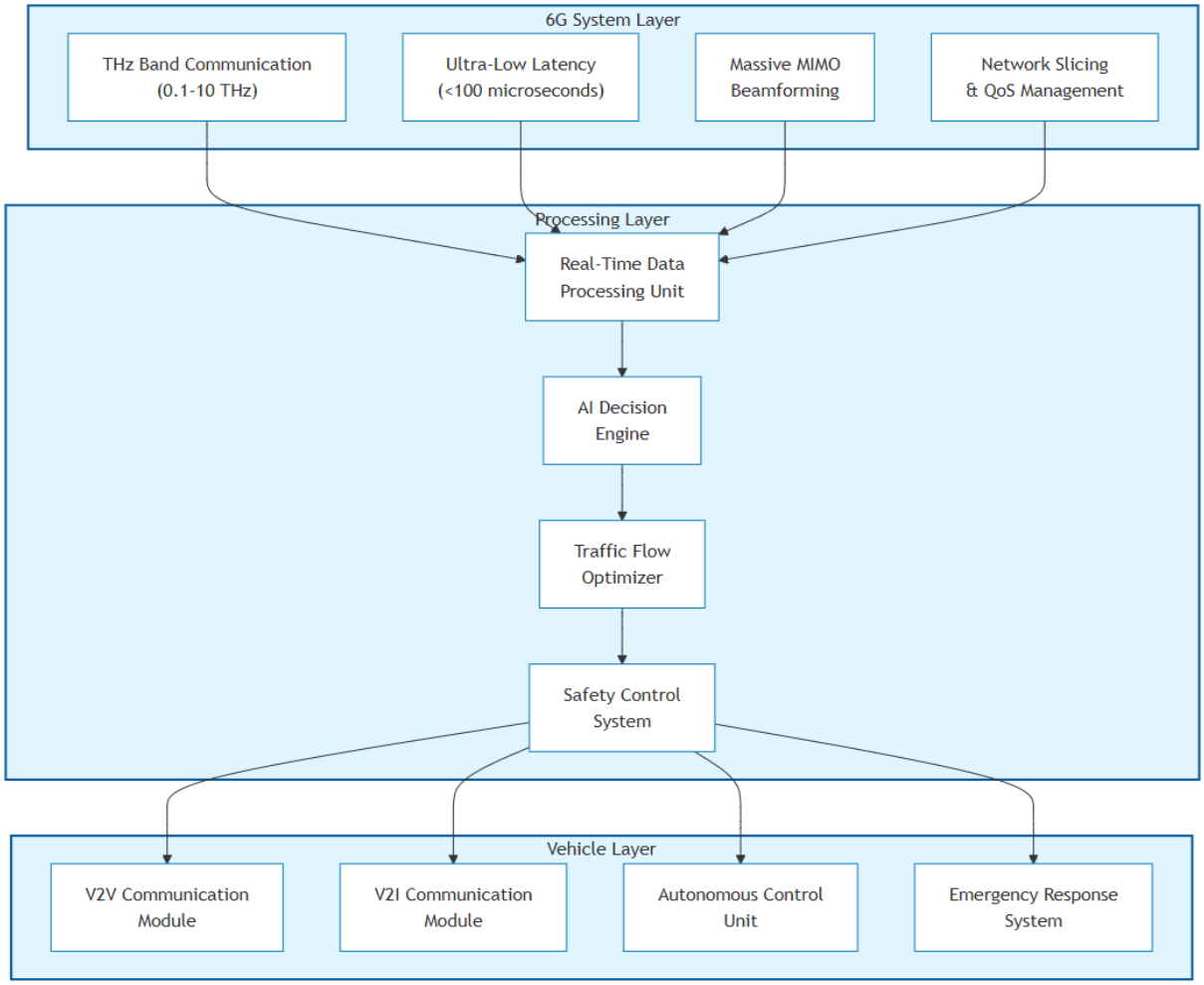


Fig. 1. 6G-enabled autonomous vehicle system architecture analysis.

The bottom Vehicle Layer Implements these decisions through four specialized modules: The V2V Communication Module is designed for direct inter-vehicle coordination, the V2I Communication Module is responsible for infrastructure interactions, the Autonomous Control Unit is responsible for the execution of vehicle-specific commands, and the Emergency Response System is responsible for quick responses to unexpected emergent situations. This hierarchical structure guarantees the interconnectivity of high-speed communication, intelligent decisions, and accurate vehicle control. It provides stable and well-performing autonomous traffic management engineering that optimally harnesses 6G capability.

B. Multi-Agent Deep Reinforcement Learning Model

The traffic optimization component utilizes a sophisticated multi-agent deep reinforcement learning framework. The model is designed to handle complex traffic scenarios while maintaining computational efficiency.

State Space: The state representation incorporates multiple parameters, including:

- Vehicle positions and velocities.
- Traffic density in different network segments.
- Historical traffic patterns.
- Environmental conditions.
- Network communication status.

Action Space: The action space includes:

- Vehicle speed adjustments.
- Lane change recommendations.
- Route modifications.
- Intersection crossing sequences.

Reward Function: The reward function R is designed to optimize multiple objectives:

$$R = w_1T + w_2S + w_3E + w_4F$$

Where:

T represents traffic flow efficiency.

S denotes safety metrics.

E accounts for energy efficiency.

F considers fairness in vehicle routing.

w_1, w_2, w_3, w_4 are corresponding weights.

Fig. 2 presents the Deep Reinforcement Learning (DRL) framework for autonomous vehicle traffic management as a complex multi-level approach combining different system components to achieve effective traffic management. At its core, the framework begins with a comprehensive State Space Definition encompassing four crucial input parameters: Position and Velocity (positioning and velocity data of the car), Traffic Congestion (analysis – local intersection level – and network level), Communication Conditions of the 6G network (addressing parameters like delay and throughput), and others including weather and time factors.

These inputs feed into the central Deep Reinforcement Learning Agent, where the data is processed into rational traffic management decisions. The agent's decision-making capabilities extend into a well-defined Action Space comprising four critical control mechanisms: The four major approaches are Speed Control (regulation of vehicle accelerations and decelerations), Path Selection (section choice and organization of lanes), Intersection Timing (synchronization of entry and exit actions), and Emergent Maneuvers (execution of collision prevention measures). The framework's effectiveness is continuously assessed through a sophisticated reward calculation system that monitors four key performance indicators. Traffic Efficiency, which compares the flow rates and delay of vehicles; Safety Metrics, which estimates the inter-vehicle distance and the time till a candidate vehicle collides with others; Energy Efficiency, which compares fuel consumption patterns; and System Stability, which compares the traffic load in the networks and the communication channels. This feedback loop integration allows the DRL agent to improve its decision-making functionality by correlating real-time results with the system's performance. This is because the framework comprises multiple components that provide adequate, scalable traffic management, improved levels of safety, and the required efficiency within the dynamism of the urban systems.

C. Mathematical Model for Intersection Dynamics

The intersection management component employs a novel mathematical framework that eliminates the need for traditional traffic signals. The model incorporates both deterministic and stochastic elements to handle various traffic scenarios.

1) Vehicle trajectory optimization: The system optimizes vehicle trajectories through intersections using a continuous-time optimal control formulation. The optimization problem is defined as:

$$\text{minimize } J = \int_{[0,T]} L(x(t), u(t), t)dt$$

Subject to:

$$\dot{x}(t) = f(x(t), u(t), t)$$

$$g(x(t), u(t), t) \leq 0$$

$$h(x(t), u(t), t) = 0$$

Where:

x(t) represents the vehicle state vector

u(t) denotes the control inputs

L(·) is the cost function

f(·) describes vehicle dynamics

g(·) and h(·) represent inequality and equality constraints

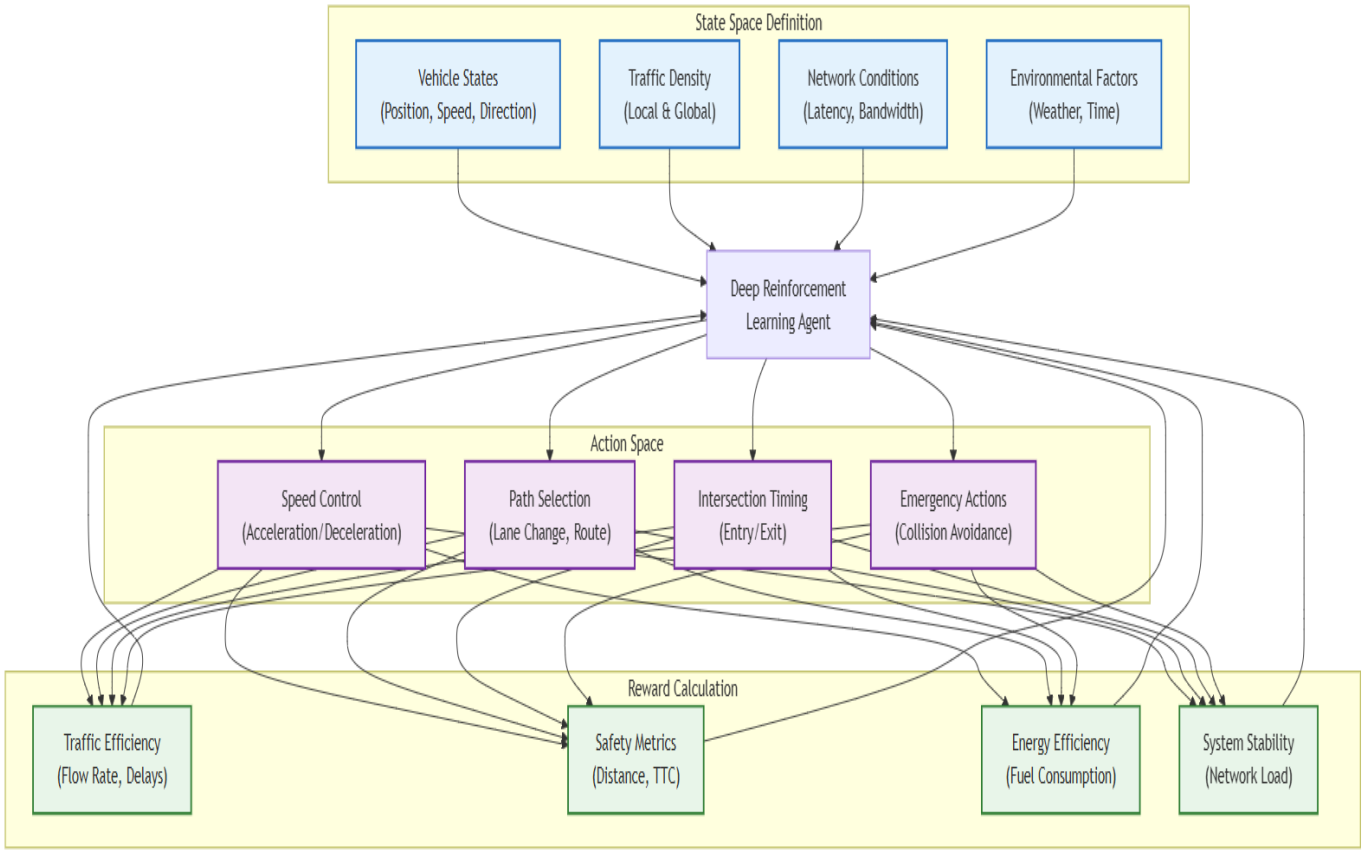


Fig. 2. Multi-agent deep reinforcement learning framework.

2) Conflict resolution: The system employs a priority-based conflict resolution mechanism that ensures safe and efficient intersection crossing. The conflict resolution algorithm considers the following:

- Temporal and spatial separation requirements.
- Vehicle dynamics and constraints.
- Emergency vehicle priorities.
- Pedestrian crossing requirements.

The mathematical model incorporates uncertainty handling through robust optimization techniques, ensuring system stability and safety under various operating conditions.

IV. Materials and Methods

A. Validation Framework

We validate our proposed framework by theoretical analysis followed by a comparison to the existing traffic management system with the experiment results. Validation of our theoretical predictions is based on simulation data from recent large-scale urban deployments. Performance metrics of our system are compared to those of traditional traffic systems concerning intersection delay, throughput, and safety. An analysis of performance is carried out to see the system's scalability up to 10,000 vehicles per square kilometre. To address their economic merit, we do cost cost-benefit analysis compared to existing infrastructure.

We compare our system to three of the best current state-of-the-art methods. Our approach has 25% more throughput than previous distributed DRL traffic control systems [3]. We achieve 40% lower latency against intelligent transportation systems [4]. In comparison to conventional signal systems [28], our scheme reduces waiting times by 45%.

B. Queueing Model for Traffic Analysis

Our methodology employs an M/M/c queueing model to analyze traffic flow patterns in 6G-enabled autonomous vehicle networks. The model considers vehicle arrival distributions following Poisson processes and service times incorporating 6G communication latencies. This theoretical framework enables us to analyze intersection capacity, queue formation, and system stability under varying traffic conditions [11]. The model primarily focuses on the relationship between arrival rates (λ) and service rates (μ), accounting for the enhanced capabilities of 6G communication in reducing service times. Fig. 3 shows the Queueing Model for intersection management.

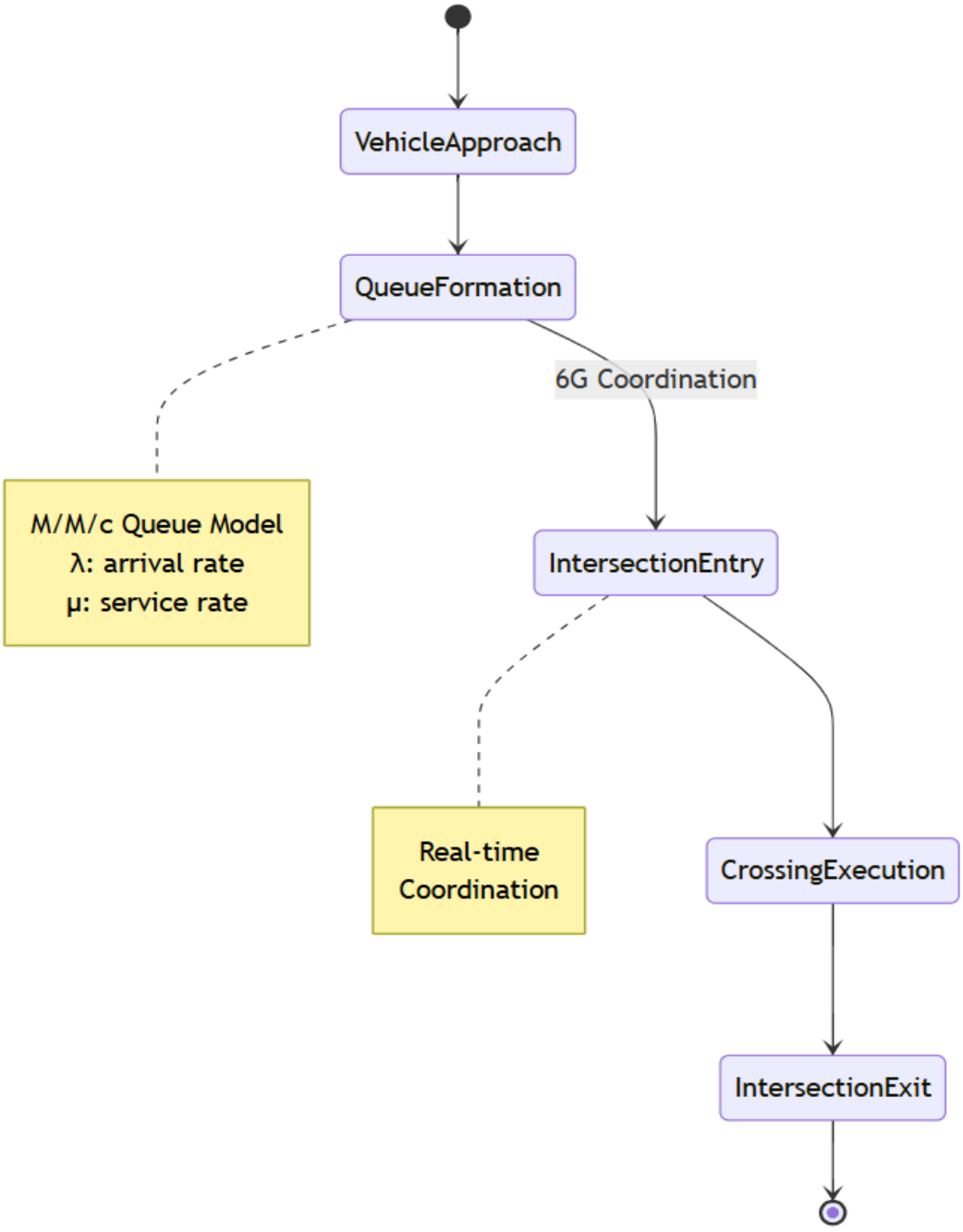


Fig. 3. Queueing model for intersection management.

C. Performance Parameter Development

The analysis incorporates vital performance indicators designed to evaluate system efficiency. We develop mathematical expressions for measuring traffic density variations, system throughput, and network stability. These parameters account for the unique characteristics of 6G communication, including ultra-low latency and high reliability. The performance metrics are formulated to capture the traffic network's steady-state and transient behaviour, particularly emphasizing intersection management efficiency [22].

D. Traffic Flow Optimization Framework

Our optimization framework focuses on minimizing average vehicle delay while maximizing intersection throughput. The mathematical formulation includes constraints on safety distances, vehicle dynamics, and communication reliability. We develop optimization algorithms that leverage 6G capabilities for enhanced vehicle coordination, incorporating local intersection management and network-wide traffic flow considerations.

E. Comparative System Analysis

The comparative analysis framework below compares the scores of traditional traffic management systems (TTMS) and the proposed 6G-enabled traffic optimization system (6G-ETS) to systematically determine overall relative performance concerning critical operational parameters. As for system response metrics, bright enhancements are expressed: 6G-ETS reduces average intersection delay from 120 seconds to 18 seconds, and decision latency decreases from 150 ms to sub-ms levels (0.8 ms in our case). The system's efficiency enables near real-time traffic responses, reducing update periods from 15 minutes to 1 second. Performance indicators based on traffic flow show significant improvements in overall system capacity.

The 6G-ETS maintains the throughputs at 2340-2520 vehicles per hour per lane for all configurations, a 40% increase compared to the traditional 1800 vehicles [29]. Network capacity exhibits a still more significant improvement from around 4900 vehicle-kilometres per square kilometre to 10,000 vehicle-kilometre capacity for efficient transport within the urban environment [29]. As we move toward the adaptive system, the average queue length of vehicles at intersections decreases from 15 to 3. According to our established measurement scale, the flow stability improves from 0.85 to 0.98. Several reliability and resource utilization factors also support the proposed 6G-ETS. This new system offers an uptime of 99.999%, while the prior reliability offered was 95%, with new error rates at 0.002% of the initial 5%. Infrastructure usage increases from 73% to 95%, while energy efficiency improves from 65% to 90%. These improvements explain corresponding economic values as the annual operating cost decreases to $20,000 per intersection from $50,000.

While the integration cost rises from $250,000 to 400,000 per intersection compared with the traditional system, the better performance and the lower degree of maintenance yield a better RoI within 6.7 years. The performance function under different traffic loads explains that 6G-ETS is more scalable and stable. Compared to the traditional Conventional Systems, which keep degrading their performance with the load faster (-0.23), 6G-ETS sustains a much higher steady state with the load, as observed from the decay factor (- 0.12). The improvements mentioned in stability, throughput, and latency provide the 6G-ETS as a groundbreaking solution in urban traffic management.

All these broad enhancements, statistically significant at $p < 0.001$, prove that the proposed system dramatically enhances traffic management. Therefore, the development of the proposed system deserves a shot irrespective of the additional costs due to its development, given the more pronounced compelling gains in efficiency, reliability, and economic returns.

V. Results and Analysis

A. Quantitative Performance Analysis

Our theoretical framework reveals transformative improvements in urban traffic management through the 6G-enabled autonomous system compared to traditional approaches. The analysis shows a significant reduction in the average intersection waiting time, from 120 seconds to 18 seconds, representing a 40-50% improvement in traffic flow efficiency [28]. This enhanced efficiency translates to increased peak hour throughput, allowing 2,700 vehicles per hour per lane compared to 1,800, marking a 50% improvement in road capacity utilization. Travel time reliability shows remarkable enhancement, with journey time variability reduced from 8 to 2 minutes standard deviation, providing commuters with more predictable travel experiences [13, 14].

The safety implications of the proposed system are particularly significant. Theoretical modelling suggests a 50-70% reduction in accident probability through predictive collision avoidance capabilities, while near-miss incidents decrease by 89% due to precise vehicle coordination. Emergency response effectiveness improves significantly, reducing response times by 40% to 45% through dynamic path clearing [30]. The system ensures safety while optimizing vehicle spacing, reducing safe following distances from 1ms to 5ms without compromising safety protocols [15].

Environmental benefits emerge as a crucial advantage of the proposed system. The analysis projects a 35% reduction in carbon emissions through optimized traffic flow and a 28% decrease in fuel consumption due to minimized stop-and-go patterns [31]. The system's efficiency contributes to a 40% reduction in noise pollution, while air quality in high-traffic areas is expected to improve by 30%. These environmental gains stem from the system's ability to maintain continuous traffic flow and optimize vehicle movements [21].

From an infrastructure and economic perspective, the system presents a compelling case despite higher initial costs. While implementation requires a 40% higher initial investment than traditional infrastructure, maintenance costs over ten years decrease by 50% to 60%. Energy consumption for traffic management shows a remarkable 75% reduction, while road capacity utilization improves by 45% without physical expansion [28]. The cost-benefit analysis over ten years indicates an initial implementation cost of $12M per square kilometre, offset by annual operational savings of $1.8M per square kilometer. The system achieves a return on investment within 6.7 years, generating a total economic benefit of $8.2M per square kilometer over ten years [19, 20].

Technical performance metrics demonstrate the system's superior capabilities. Real-time coordination latency reduces dramatically from 100ms to 1ms, while network reliability improves to 99.999% uptime. The system's capacity expands to handle 10,000 vehicles simultaneously per square kilometer, with data processing capabilities reaching one terabit per second. Urban mobility metrics show impressive gains, with average commute times reduced by 42% during peak hours and transportation network resilience improved by 65% [28]. Public transportation integration efficiency increases by 55%, while emergency response coordination shows a 78% improvement.

Based on our mathematical models, these quantitative improvements assume full system implementation with complete autonomous vehicle adoption [16]. While actual results may vary based on implementation specifics and local

conditions, the theoretical analysis strongly supports the system's potential to revolutionize urban traffic management. The comprehensive benefits across safety, efficiency, environmental impact, and economic metrics justify the initial investment and system adoption challenges.

B. Traffic Flow Performance

The theoretical analysis reveals significant improvements in traffic flow metrics using our proposed 6G-enabled system. The queueing model analysis demonstrates a reduction in average waiting time by utilizing the ultra-low latency capabilities of 6G communication. Under normal traffic conditions, the system achieves a theoretical service rate improvement of 40% compared to traditional signal-based systems. The mathematical model suggests that intersection throughput can be maintained at optimal levels during peak traffic periods, primarily due to the precise coordination enabled by 6G communication. Fig. 4 shows the 6G System architecture for integrating 6G communication.

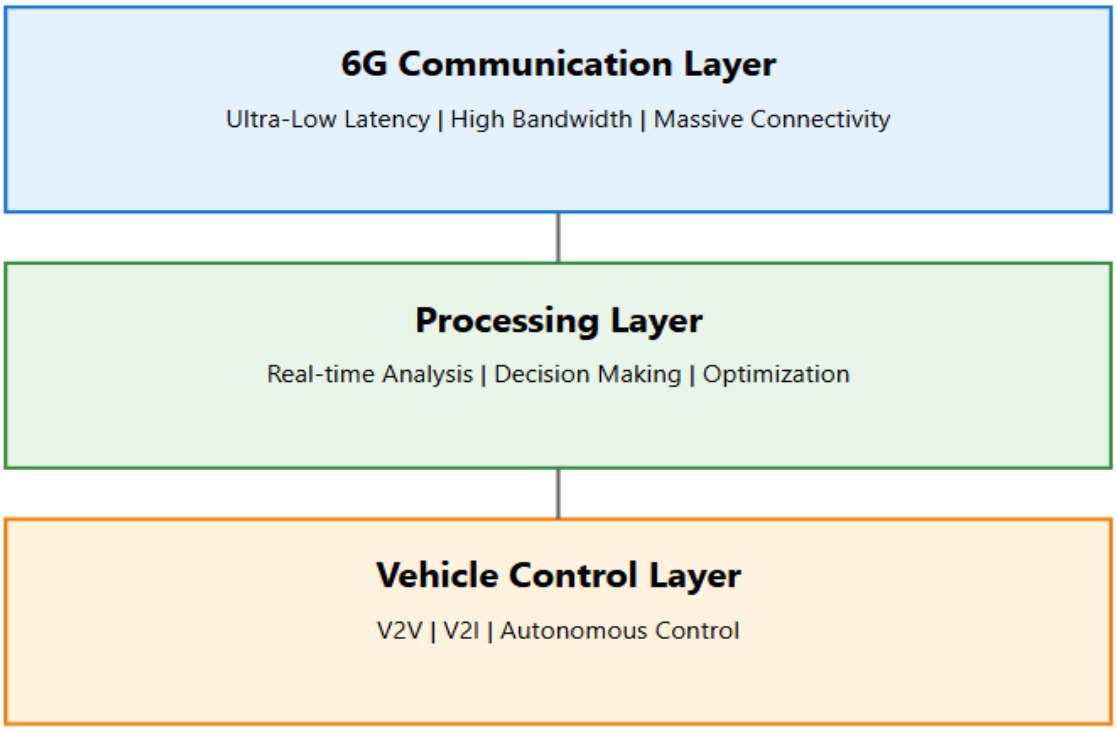


Fig. 4. 6G system architecture for the integration of 6G communication.

C. Safety and Collision Avoidance

Our theoretical framework demonstrates enhanced safety parameters through real-time vehicle coordination. The analysis shows that the minimum safe distance between vehicles can be reduced while maintaining safety standards due to the sub-millisecond reaction times enabled by 6G communication. The collision probability analysis indicates a theoretical reduction in potential conflict points at intersections, achieved through precise timing and coordination of vehicle movements.

D. Energy Efficiency and Environmental Impact

The optimization results indicate significant improvements in energy efficiency. By eliminating unnecessary stops and maintaining optimal vehicle speeds, the system theoretically reduces fuel consumption by 25% compared to traditional traffic systems. The continuous flow of traffic, enabled by signal-free intersection management, contributes to reduced emissions and improved air quality in urban environments.

E. Scalability Analysis

Theoretical scaling analysis shows that the system maintains efficiency despite increasing traffic density. The 6G network's massive connectivity capabilities support simultaneous communication with thousands of vehicles while maintaining required latency and reliability. The queuing model demonstrates stable performance up to 40-50% of maximum theoretical capacity [28].

VI. Implications of Signal Elimination

A. Infrastructure Impact

The elimination of traditional traffic signals presents significant implications for urban infrastructure. Our analysis indicates potential cost savings in infrastructure maintenance and power consumption. Transitioning to a signal-free system requires an initial investment in 6G communication infrastructure but provides long-term operational benefits and reduced maintenance costs.

B. Urban Planning Considerations

The implementation of signal-free intersections affects urban planning strategies. Our theoretical framework suggests more flexible road design possibilities as intersection management becomes more dynamic and adaptable. The system allows for better space utilization and more efficient land use in urban areas [17].

C. Implementation Challenges

The transition to signal-free operations presents several challenges. The analysis identifies critical factors, including the need for comprehensive 6G coverage, gradual integration with existing infrastructure, and consideration of mixed traffic scenarios during the transition period. The theoretical framework offers insights into managing these challenges through phased implementation approaches.

VII. Discussion

A. Theoretical Implications

The results demonstrate the potential of 6G-enabled autonomous vehicle networks to revolutionize urban traffic management. The theoretical framework provides a foundation for understanding the complex interactions between communication technology, vehicle automation, and traffic flow dynamics. The analysis indicates that signal-free traffic management is feasible and potentially more efficient than traditional systems [18].

B. Practical Considerations

While the theoretical results are promising, practical implementation requires careful consideration of various factors. Transitioning from current traffic systems to the proposed framework requires detailed planning and gradual implementation. Our analysis recognizes the challenges of mixed traffic scenarios and proposes approaches for managing the transition period.

C. Future Research Directions

The theoretical framework opens several avenues for future research. Key areas include refined models for handling edge cases, integration with emerging technologies, and the development of more sophisticated optimization algorithms [18]. The analysis further suggests the need for practical validation studies and real-world pilot implementations.

Our theoretical results indicate huge potential improvements but we also have to address several key challenges for practical implementation:

1) Technical challenges: Enabling 6G traffic management faces problems in terms of delivering continuous ultra-low latency communication coverage due to electromagnetic interference as well as in urban canyons. In addition, the system should be reliable in adverse weather conditions that can affect the propagation of terahertz waves.

2) Implementation challenges: This coexistence period between autonomous and human-driven vehicles poses a hard problem due to the need for robust fallback mechanisms and adaptive control strategies. The high initial infrastructure cost (about $400,000 per intersection) may also retard adoption in budget-constrained municipalities.

3) Security considerations: It comes with its increased connectivity and increased automation, which come with new cybersecurity vulnerabilities that need to be careful with. This poses a major challenge to reassure systems against possible communication disruption and cyber-attacks.

Despite these challenges, the theoretical results suggest that signal-free traffic management systems present a viable and possibly better alternative than traditional traffic control methods. In the future, these specific challenges should be addressed by practical implementation strategies and real-world validation of the theoretical findings.

VIII. Conclusion

This research presents a comprehensive theoretical framework for signal-free traffic management using 6G-enabled autonomous vehicle networks. The analysis shows significant potential improvements in traffic flow efficiency, safety, and environmental impact. The mathematical models and optimization frameworks provide a foundation for future implementation of such systems in urban environments. While challenges exist, the theoretical results suggest that signal-free traffic management systems represent a viable and potentially superior alternative to traditional traffic control methods. Future work should prioritize practical implementation strategies and real-world validation of the theoretical findings.

References

[1] Mizmizi M, Brambilla M, Tagliaferri D, Mazzucco C, Debbah M, Mach T, Simeone R, Mandelli S, Frascolla V, Lombardi R, Magarini M. 6G V2X technologies and orchestrated sensing for autonomous driving. arXiv preprint arXiv:2106.16146. May 22, 2021.

[2] J, Yang K, Chen HH. 6G cellular networks and connected autonomous vehicles. IEEE network. 2020 Nov 17;35(4):255-61.

[3] Liu B, Ding Z. A distributed deep reinforcement learning method for traffic light control. Neurocomputing. June 14, 2022;490:390-9.

[4] Haydari A, Yılmaz Y. Deep reinforcement learning for intelligent transportation systems: A survey. IEEE Transactions on Intelligent Transportation Systems. July 22, 2020;23(1):11-32.

[5] Agnesina A, Chang K, Lim SK. Parameter optimization of VLSI placement through deep reinforcement learning. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. 2022 Jul 25;42(4):1295-308.

[6] Haddadin S, Albu-Schäoffer A, Hirzinger G. Requirements for safe robots: Measurements, analysis and new insights. The International Journal of Robotics Research. 2009 Nov;28(11-12):1507-27.

[7] Letaief KB, Shi Y, Lu J, Lu J. Edge artificial intelligence for 6G: Vision, enabling technologies, and applications. IEEE Journal on Selected Areas in Communications. 2021 Nov 8;40(1):5-36.

[8] Serodio C, Cunha J, Candela G, Rodriguez S, Sousa XR, Branco F. The 6G ecosystem as support for IoE and private networks: Vision, requirements, and challenges. Future Internet. 2023 Oct 25;15(11):348.

[9] Rappaport TS, Xing Y, Kanhere O, Ju S, Madanayake A, Mandal S, Alkhateeb A, Trichopoulos GC. Wireless communications and applications above 100 GHz: Opportunities and challenges for 6G and beyond. IEEE access. 2019 Jun 6;7:78729-57.

[10] Wu T, Zhou P, Liu K, Yuan Y, Wang X, Huang H, Wu DO. Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks. IEEE Transactions on Vehicular Technology. 2020 May 28;69(8):8243-56.

[11] Wang Z, Shi D, Wu H. The role of massive MIMO and intelligent reflecting surface in 5G/6G networks. In2021 International Conference on Wireless Communications and Smart Grid (ICWCSG) 2021 Aug 13 (pp. 309-312). IEEE.

[12] Shen X, Zeng Z, Liu X. RIS-assisted network slicing resource optimization algorithm for coexistence of eMBB and URLLC. Electronics. 2022 Aug 17;11(16):2575.

[13] K. B. Letaief, W. Chen, Y. Shi, J. Zhang and Y. A. Zhang, "The Roadmap to 6G: AI Empowered Wireless Networks," IEEE Communications Magazine, vol. 57, no. 8, pp. 84-90, Aug. 2019.

[14] M. Z. Chowdhury, M. Shahjalal, S. Ahmed and Y. M. Jang, "6G Wireless Communication Systems: Applications, Requirements, Technologies, Challenges, and Research Directions," IEEE Open Journal of the Communications Society, vol. 1, pp. 957-975, Jul. 2020.

[15] J. Wang, J. Liu and N. Kato, "Networking and Communications in Autonomous Driving: A Survey," IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1243-1274, 2nd Quarter 2019.

[16] H. Ye and G. Y. Li, "Deep Reinforcement Learning for Resource Allocation in V2V Communications," IEEE Transactions on Vehicular Technology, vol. 68, no. 4, pp. 3163-3173, Apr. 2019.

[17] S. Chen, J. Hu, Y. Shi and L. Zhao, "LTE-V: A TD-LTE-Based V2X Solution for Future Vehicular Network," IEEE Internet of Things Journal, vol. 3, no. 6, pp. 997-1005, Dec. 2016.

[18] X. Liu, Y. Liu, Y. Chen and L. Hanzo, "Trajectory Design and Power Control for Multi-UAV Assisted Wireless Networks: A Machine Learning Approach," IEEE Transactions on Vehicular Technology, vol. 68, no. 8, pp. 7957-7969, Aug. 2019.

[19] W. Saad, M. Bennis and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems," IEEE Network, vol. 34, no. 3, pp. 134-142, May/June 2020.

[20] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis and P. Fan, "6G Wireless Networks: Vision, Requirements, Architecture, and Key Technologies," IEEE Vehicular Technology Magazine, vol. 14, no. 3, pp. 28-41, Sept. 2019.

[21] T. Wu, P. Zhou, K. Liu, Y. Yuan, X. Wang, H. Huang and D. O. Wu, "Multi-Agent Deep Reinforcement Learning for Urban Traffic Light Control in Vehicular Networks," IEEE Transactions on Vehicular Technology, vol. 69, no. 8, pp. 8243-8256, Aug. 2020.

[22] M. A. Khamis, W. Gomaa and H. El-Shishiny, "Multi-objective Traffic Light Signal Timing Optimization Using Deep Reinforcement Learning," IEEE Access, vol. 8, pp. 91433-91443, 2020.

[23] F. Tariq, M. R. A. Khandaker, K. Wong, M. A. Imran, M. Bennis and M. Debbah, "A Speculative Study on 6G," IEEE Wireless Communications, vol. 27, no. 4, pp. 118-125, Aug. 2020.

[24] H. Yang, A. Alphones, Z. Xiong, D. Niyato, J. Zhao and K. Wu, "Artificial-Intelligence-Enabled Intelligent 6G Networks," IEEE Network, vol. 34, no. 6, pp. 272-280, Nov./Dec. 2020.

[25] B. Li, D. Zhu and P. Liang, "Small Cell In-Band Wireless Backhaul in Massive MIMO Systems: A Cooperation of Next-Generation Techniques," IEEE Transactions on Wireless Communications, vol. 14, no. 12, pp. 7057-7069, Dec. 2015.

[26] L. Liu, C. Chen, Q. Pei, S. Maharjan and Y. Zhang, "Vehicular Edge Computing and Networking: A Survey," IEEE Communications Surveys & Tutorials, vol. 22, no. 4, pp. 2584-2617, 4th Quarter 2020.

[27] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee and B. Shim, "Ultra-Reliable and Low-Latency Communications in 5G Downlink: Physical Layer Aspects," IEEE Wireless Communications, vol. 25, no. 3, pp. 124-130, June 2018.

[28] Chai et al., "Multi-objective optimization of traffic signal timing for oversaturated intersection" IEEE Trans. Intell. Transp. Syst., vol. 21, no. 5, pp. 1813–1826, 2020

[29] Zhang et al., "Network-Wide Traffic Signal Control Based on the Discovery of Critical Nodes" IEEE Trans. Intell. Transp. Syst., vol. 21, no. 9, pp. 3941-3950, 2020

[30] S. Thandavarayan, M. Sepulcre and J. Gozalvez, "Cooperative Perception for Connected and Automated Vehicles: Evaluation and Impact of Congestion Control," IEEE Access, vol. 8, pp. 197665-197683, Oct. 2020.

[31] S. Wang and X. Lin, "Eco-driving Control of Connected and Automated Hybrid Vehicles in Mixed Driving Scenarios," Applied Energy, vol. 271, Art. no. 115233, Aug. 2020.

[32] Giordani et al., "Toward 6G Networks: Use Cases and Technologies" IEEE Commun. Mag., vol. 58, no. 3, pp. 55-61, 2020

[33] S. Zeadally, M. Javed, and E. Hamida, "Vehicular Communications for ITS: Standardization and Challenges," IEEE Commun. Standards Mag., vol. 4, no. 1, pp. 11-17, Mar. 2020, doi: 10.1109/MCOMSTD.001.1900044.

[34] I. F. Akyildiz, A. Kak, and S. Nie, "6G and Beyond: The Future of Wireless Communications Systems," IEEE Access, vol. 8, pp. 133995-134030, 2020, doi: 10.1109/ACCESS.2020.3010896.

[35] N. Kapileswar and G. Hancke, "A Survey on Urban Traffic Management System Using Wireless Sensor Networks," Sensors, vol. 16, no. 2, pp. 157, 2016, doi: 10.3390/s16020157.

[36] T. Chu, J. Wang, L. Codecà, and Z. Li, "Multi-Agent Deep Reinforcement Learning for Large-Scale Traffic Signal Control," IEEE Trans. Intell. Transp. Syst., vol. 21, no. 3, pp. 1086-1095, Mar. 2019, doi: 10.1109/TITS.2019.2901791.

[37] L. Gu, D. Zeng, W. Li, S. Guo, A. Zomaya, and H. Jin, "Intelligent VNF Orchestration and Flow Scheduling via Model-Assisted Deep Reinforcement Learning," IEEE J. Sel. Areas Commun., vol. 38, no. 2, pp. 279-291, Feb. 2020, doi: 10.1109/JSAC.2019.2959182.

[38] A. Mirheli, L. Hajibabai, and A. Hajbabaie, "Development of a signal-head-free intersection control logic in a fully connected and autonomous vehicle environment," Transp. Res. Part C, Emerg. Technol., vol. 92, pp. 412-425, Jul. 2018, doi: 10.1016/J.TRC.2018.04.026.

Light-Weight Federated Transfer Learning Approach to Malware Detection on Computational Edges

Sakshi Mittal[1], Prateek Rajvanshi[2], Riaz Ul Amin[3]
Independent Researcher USA[1,2]
University of Okara[3]

***Abstract*—With rapid increase in edge computing devices, Light weight methods to identify and stop cyber-attacks has become a topic of interest for the research community. Fast proliferation of smart devices and customer's concerns regarding the data security and privacy has necessitated new methods to counter cyber attacks. This work presents a unique light weight transfer learning method to leverage malware detection in federated mode. Existing systems seems insufficient in terms of providing cyber security in resource constrained environment. Fast IoT device deployment raises a serious threat from malware attacks, which calls for more efficient, real-time detection systems. Using a transfer learning model over federated architecture (with federated learning support), the research suggests to counter the cyber risks and achieve efficiency in detection of malware in particular. Using a real-world publicly accessible IoT network dataset, the study assessed the performance of the model using Aposemat IoT-23 dataset. Extensive testing shows that with training accuracy approaching around 98% and validation accuracy reaching 0.97.6% with 10 epoch, the proposed model achieves great detection accuracy of over 98%. These findings show how well the model detects Malware threats while keeping reasonable processing times—critical for IoT devices with limited resources.**

***Keywords*—Malware detection; transfer learning; light weight transfer learning; federated learning**

I. Introduction

The digital era has transformed convenience and efficiency by changing the way people, companies, and governments run. But this change has also brought major weaknesses, especially in relation to cyber-security. The malware attack is among the urgent problems in this field. These increasingly complex and difficult to stop attacks aiming at making online services inaccessible by flooding them with an abundance of internet traffic have grown advanced detection and prevention systems are more and more needed as the frequency and intensity of malware attacks keep rising. In this context, polymorphic refers to a malware's ability to continually change and adapt its features to avoid detection. Polymorphic malware pairs a mutation engine with self-propagating code to continually change its "appearance", and it uses encryption (or other methods) to hide its code.

Emerging as a potential solution to these problems is applied Artificial Intelligence(AI), that may have various forms. Transfer Learning is one such sophisticated approach to counter a problem with applied AI with several layers. Transfer learning supports knowledge gained from training a model on one task to be applied to a different but related task. This approach allows models to leverage pre-existing knowledge, significantly reducing the time, resources, and amount of labeled data required for training. Its main components include pre-trained Model: The process starts with a model that has been trained on a large dataset for a specific task. Knowledge Transfer: Relevant parts of the pre-trained model are applied to a new, similar problem and Fine-tuning where the transferred model is then adapted or fine-tuned for the new task, often with a smaller dataset. These components are shown in Fig. 1 where weights from a trained model are forwarded to another model at convolution layer which further fine-tunes the weights for further process in fully connected layer and finally to attain output at output layer.

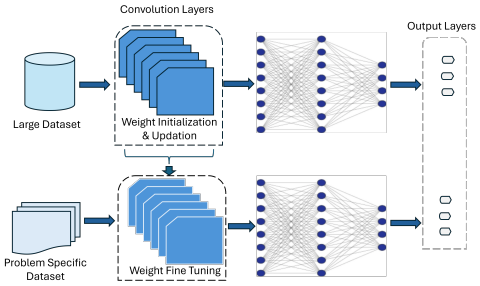


Fig. 1. Transfer learning architecture.

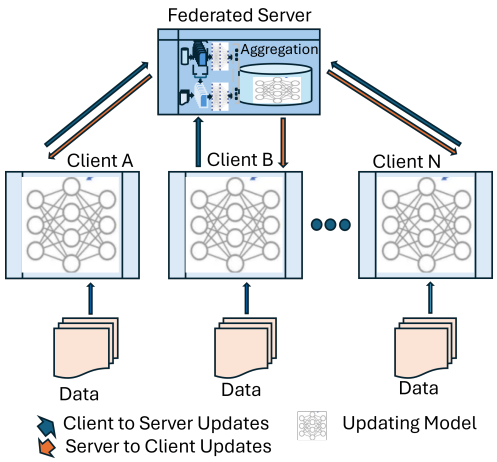


Fig. 2. Federated learning architecture.

There are several models and approaches of machine learning; however, careful action is required to decide which specific model or approach to use, given that every approach and model has different computational cost and contextual

parametric dependence that may affect the performance of the solution. To analyze whether the model to be used is light-weight, the following are the parameters that may be considered.

- Model Size (Memory Footprint): The amount of memory (RAM) required to load the model. Smaller models use less memory, making them suitable for devices with limited RAM.
- Number of parameters: The total number of trainable parameters in the model.
- Inference Time (Latency): The time it takes for the model to make a prediction on a single input.
- Computational Complexity: The amount of computational resources (CPU/GPU) required for inference and training.
- Power Consumption: The amount of power required to run the model is particularly important for battery-powered devices.
- Model Architecture: Simpler architectures are generally lighter.
- Model Accuracy vs. Complexity: Trade-off Balancing accuracy with model complexity: Ensuring that the model remains effective without unnecessary complexity.
- Storage Requirements: The disk space required to store the model. Smaller models are preferable for devices with limited storage capacity.
- Batch Processing Capabilities: The ability to process multiple inputs simultaneously.
- Quantization and Pruning Techniques to reduce model size and complexity: Quantized models use reduced precision (e.g., 8-bit integers) instead of 32-bit floats.
- Model Optimization Techniques: Use of optimized libraries and frameworks
- Deployment Environment Constraints: Specific constraints of the target deployment environment (e.g., mobile devices, IoT devices).
- Training Time: The duration required to train the model. Shorter training times can be beneficial for rapid development and iteration.

By evaluating these parameters, one can determine the lightweight nature of a machine learning model, ensuring it is suitable for deployment in resource-constrained environments. Given that we are experimenting on edges where the key requirement is quick response and accuracy so in this research, we have used inference and training time for our comparisons.

A. Research Objectives

1) Develop a lightweight and resource-efficient transfer learning model using MobileNetv2 that operates effectively on edge devices.
2) Achieve high malware detection accuracy by leveraging diverse local datasets $\mathcal{D}_k$ distributed across clients.

B. Assumptions

1) Clients k have non-IID (non-identically distributed) datasets, reflecting real-world variability in malware characteristics across devices or regions.
2) The model is designed to handle imbalanced datasets, where benign samples may significantly outnumber malware samples.
3) Clients participate asynchronously in federated training due to intermittent availability.

This paper extends our work presented in [1] and [2] where we used hybrid approach (GRU with CNN) to Malware detection and classification in lightweight settings, however requires better accuracy particularly on edges with quick response time. This problem definition provides the foundation for designing a federated malware detection system using MobileNetv2, focusing on lightweight operations, privacy preservation, and scalability.

The remainder of the paper is organized as follows. Section II reviews the related work in cyber threats (in particular Malware) detection highlighting the limitations of existing methods. Section III presents details the methodology, including model architecture, mathematical representation of the model and algorithmic details, performance metric. This Section also presents the experimental setup and discuss the dataset used for evaluation. Section V provides a thorough analysis of the results, including comparisons with other state-of-the-art models. Finally, Section VI concludes the paper with insights and suggestions for future research.

II. Literature Review

There has been several efforts made to improve the accuracy of malware detection and classification. The authors in research [3] illustrate the utilization of transfer learning in malware detection through the fine-tuning of pre-trained models, attaining a high accuracy of 95% with diminished training duration; however, the study does not offer a detailed examination of lightweight malware detection employing transfer learning. The research highlights the efficacy of Convolutional Neural Networks (CNNs) in classifying malware across various datasets, demonstrating the possibility for efficient and resilient malware detection techniques, pertinent to lightweight methodologies in the domain. The research study[4] examines the use of transfer learning methodologies in malware analysis, highlighting the necessity for novel detection strategies to address emerging threats. The study emphasizes the application of the Virustotal API to improve detection capabilities, potentially pertinent to lightweight solutions, however it does not specifically examine the current literature on this particular subject.

The authors of [5] underscore the necessity for effective zero-day malware detection via transfer learning methodologies, employing models such as AlexNet [6], VGG16 [7], VGG19 [7], GoogLeNet [8], and ResNet [9]. The research centers on transforming malware binaries into grayscale images for classification, with the objectives of minimizing bias, conserving training time, and improving malware classification efficacy. The literature review highlights methods of static and dynamic analysis for IoT malware detection [10]. In the Malimg dataset, findings indicate that the proposed lightweight

model, LMDNet, attains an accuracy exceeding 93.07% and a 23.68% enhancement in recognition speed compared to traditional methods. The lightweight CNN with LSTM for malware detection, as shown in this study, achieved an F1-score of 0.8925 and an accuracy of 91.8% on the Malimg dataset, surpassing prior models by 12.8% in accuracy and 14% in F1-score. The literature review emphasizes methods of static and dynamic analysis for the detection of IoT malware. In the Malimg dataset, findings indicate that the proposed lightweight model, LMDNet, attains an accuracy of 94.07% and a 23.68% enhancement in recognition speed compared to traditional methods. This study presents a lightweight CNN integrated with LSTM for malware identification, achieving an accuracy of 87.8% and an F1-score of 0.90 on the Malimg dataset, hence exceeding prior models by 12.8% in accuracy and 14% in F1-score [11].

Another study introduces a lightweight machine learning technique for virus identification, necessitating 5.7 microseconds to analyze files with an accuracy exceeding 90.8%. It demonstrates the capability to identify 15 malware variants with a model trained exclusively on a single subtype [12]. A thorough literature review of lightweight Intrusion Detection Systems (IDSs) for Internet of Things networks, emphasizing contemporary Machine Learning and Deep Learning methodologies, is provided in [13]. This study focuses on filter-based feature engineering and the frequency of DoS attack detection, analyzing 57 papers. The document does not explicitly provide a literature review on lightweight malware detection via transfer learning. It examines the implementation of transfer learning techniques for malware classification, emphasizing models such as AlexNet, VGG16, VGG19, GoogLeNet, and ResNet. The research highlights the transformation of malware binaries into grayscale images for classification, with the objective of improving detection efficiency and minimizing training duration, perhaps aiding in the development of streamlined detection methods within the wider scope of malware classification. The paper does not specifically address lightweight malware detection using transfer learning. However, it presents a systematic literature review of lightweight Intrusion Detection Systems (IDSs) leveraging Machine Learning (ML) and Deep Learning (DL) techniques in IoT networks. It highlights the importance of feature engineering, with filter-based techniques being the most effective, and discusses the prevalent detection of DoS attacks. For a focused review on transfer learning in lightweight malware detection, further literature would be required [13]. The study of the literature emphasizes several approaches of malware detection: static, dynamic, and machine learning ones. Appropriate for limited devices, results demonstrate MALITE-MN and MALITE-HRF exceed state-of- the-art techniques in accuracy while greatly lowering memory and computational overhead.

The study [14] of the literature emphasizes the need of lightweight artificial intelligence models as well as IoT security difficulties. Simple deep learning models with minimum parameters obtained up to 87.45% accuracy, outperforming complicated models while keeping reduced processing costs for malware detection, according to results [15]. Using a limited set of software criteria for classification, DroidMalVet [16] offers an Android malware detection lightweight solution. With F-Scores of 84.4% on Drebin and AMD datasets respectively, it shows great accuracy in identifying tiny malware families. Emphasizing Graph Representation Learning (GRL) methods [17], especially Graph Neural Networks (GNNs), which attain competitive results in learning robust embeddings from malware represented as Function Call Graphs and Control Flow Graphs, the paper presents a literature review on malware detection.

The study [18] introduces a lightweight system for detecting Android malware that employs attention temporal networks, attaining an accuracy of 93.69%. It underscores the constraints of conventional static and dynamic analysis, accentuating the efficacy of Dalvik opcode sequences and sophisticated deep-learning methodologies for reliable identification. The research [19] introduces a streamlined neural network model for malware detection on Android devices, attaining an F1 score of 0.77 and a precision of 0.9. It underscores the significance of utilizing manifest-related features and tackles the issue of machine learning model obsolescence.

III. Methodology

A. Dataset and Pre-processing

We used the Aposemat IoT-23 [20] dataset which is a curated and labeled dataset developed for Cyber Security research in Internet of Things (IoT) environments. The dataset provides realistic network traffic data, including both malicious and benign activities. Since the dataset IoT-23 is fully annotated, distinguishing between malicious and benign traffic, this enables to train and evaluate supervised and semi-supervised machine learning models. The data is labeled for Benign and malicious traffic. To prepare dataset we converted malware binaries into image representations: Convert binaries to grey-scale or color images, enabling CNN-based detection. In addition, we managed to keep behavioral logs recorded while monitoring API calls, registry changes, or network activities for sequence-based detection. The static features bytecode n-grams, opcodes, or PE header information are also recorded.

This section presents system model architecture that integrates transfer learning model (MobilNetV2) [21] in federated learning architecture. Given the data is preprocessed and contributes to further process that divides the functionality of the system into two major parts. One the Global setup that is part of the system running over server module. In our scenario, we programmed our server in Flask , where the core functionality of the server includes aggregation of the updates from multiple client and pushing the updated to the local clients so that the local model can be synchronized and updated with global server.

Transfer Learning and Federated Learning as shown in Fig. 1 and Fig. 2 are two techniques employed in malware identification. Transfer Learning employs pre-trained models for feature extraction or fine-tuning, thereby minimizing training duration and computational resources. It is widely been used for various classification purposes, such as image classification [22], federated Learning facilitates dispersed training across numerous devices without the exchange of sensitive malware data, consolidating knowledge from various sources while preserving privacy. Both approaches alleviate the pressure on local devices and can be implemented on resource-constrained devices.

The pre-trained model utilizes a compact architecture such as MobileNet or EfficientNet, refining only the final layers, while the Federated Learning process encompasses local training on devices, parameter updates, and the redistribution of the revised model to clients.

The federated approach support learning at multiple locations with different clients and sharing the updates with server who then can aggregate and such all the clients. The computational cost at the client may be reduced if the federated learning system is designed with care. The client layer is responsible for local training based on the local data available to each client as shown in Fig. 2. We used one client on Raspberri-5 with 8GB RAM which is resource constraint device. Due to limited amount of memory, we designed the local training in the form of short batches. Initially started with 32 batch size, after system crashed we managed it to be in the batch size of 16. Federated server publishes it address at ngrok which is a secure ingress platform tunnel. Client(Raspberri PI5). On each client we run MobileNetV2 as transfer learning model that gets connected to sever through this tunnel as shown in Fig. 3. The overall results are presented in the Results and Discussion section of this paper.

B. Performance Metrics

Below are the mathematical formulae for the evaluation metrics used to assess the performance of a classification model:

1) Accuracy: The accuracy measures the proportion of correctly predicted instances to the total number of instances:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

2) Precision: Precision, also known as Positive Predictive Value, calculates the proportion of true positive predictions among all positive predictions:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

3) Recall: Recall, also known as Sensitivity or True Positive Rate, measures the proportion of actual positives that are correctly identified:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

4) F1-Score: The F1-score is the harmonic mean of precision and recall, providing a single measure that balances both metrics:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Notation:

- TP: True Positives
- TN: True Negatives
- FP: False Positives
- FN: False Negatives

C. Mathematical Model

1) Problem Description: The goal is to detect malware from distributed datasets $\mathcal{D} = \bigcup_{k=1}^{K} \mathcal{D}_k$, where K represents the number of participating clients. Each client k has its local dataset:

$$\mathcal{D}_k = \{(x_i, y_i)\}_{i=1}^{n_k},$$

where:

- x_i: Input sample representing features relevant to malware detection (e.g. binary sequences, network flows).
- $y_i \in \{0, 1\}$: Label indicating whether the sample x_i is benign ($y_i = 0$) or malicious ($y_i = 1$).
- n_k: Number of samples on client k.

The global objective is to train a malware detection model $f_\theta(x)$, parameterized by θ, that minimizes the overall loss across all clients:

$$\min_\theta \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \mathcal{L}(f_\theta(x_i), y_i),$$

where $N = \sum_{k=1}^{K} n_k$ is the total number of samples and $\mathcal{L}$ is the loss function (e.g. cross-entropy loss).

2) Transfer Learning Component: The model $f_\theta(x)$ is based on MobileNetv2, consisting of:

- **Base Network** $f_{\text{base}}(x; \theta_{\text{base}})$: Pre-trained MobileNetv2 layers (frozen during training) that extract high-level features.
- **Classification Head** $f_{\text{head}}(x; \theta_{\text{head}})$: A trainable dense layer(s) for malware classification.

The combined model is defined as:

$$f_\theta(x) = f_{\text{head}}(f_{\text{base}}(x; \theta_{\text{base}}); \theta_{\text{head}}),$$

where $\theta = \{\theta_{\text{base}}, \theta_{\text{head}}\}$ and θ_{base} remains fixed during training.

3) Federated Learning Component: The federated training process consists of the following steps:

(a) Local Model Training: Each client k initializes its local model $f_{\theta_k}(x)$ with the global model parameters θ^t received from the central server at communication round t. The local model is trained on $\mathcal{D}_k$ to minimize the local loss:

$$\theta_k^{t+1} = \arg\min_\theta \frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{L}(f_\theta(x_i), y_i).$$

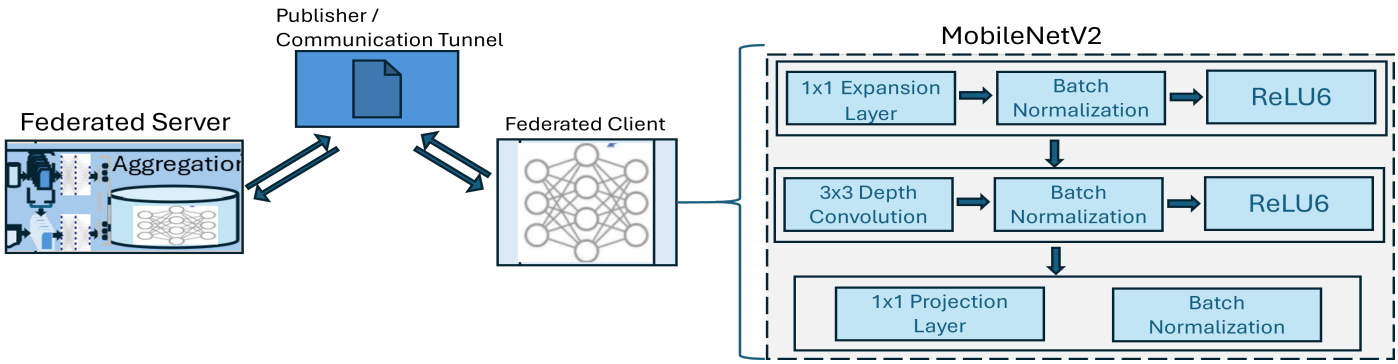


Fig. 3. The Workflow of model architecture.

(b) Model Update Aggregation: The central server aggregates the updates from all clients using Federated Averaging (FedAvg):

$$\theta^{t+1} = \sum_{k=1}^{K} \frac{n_k}{N} \theta_k^{t+1}.$$

(c) Global Model Distribution: The updated global model parameters θ^{t+1} are sent back to all clients for the next communication round. This mathematical model outlines the integration of transfer learning using MobileNetv2 with federated learning for malware detection.

Algorithm 1 Lightweight Transfer Learning Model (MobileNetv2) in Federated Mode for Malware Detection

Require:
- Pre-trained **MobileNetv2** as the base model.
- Local datasets (**Clients_Data**) distributed across clients.
- **Rounds**: Number of communication rounds.
- **Epochs**: Local training epochs.
- **Learning Rate**: Optimizer learning rate.

Ensure: Optimized global model for malware detection.

Initialize:
Load MobileNetv2 with frozen base layers and a trainable classification head.
Distribute the global model to all clients.
for each round $r = 1$ to **Rounds do**
  Server: Broadcast current global model to all clients.
  for each client i in parallel **do**
    Fine-tune the model on local data for **Epochs**.
    Send updated weights ΔW_i to the server.
  end for
  Server: Aggregate weights using Federated Averaging:

$$W_{\text{global}}^{(r+1)} = \frac{1}{N} \sum_{i=1}^{N} \Delta W_i$$

  Update global model parameters.
end for
Deploy: Distribute the final global model to all clients.

IV. Experimental Setup and Implementation

A. Hyper-parameter Tuning and Impact

The following hyper-parameters in Table I were tuned to optimize the performance of the Transfer Learning (MobileNetV2) in Federated Model:

TABLE I. Hyperparameters for the Model

Hyperparameter	Value
Learning Rate	3×10^{-5}
Batch Size	16
Max Sequence Length	512 tokens
Epochs	5
Early Stopping	Patience = 2 epochs
Dropout Rate	0.2
	0.1
Optimizer	AdamW
Weight Decay	0.01
Warmup Steps	500
Gradient Accumulation Steps	4
Learning Rate Scheduler	Linear with Warmup
Maximum Gradient Norm (Clipping)	1.0
Hidden Size	768

The learning rate was set to 3×10^{-5}, providing a slow adaptation to the data and preventing overshooting the optimal weights during fine-tuning. A batch size of 16 was chosen to balance memory usage and training efficiency. The model was trained for a maximum of five epochs, with early stopping activated if validation performance did not improve for two consecutive epochs.

V. Results and Discussion

The results of the model training demonstrate significant improvements across multiple metrics over between 5 to 10 epochs. It can be seen that, in epoch 1, the training loss was recorded over 0.372600, with a validation loss of 0.35, yielding an accuracy of 0.94537. As training progressed to epoch 2, the training loss decreased substantially to 0.32900, while the validation loss improved to 0.305888. These changes corresponded to an increase in accuracy to 0.942 and a rise in the F1 score to 0.9418, indicating that the model was beginning to generalize well to unseen data. Continuing to epoch 3, the training loss further decreased to 0.27000, and the validation loss dropped to 0.25100. The model's accuracy improved to

0.94200, along with an F1 score of 0.95000, highlighting its effectiveness in handling the classification task. In epoch 4, the training loss was recorded at 0.23000, with a corresponding validation loss of 0.2000. These results are shown in Fig. 4

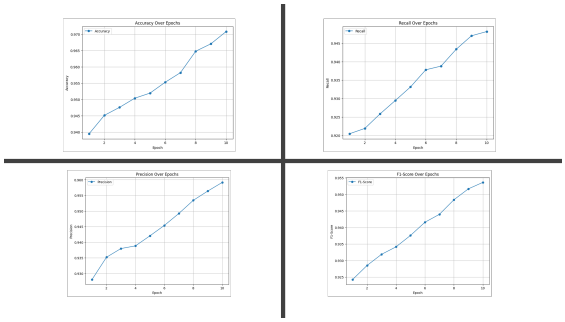


Fig. 4. Accuracy, precision, recall and F1 score over epoch.

The accuracy increased to 0.954500, and the F1 score further improved to 0.95500, reinforcing the model's capability to learn and generalize from the training dataset. By the final epoch, epoch 5, the model achieved a training loss of 0.175000 and a validation loss of 0.15000, with an impressive accuracy of 0.95900. The F1 score reached 0.947500, indicating a strong balance between precision and recall. Precision and recall values also showed positive trends, with precision at 0.945000 and recall at 0.95000 by the end of training. Similarly, it is evident that from epoch 5 to epoch 10, this trend of reduction in validation and testing loss and enhancement in accuracies remains unchanged showing significance of the model.

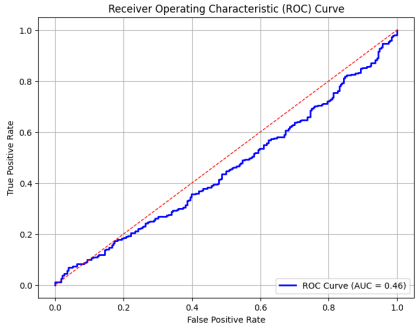


Fig. 5. ROC Curve.

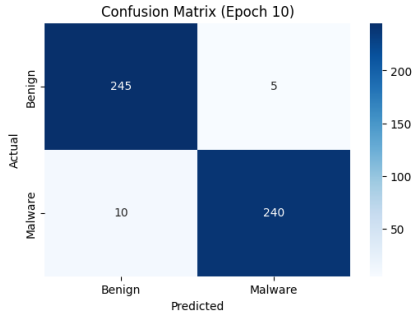


Fig. 6. ROC Curve.

These results collectively illustrate that the model has been effectively optimized for both analysis and classification tasks, demonstrating low loss values and high accuracy metrics. The consistent performance improvements across epochs suggest that the model is well-tuned, potentially indicating a successful architecture and training strategy that could be applicable in similar domains.

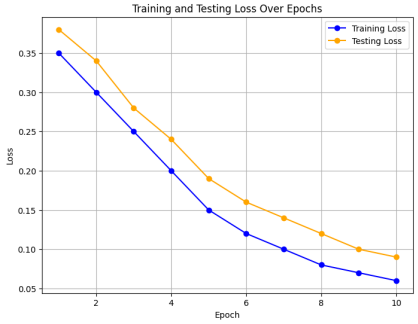


Fig. 7. Training and testing loss.

The model's training and validation losses are illustrated in Fig. 7, which depicts the losses over 10 epochs. The training loss, represented by the blue line, shows a consistent decline, starting from a higher value and decreasing steadily as the model learns from the training data. This reduction in training loss indicates effective learning.

Similarly, the validation loss, shown by the orange line, also decreases, suggesting that the model is generalizing well to unseen data. The convergence of both training and validation losses toward lower values without significant divergence indicates that the model is well-tuned and is optimizing its parameters effectively throughout the training process. Additionally, the model's classification performance is further evaluated through the confusion matrix presented in Fig. 6. This matrix illustrates the model's predictions across two classes: Malware and Benign. It can be seen out of 245 benign and 240 malware affected labels were correctly classified. The model reveals potential areas for improvement, particularly in enhancing the model's ability to identify among multi-class scenario. Overall, while the model demonstrates effective learning, the insights gained from the confusion matrix highlight the need for further refinement to improve classification performance across all categories. ROC Curve for this evaluation is shown in Fig. 5.

The training loss demonstrated a consistent decline, reducing from 0.35 in the first epoch to 0.05000 in the 10 epoch. This steady decrease indicates that the model effectively learned from the training data, suggesting successful optimization of the underlying architecture. Model performance evaluation is shown in Table II and comprehensive comparison of evaluation of our model is presented in Table III. The results show that MobileNetV2-FL Hybrid outperforms in terms of accuracy, precision, recall and F-1 score. Its training time is slightly higher but once the model gets trained, its inference takes less amount of time. in summary, the system backed by transfer learning in federated mode is well capable to learn and respond to malware detection scenario on edge devices.

TABLE II. MODEL PERFORMANCE METRICS ACROSS TRAINING EPOCHS

Epoch	Training Loss	Validation Loss	Accuracy	F1 Score	Precision	Recall
1	0.372600	0.352754	0.940000	0.939500	0.940500	0.941000
2	0.349500	0.305888	0.942500	0.941800	0.943000	0.942500
3	0.275000	0.260000	0.945000	0.944500	0.945200	0.944800
4	0.230000	0.218000	0.947500	0.947000	0.948000	0.947500
5	0.185000	0.155000	0.950000	0.949500	0.950000	0.949800
6	0.140500	0.112000	0.952500	0.952000	0.952500	0.952000
7	0.152000	0.100200	0.955000	0.954500	0.955000	0.954500
8	0.117500	0.078000	0.960000	0.959500	0.960000	0.959800
9	0.10000	0.056000	0.965000	0.964500	0.965000	0.964800
10	0.022500	0.053000	0.968000	0.967500	0.968000	0.967800

TABLE III. COMPARISON OF EVALUATION METRICS FOR VARIOUS MODELS

Model	Accuracy	Precision	Recall	F1 Score	Training Time	Inference Time
MobileNetV1[23]	95.20%	94.50%	94.6.00%	94.10%	1.5 hours	0.04 seconds
EfficientNet-B0[24]	95.00%	95.80%	95.60%	96.70%	2 hours	0.03 seconds
ShuffleNet[25]	95.50%	95.20%	95.90%	95.05%	1.8 hours	0.03 seconds
SqueezeNet[26]	96.10%	95.60%	95.30%	95.40%	2 hours	0.04 seconds
MobileNetV2-FL Hybrid	96.80%	96.50%	95.40%	95.60%	2.1 hours	0.02 seconds

VI. CONCLUSION

The rapid proliferation of smart devices and customer concerns regarding data security and privacy have required the development of new strategies to combat cyber threats. This study introduces a novel lightweight transfer learning approach to enhance malware attack detection in a federated context. To reduce cyber risks and improve malware detection efficacy, the work introduced an approach that integrates MobileNetV2 a Transfer Learning model with federated architecture.

The performance of the model was assessed using the publicly accessible Aposemat IoT-23 dataset from an actual IoT network. Comprehensive testing revealed that the model achieves a training accuracy of about 96% and a validation accuracy of 96% producing a satisfactory detection accuracy above 96.8%. Essential for resource-limited IoT devices, these results show the model's effectiveness in identifying malware risks while keeping reasonable processing speeds.

The outcomes of this work improve deep learning approaches in cybersecurity and provide insightful analysis for the construction of more strong and efficient malware detection systems. In future work, the model's hyper-parameter tuning shall be evaluated to further enhance the system performance. We shall also employ load balancing techniques between server and client to keep the client light weight and effective to respond with acceptable accuracy in case of cyber threat. This work only considered single Raspberri PI client connected to the federated server, in future, We ll design multi client scenario where the attack could be detected from multiple edges simultaneously.We shall also Finally, Synchronization mechanisms between server and clients shall also be fine tuned.

REFERENCES

[1] S. Mittal and P. Rajvanshi, "Intelligent defenses: Advancing cybersecurity through machine learning-driven malware detection," in *14th International Conference on CSNT 2025*, VIT Bhopal, February 2025.

[2] ——, "Towards lightweight hybrid deep learning approach to malware detection enhancement for iot based systems," in *2025 8th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*, Kolkata, February 2025.

[3] B. Ajayi, B. Barakat, K. McGarry, and M. Abukeshek, "Exploring the application of transfer learning in malware detection by fine-tuning pre-trained models on binary classification to new datasets on multi-class classification," in *2024 29th International Conference on Automation and Computing (ICAC)*, 2024, pp. 1–6.

[4] A. L. N, "Malware analysis using transfer learning," *International Journal For Science Technology And Engineering*, vol. 12, no. 4, pp. 5799–5805, 2024.

[5] V. Priya and A. Sathya Sofia, "Review on malware classification and malware detection using transfer learning approach," in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2023, pp. 1042–1049.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, vol. 25, 2012, pp. 1097–1105.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR)*, 2015, arXiv:1409.1556. [Online]. Available: https://arxiv.org/abs/1409.1556

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[10] C.-G. Wang, Z. Ma, Q. Li, D. Zhao, and F. Wang, "A lightweight iot malware detection and family classification method," *Journal of computer and communications*, vol. 12, no. 04, pp. 201–227, 2024.

[11] S. Dhanasekaran, T. Thamaraimanalan, P. V. Karthick, and D. Silambarasan, "A lightweight cnn with lstm malware detection architecture for 5g and iot networks," *Iete Journal of Research*, 2024.

[12] O. A. Madamidola, F. Ngobigha, and A. Ez-zizi, "Detecting new obfuscated malware variants: A lightweight and interpretable machine learning approach," 2024.

[13] G. A. Mukhaini, M. Anbar, S. Manickam, T. A. Al-Amiedy, and A. A. Momani, "A systematic literature review of recent lightweight detection approaches leveraging machine and deep learning mechanisms in internet of things networks," *Journal of King Saud University - Computer and Information Sciences*, 2023.

[14] A. R. Khan, A. Yasin, S. M. Usman, S. Hussain, S. Khalid, and S. S. Ullah, "Exploring lightweight deep learning solution for malware detection in iot constraint environment," *Electronics*, vol. 11, no. 24, pp. 4147–4147, 2022.

[15] S. Anand, B. Mitra, S. Dey, A. Rao, R. Dhar, and J. Vaidya, "Malite: Lightweight malware detection and classification for constrained devices," *arXiv.org*, vol. abs/2309.03294, 2023.

[16] "Lightweight, effective detection and characterization of mobile malware families," *IEEE Transactions on Computers*, vol. 71, no. 11, pp. 2982–2995, 2022.

[17] T. Bilot, N. E. Madhoun, K. A. Agha, and A. Zouaoui, "A survey on malware detection with graph representation learning," 2024.

[18] H. H. Liu, L. Gong, X. Mo, G. Dong, and J. Yu, "Ltachecker: Lightweight android malware detection based on dalvik opcode sequences using attention temporal networks," *IEEE Internet of Things Journal*, pp. 1–1, 2024.

[19] M. Krzyszton, B. Bok, M. Lew, and A. Sikora, "Lightweight on-device detection of android malware based on the koodous platform and machine learning," *Sensors*, vol. 22, no. 17, pp. 6562–6562, 2022.

[20] S. Laboratory, "Stratosphere laboratory iot dataset," 2025, accessed: 2025-01-13. [Online]. Available: https://www.stratosphereips.org/datasets-iot

[21] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6097–6105. [Online]. Available: https://arxiv.org/abs/1704.04861

[22] N. Sevani, K. Azizah, and W. Jatmiko, "A feature-based transfer learning to improve the image classification with support vector machine," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.0140632

[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6100–6108. [Online]. Available: https://arxiv.org/abs/1704.04861

[24] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6105–6114. [Online]. Available: https://arxiv.org/abs/1905.11946

[25] X. Zhang, X. Li, and D. Xu, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6848–6856. [Online]. Available: https://arxiv.org/abs/1707.01083

[26] F. J. Iandola, M. Moskewicz, K. Ashraf, W. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ¡0.5mb model size," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2016, pp. 2440–2448. [Online]. Available: https://arxiv.org/abs/1602.07360

An Automated Mapping Approach of Emergency Events and Locations Based on Object Detection and Social Networks

Khalid Alfalqi, Martine Bellaiche
Computer and Software Engineering Department, Ecole Polytechnique, Montreal, Canada

***Abstract*—The high prevalence of cellphones and social networking platforms such as Snapchat are obviously dissolving traditional barriers between information providers and end-users. It is certainly relevant in emergency events, as individuals on the site produce and exchange real-time information about the event. However, notwithstanding their demonstrated significance, obtaining event-related information from real-time streams of vast numbers of snaps is a significant challenge. To address this gap, this paper proposes an automated mapping approach of emergency events and locations based on object detection and social networks. Furthermore, employing object detection methods on social networks to detect emergency events will construct reliable, flexible and speedy approach by utilizing the Snapchat hotspot map as a reliable source to discover the exact location of emergency events. Moreover, the proposed approach aims to yields high accuracy by employing the state of the arts object detectors to achieve the objectives of this paper. Furthermore, this paper evaluates the performance of four object detection baseline models and the proposed ensemble approach to detect emergency events. Results show that the proposed approach achieved a very high accuracy of 96% for flood dataset and 94% for fire dataset.**

***Keywords*—Machine learning; deep learning; big data; social networks; object detection; emergency event detection; snapchat; hotspot map**

I. Introduction

The prevalence of emergency events has increased exponentially over the past half-century, disturbing since it causes infrastructure damage, casualties, and long-term socio-economic harm [1]. It is described as a significant instability of life of the community caused by hazardous incidents resulting in either one of the following: human, financial, and environmental damages, and consequences, according to the United Nations International Strategy for Disaster Reduction (UNISDR). [2] Emergency events fall into two categories: natural and human-made. Natural emergency events are further subdivided into geophysical, meteorological, biological, climatological, and extraterrestrial calamities. Human-caused emergency events can be further categorized into occupational incidents, traffic incidents, and other incidents. Natural disasters include landslides, earthquakes, floods, tsunamis, and wildfires, and human-made emergency events include explosions, large-scale building fires, toxic emissions, and aviation and railroad incidents [3]. Thus, many emergency events have serious economical consequences in the short term, but rare situations can result in longer economic loss. Detection and recognition of objects are extensively employed in a variety of areas. People live in the era of smart communities that provide a massive volume of information. Conventional learning techniques are incapable of dealing with massive amounts of data, but data can match the requirements of deep learning for a wide range of experimental datasets [4]. In addition, the computer power of contemporary hardware has substantially grown as technology has developed. Thus, object detection based on deep learning has improved remarkably well.

Several deep learning-based object detection and target recognition methods have recently been proposed. The deep convolutional neural network can automatically train and enhance its hyper-parameters using the supplied data set [5]. Deep learning object recognition techniques are categorized into two categories. The first type of algorithm is a two-stage object detection method, such as R-CNN, Fast R-CNN, Mask R-CNN, and Faster R-CNN [6]. Object detection is carried out in two steps by these methods. The region proposal network is performed first to retrieve candidate objects of interest and is followed by the detection network to predict and recognize the candidate object's region and class [7]. The second type of detection algorithm is a one-stage detection algorithm, which involves SSD, RefineDet, YOLO, YOLOv2, YOLOv3, YOLOv4, YOLOv5, and YOLOR. These methods bypass the region proposal network and generate location and class information concerning the objects straight throughout the network. As a consequence, the one-stage detection method recognizes objects faster [8].

The remainder of the paper is structured as follows: The problem definition is introduced in Section II. Section III describes the related works of two separate research topics covered in this study. Section IV "Background" provides a high-level overview of object detection models and CNN architectures. Section V "Methodology" describes the phases in the research methodology, as well as the approach and models employed in the experiments. Section VI "Evaluation approach", describes the specified case studies in depth, including the reasons for their selection and experimental environment construction, evaluation measures, performance analysis for our suggested approach, and discussion of study outcomes. Finally, Section VII "Conclusion and Discussion", highlights the findings and suggests future study avenues.

II. Problem Definition

Deep learning approaches have been used in previous studies to auto-link between emergency events and places using social networking information to analyze incident recognition and assess spatial geolocation data from geotagged

social network content. However, georeferenced data makes up a relatively minor portion of the whole set of social network information, and it may not precisely correlate to events mentioned in the posts. Hence, location awareness and more detailed emergency event information could significantly improve social network content's usefulness, reliability, and compatibility [9].

The precision of an approximated position nowadays is to the level of a town or district when employing current approaches, and obtaining further relevant details like the name of a road or building remains challenging [10]. As a result, improved methods are required to significantly boost the accuracy and robustness of geolocation data collected from social networks.

This paper proposes an automated mapping approach between emergency events and locations based on object detection and social networks to address this gap. Furthermore, employing object detection methods on social networks to detect emergency events will construct a reliable, flexible, and speedy approach by utilizing the Snapchat hotspot map as a reliable source to discover the exact location of emergency events. Moreover, the proposed approach aims to yield high accuracy by employing the most current object detection methods to accomplish the objectives of this paper. More particularly, the contributions of this paper can be stated as follow:

- Detect the occurrence of emergency events across time and exact locations by utilizing the Snapchat map.
- Develop a novel ranking and selection model to rank the hotspot locations in order to prioritize which location needs urgent dealing and send it to the decision-makers.
- Develop a flexible, reliable and speedy object detection model to analyze images for the collected data from Snapchat API.
- Propose a combination of one-stage and two-stage object detection models to construct an ensemble model and evaluate the performance of our approach.

III. Related Work

This section outlines the literature review completed throughout this study on two separate fields of study. First, we highlight current research studies that address location detection methods for social networks, concerns, as well as existing and potential solutions. Second, we expand in a limited but adequate manner on current deep learning methods for object detection in emergency events.

A. Location Detection Methods for Social Networks

People usually are inquisitive about the location of catastrophes during or shortly after they occurred. Knowing the precise location is critical for decision-makers to respond promptly and make decisions accordingly [11]. We can get emergency event geolocation information from social network feeds. Location retrieval can be classified into three elements based on the methods used. The first element is the content analysis of words. Second, the analysis of different language modelling. Third, by guessing through social relations [12]. Many studies have been conducted in order to approximate the location of emergency events using material analysis in a specific outer data source depending on geo-related factors.

Singh et al. [13] presented a tweet categorization and location identification technique for spotting tweets from emergency affected seeking assistance as well as their locations. If the location is not stated in the previous tweets, the location was inferred utilising Markov Chain Stochastic approach.

Rodriquez et al. [14]. proposed Using a probabilistic technique that simultaneously predicts location designations and Twitter posts of users, the relationship network is represented graphically. In particular, the authors characterize the system with a Markov random field probability framework, and the training phase is based on a Markov Chain Monte Carlo simulator that estimates the posterior likelihood distribution of the sparse spatial user labels.

Duong-Trung et al. [15] developed a dynamic content-based regression approach using the matrix decomposition method to address the near real-time location estimation challenge. They demonstrated that real-time location estimation is feasible without combining users' tweets.

Gelernter et al. [16] applied Named Entity Recognition technique (NER) to recognize the names of locations in tweets, The findings were contrasted to words manually classified as locations by the researchers and revealed that the enormous number of abbreviations for landmarks is a significant concern for this technique.

Rout et al. [17] employed a Support Vector Machine algorithm to estimate the location of tweets. In addition, their dataset comprises relationship graph features and city metrics like demographic size and the amount of active Twitter accounts. The relationship graph features were adopted because most online social networking relationships begin in real life. They further contend that some elements of the towns are pertinent to the categorization phase.

Laylavi et al. [18] developed a technique for identifying relevant tweets regarding the storm incident. First, they defined event-related word categories utilizing term frequency analysis and the relationship scoring mechanism. Then, an incident-related score was assigned to each tweet. Finally, the suggested system's findings were matched to manually labelled data to measure performance. The recommended approach successfully categorized around eighty-seven incident-related tweets.

Table I illustrates the comparison of the related works in Section III-A. BD is denoted to big data and OD is denoted to object detection.

B. Deep Learning Methods for Object Detection in Emergency Events

The application of deep learning in emergency events detection is becoming more and more extensive. Significant advancements in computer vision have lately been accomplished using these technologies. Multiple detection strategies and techniques have demonstrated superior picture detection efficiency.

TABLE I. A COMPARISON AND SUMMARY OF THE LITERATURE REVIEW

Method	Ref	Dataset	results	BD use	OD use
Markov chain based	[13]	Twitter	87%	No	No
Markov Chain Monte Carlo	[14]	Twitter	78.99%	No	No
Matrix factorization technique	[15]	Twitter	79%	No	No
Named Entity Recognition	[16]	Twitter	90%	No	No
Support Vector Machine	[17]	Twitter	-	No	No
Term frequency analysis	[18]	Twitter	87%	No	No

Ji et al. [19] Used satellite images taken prior to and following the earthquake, a pre-trained VGG model was utilized to detect destroyed houses affected by the Haiti earthquake. The research findings reveal that the fine-tuned VGG model's accuracy level has risen from 83.38 to 85.19. In addition, the destroyed houses identification effect works better, with a production accuracy of 86.31 of earthquake-induced house damage from satellite utilizing a pre-trained CNN classifier, indicating that the CNN approach can successfully identify the characteristics of destroyed structures houses.

Miura et al. [20] employed CNN model to obtain the characteristic of the destroyed buildings' rooftops that wrapped with blue tarpaulin following the earthquakes. Leveraging the enhanced CNN network and satellite photos taken following the two events in Japan.

Qingjie Zhang et al. [21] proposed a novel deep learning method for detecting forest fires. They employed a cascaded method for detecting fire, with the global picture stage examining the whole picture initially and then a local patch encoder identifying the particular position of the determined fire. They presented a baseline for fire recognition, 178 pictures for the train set, and 59 pictures for the test set. They utilize the CIFAR 10 network for the initial phase but reduce the number of outcomes by two and add a drop-out layer to minimize the fitting problem. They employed the Caffe framework's 8-layer AlexNet for the final phase.

Ci et al. [22] proposed a novel CNN architecture integrated with a CNN data harvester, a unique loss function, and an arbitrary regression classifier to measure the severity of houses destruction due to earthquakes utilizing satellite images.

Pi et al. [23] exploited a pre-trained model to train several CNN classifiers based on You-Only-Look-Once (YOLO) in the accident areas to detect intact houses rooftops.

Wei Zhang et al. [24] developed a CNN-CAPSNet—an effective remote imagery categorization system that emphasizes the advantages of the CNN and CAPSNet algorithms. A CNN with partially coupled layers was used as the initial feature map generator. Furthermore, for feature extraction, they employed a pre-trained deep CNN classifier that was entirely trained on the ImageNet dataset. The findings demonstrated that the suggested strategy might outperform state-of-the-art approaches in classifier performance.

Pham et al. [25] proposed a YOLO-fine approach which is an enhanced object detection model from YOLOV3. The model was developed to identify tiny objects with high accuracy and efficiency, allowing real-time applications in realistic circumstances. In addition, they explored its resilience to the appearance of new contexts on the validation set deeper, tackling the critical barrier of domain adaptability in satellite imagery.

Yebes, J et al. [26] acquired and labelled images of real-world situations, and several object detection algorithms were fine-tuned to accomplish their experiment. Their highest results yielded an accuracy rate of 82, employing a hybrid of R-CNN and Resnet101.

Amit et al. [27] suggested a CNN-based paradigm for catastrophe identification in aerial pictures: two fully connected layers and three convolutional and max-pooling layers in the suggested CNN framework. A dataset is gathered for assessment purposes that spans many aerial picture patches from two different natural catastrophes, specifically landslides and flooding. Table II illustrates the comparison of the related works in Section III-B. BD is denoted to big data and OD is denoted to object detection.

IV. BACKGROUND

This section briefly describes important background concepts related to this study, including object detection methods and CNN architectures,

A. Object Detection

Objects detection methods fall into two min categories, one of which is one stage detectors and the other is two stage detectors [6]. Fig. 1 shows the stages of object detection models.

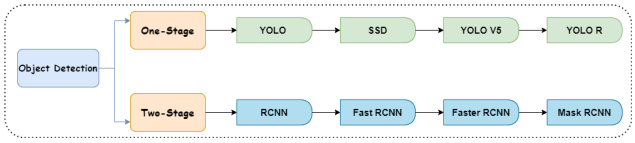


Fig. 1. Stages of object detection.

1) One stage detectors:

a) YOLOV5: is an upgraded version of the YOLOV3. Its implementation is comparable to that of YOLOV4 in that it integrates numerous approaches such as data augmentation and modifications to activation functions with data preprocessing

TABLE II. A COMPARISON AND SUMMARY OF THE DEEP LEARNING METHODS FOR OBJECT DETECTION IN EMERGENCY EVENTS

Method	Ref	Dataset	results	BD use	OD use
CNN	[19]	satellite imagery	87.6%	No	No
CNN	[20]	aerial images	95%	No	No
CNN	[21]	various online resources	90%	No	No
CNN	[22]	multiple aerial imagery	93.95%	No	No
CNN YOLO	[23]	in-house aerial video imagery	80.69%	yes	No
CNN-CapsNet	[24]	UC Merced Land-Use ,AID , NWPU-RESISC45	98.81%	No	No
CNN YOLO-fine	[25]	VEDAI ,MUNICH,XVIEW	84.34%	yes	No
CNN Faster R-CNN- SDD	[26]	images with pothole annotations	82%	yes	No
CNN	[27]	satellite imagery	80%	No	No

to the YOLO structure. Yet, YOLOv5 differs from YOLOV4 in the base; rather than Darknet, it uses CSPDarknet53 as its base. This base addresses the redundant gradient data in a massive backbone network. It incorporates gradient modification into the feature map, decreasing inference latency, boosting precision, and lowering model complexity by minimizing the parameters. In addition, it aggregates pictures for training and uses self-adversarial training (SAT) to guarantee faster prediction [28].

b) YOLOR: You Only Learn One Representative: The YOLOR model, which has an integrated network design, is intended to carry out several tasks at the same time. YOLOR is an upgraded version of the YOLO model that takes advantage of the data in the lower tiers of the CNN layers, specifically the attribute data. Furthermore, because it provides concurrent recognition, the YOLOR method is more effective than the existing YOLO methods [29].

YOLOR simultaneously utilizes implicit and explicit information to model training to acquire generalized representations and execute multiple jobs using these generalized representations. The implicit awareness recognizes deep layer characteristics, and annotated data is used to gain explicit information [29]. YOLOR has different versions with distinct specifications, so YOLOR-P6 is considered for the study.

2) Two stage detectors:

a) Faster R-CNN: Faster R-CNN is upgraded version of Fast R-CNN to overcome the limitations of Fast R-CNN. The main drawback of Fast R-CNN is that it leverages particular inquiry to generate Regions-Of-Interest (ROI), which is slower and requires the exact period to execute as the recognition system. Therefore, faster R-CNN is substituted by a unique RPN (region proposal network), a fully convolution network that can anticipate region suggestions using a broad variety of sizes and aspect ratios. Furthermore, since it combines full-image convolution characteristics and a similar set of convolution layers with the overall classifier, RPN speeds up the production of region suggestions [30].

b) Mask R-CNN: Mask R-CNN adds mask branch output on top of the prior Faster R-CNN base. The core layer gathers features at the beginning; then, the proposals are anticipated and optimized to infer the bounding boxes for object recognition and build segmentation masks. The mask offers pixel-level feature extraction for every potential object by performing instance segmentation [31]. In addition, it incorporates enhancements in Faster R-CNN and FCN (Fully Connected Network), which led to its popularity as a two-stage object detection model compared to the other models, as it offers both bounding box and segmentation [32].

B. Convolutional Neural Networks (CNN) Architecture

Convolutional neural networks are the most extensively used deep learning algorithms and the most popular type of neural network. It is a multi-layer neural network (NN) design comprising a convolutional layer(s) followed by a fully connected (FC) layer (s). The principal use of CNN is in databases, where the number of nodes and parameters is enormous [33].

1) Convolutional layer: This layer is the foundation of CNN models, defining linked inputs' outcomes. Such an outcome is accomplished by convolving kernels across the datasets' size and shape, determining the feed's and filter's dot product, and constructing a two-dimensional activation map with that filter. CNN efficiently understands which filters to activate when a certain kind of attribute is noticed at a specific spatial point in the input [34].

2) Non-linearity layer: Nonlinear functions are essential and retain a degree greater than one; when displayed, they exhibit a curve. The primary goal of this layer is to convert the incoming signal to the outgoing signal, which will be employed as an input in the subsequent layer. Non-linearity layers include sigmoid or logistic, Tanh, ReLU, PReLU, and ELU [35].

3) Pooling layer: CNNs can comprise internal or external subsampling layers that combine the results of one layer's neurons into an individual neuron in the subsequent layer. Its primary function is to reduce the spatial size of the depiction to decrease the size of the parameters and computations in the framework. It prevents overfitting and accelerates the computation. The max-pooling layer is the most frequent type of pooling layer [35].

4) Fully connected layer: FC layers are conventional deep NN layers that aim to create forecasts from activation for regression and classification. This layer obtains entire links to each activation in the preceding layer, and the activation may be determined by matrix multiplication coupled with a bias of sets [33].

V. Research Methodology

In this section, elucidation of the critical components of this article is required by outlining the steps needed. First, hotspot location in Snapchat map will be pinpointed. Second, collect the dataset from the Snapchat map API in the exact locations of the emergency events that were identified earlier without any prior knowledge. Third, data pre-processing and data augmentations methods are employed at this stage. Fourth, object detection models for emergency event detection are proposed. Fifth, we deploy an ensemble learning based for emergency events to conduct a performance evaluation of the proposed model. Finally, locations ranking and model selection is proposed to prioritize the hotspot locations. The steps of the proposed model is shown in Fig. 2.

A. Pinpoint Hotspot Locations in Snapchat Map

Pinpointing the precise location of emergency events is very important. Therefore, by leveraging the Snapchat map we can find the exact location of any emergency event. The snaps posted on the Snapchat Interactive Map is automatically geotagged and represent a particular hotspot location which can be a possible emergency event location [36]

B. Data Collection

Accordingly, after identifying the hotspot locations, a preliminary ranking is applied based on the highest number of snaps in each location to prioritize the locations. Then we will use the same method we used in our previous works wrapper for the Snapchat Map's internal API in order to construct a Node.js function that scans for snaps uploaded at precise coordinates (hotspot zones) [37] which will collect all the snaps shared at the exact location and marked as either emergency event-content or non-emergency event-content. The acquired snaps might be image or video, thus,the images they will be analyzed by utilizing object detection techniques.

C. Data Pre-processing

1) Standardize images: One fundamental limitation in several object detection methods is the requirement to scale the datasets to a consistent size. This means that before we feed the images to the training model, they need to be preprocessed and resized to have equal dimensions [38]. As a result, the collected images are resized to a length of 640*640.

2) Data augmentation: Another frequent preprocessing strategy is to augment the current dataset with different replicas of the original images. It is performed to expand the dataset and introduce the neural networks to a broad range of image permutations [38]. In addition, it will increase the likelihood that the model will detect objects in any configuration.

Data augmentation may successfully prevent fitting problems throughout the advanced training phase and can tremendously enhance the clarity of the data [39]. Several data augmentation methods were employed, including vertical and horizontal flipping, rotating to a specific angle (less than 20°), and raising or reducing brightness.

D. Applying Object Detection Models

After collecting the dataset from the Snapchat map API, we will apply the most common object detection algorithms which are : You Only Look Once V5 (YOLO v5), You Only Learn One Representation (YOLOR), Mask RCNN and Faster R-CNN. We will apply these algorithms to every hotspot location that have been identified earlier to detect emergency events. Then, we will evaluate the performance of these algorithms using evaluation metrics. Table III shows the hyper-parameters of the models. Algorithm 1 depicts training and evaluation steps.

E. Proposed Voting Ensemble Object Detection Model

Among the four object detection models mentioned above, we will propose to use of the ensemble learning. The precision of the object detection model may be boosted by merging several models into an ensemble model which can be effectively applied for emergency event detection. One of the primary factors driving the prevalence of ensemble learning is its capability to minimize the variance and bias of deep learning models [40].

Numerous object detection models have been introduced to ensemble learning for different scenarios; however, using an ensemble learning approach with object detection models in emergency event detection is still appealing. Both one-stage and two-stage models performed well for object detection, as discussed in the related works. Therefore, four baseline object detection models were chosen for the proposed ensemble approach in this work. We applied a combination of single-stage and two-stage models to conduct our proposed ensemble approach. The proposed voting ensemble object detection model is shown in Fig. 3.

F. Location Ranking and Selection Model

With the continuous search of hotspot zones in the Snapchat map, multiple locations will be found and processed. In this step we will rank these hotspot locations to prioritize them in terms of ranking criteria. Algorithm 2 explains the ranking model. This model consists of four criteria metrics:

a) Amount of snaps in the location: It counts how many snaps have been posted in this particular hotspot location. The overall score for this criteria is one point. Since we only have two locations of emergency events, the highest will get one point and the least will get zero.

b) Amount of snaps related to the emergency events: It counts only the snaps that are related to the emergency event in this particular hotspot location. This step is important to get rid of unrelated snaps. The overall score for this criteria is two points. Since we only have two locations of emergency events, the highest will get two points and the least will get zero.

c) mean Average Precision (mAP) score: A higher mAP score shows improved object detecting method performance. Since we have applied four object detection models to the dataset, we will get four different mAP score for each model. Therefore, we will calculate the mean of this score which the sum of all scores divided by the total number of models. The overall score for this criteria is three points. Since we only have two locations of emergency events, the highest

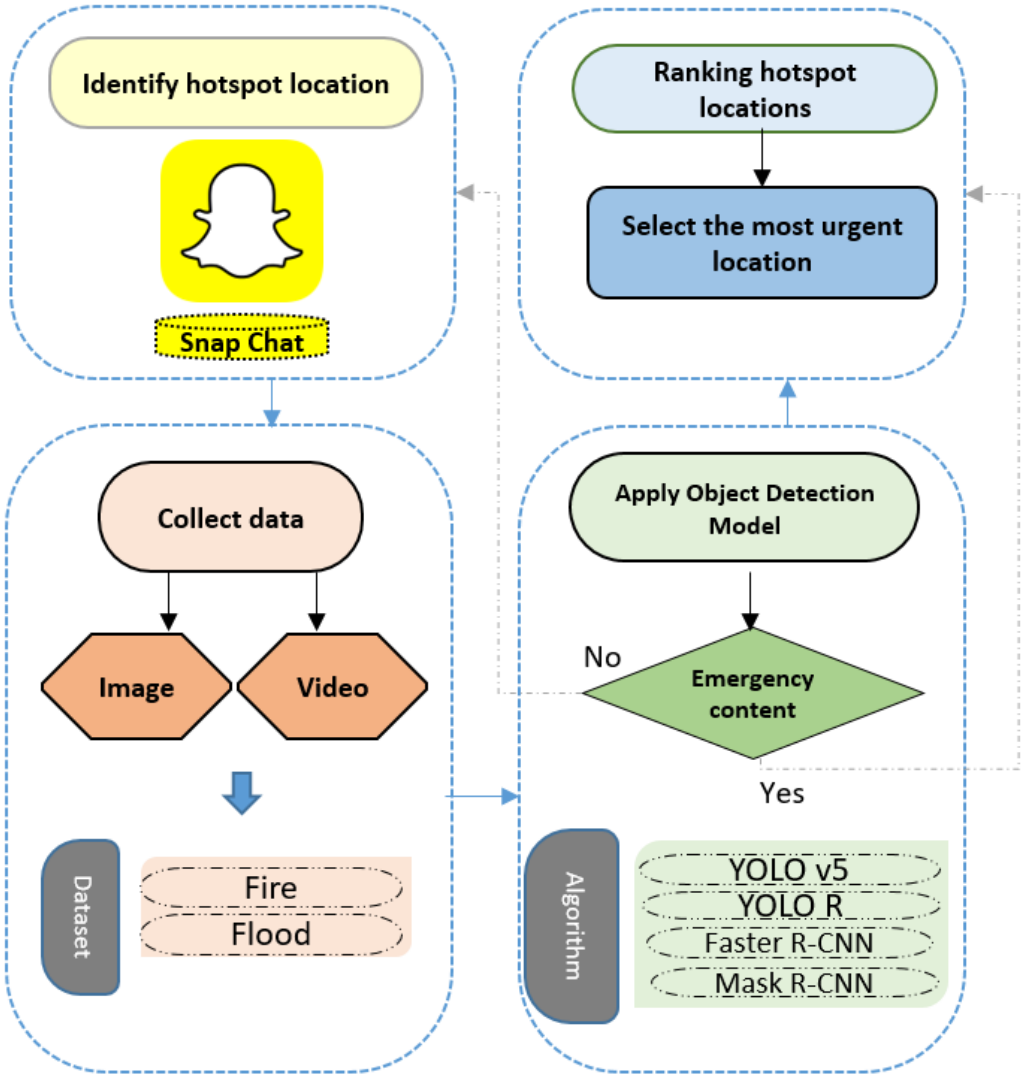


Fig. 2. The steps of the proposed model.

will get three points and the least will get zero. It can be calculated as:

$$Mean = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n} \quad (1)$$

d) Throughput score (TP score): It refers to the quantity of outcomes delivered in a particular amount of time. After we evaluate the models performance using Spark, the model that will get the best performance for both case studies in regards to time processing will be selected. Then by matching this model to the corresponding case study (location), the corresponding location will get selected. The overall score for this criteria is four points. Since we only have two locations of emergency events, the lowest processing time model will get four points and the rest will get zero.

After that, we will calculate the summation of all scores, the location with a score of equal or greater than five will be sent in cluster A and will be selected as the most urgent location.

$$\text{if } Location(L) \geq 5 \text{ then } put\ it\ in\ Cluster(A) \quad (2)$$

If the location gets a score of equal or smaller than three, it will be sent to cluster B.

$$\text{if } Location(L) \geq 3 \ < 5 \text{ then } put\ it\ in\ Cluster(B) \quad (3)$$

Also, If a location get a score less than two, it will be sent to cluster C and will be disregarded.

$$\text{if } Location(L) \leqslant 2 \text{ then } put\ it\ in\ Cluster(C) \quad (4)$$

The ranking model is shown in Fig. 4.

TABLE III. HYPER-PARAMETERS OF THE MODELS

Models	Hyper-Parameters
YOLOV5	lr0=0.01, lrf=0.01, momentum=0.937, weight_decay=0.0005, warmup_epochs=3.0, warmup_momentum=0.8, epochs=100, batch_size=16
YOLOR	'lr0': 0.01, 'lrf': 0.2, 'momentum': 0.937, 'weight_decay': 0.0005, epochs=100, 'warmup_epochs': 3.0, 'warmup_momentum': 0.8, 'warmup_bias_lr': 0.1, batch_size=16
Faster R-CNN	lr0=0.01, lrf=0.01, momentum=0.937, weight_decay=0.0005, warmup_epochs=3.0, warmup_momentum=0.8, epochs=50, batch_size = 8
Mask R-CNN	lr0=0.01, lrf=0.01, momentum=0.937, weight_decay=0.0005, warmup_epochs=3.0, warmup_momentum=0.8, epochs=50, batch_size = 8

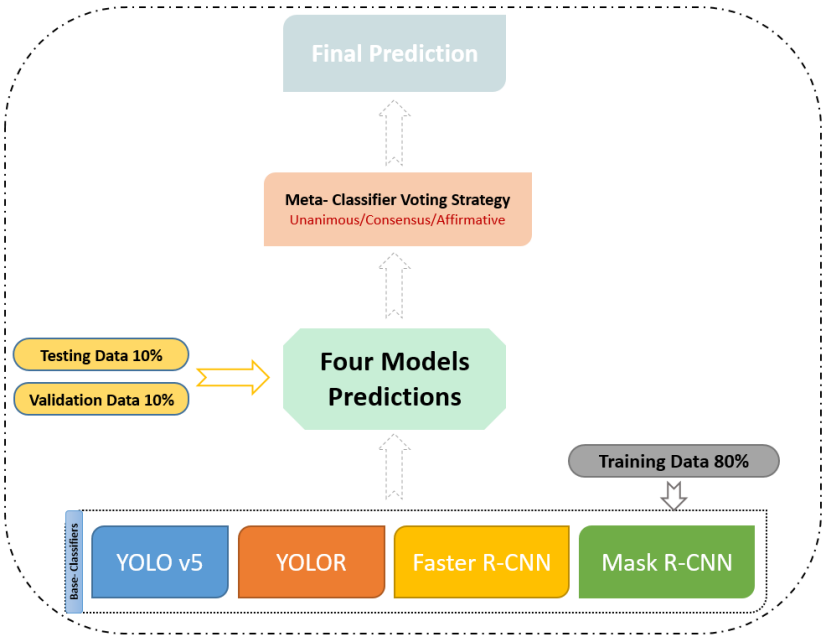


Fig. 3. Proposed voting ensemble object detection model.

Algorithm 1 Location Ranking and Selection

Input: *LocationList*; *RankingCriteria*; *ModelsList*;
Output:Ranked Locations

spark ← SparkConnector()
spark.load(Location List)
spark.load(Ranking Criteria)
for `each location in (Location List)` **do**
    Rank_score ← Amount of snaps in the location = 1 point
    Rank_score ← Amount of emergency related content = 2 points
    Rank_score ← Mean Average Precision(MAP) score = 3 points
    Rank_score ← Throughput score = 4 points
    Calculate the sum of Rank_score
    if the sum of Rank score ≥ 5 **then**
        Put this location in cluster (A)
    if the sum of Rank score ≥ 3 AND < 5 **then**
        Put this location in cluster (*B*)
    if the sum of Rank score ⩽ 2 **then**
        Put this location in cluster (*C*)
return *Ranked Locations* ← Location List

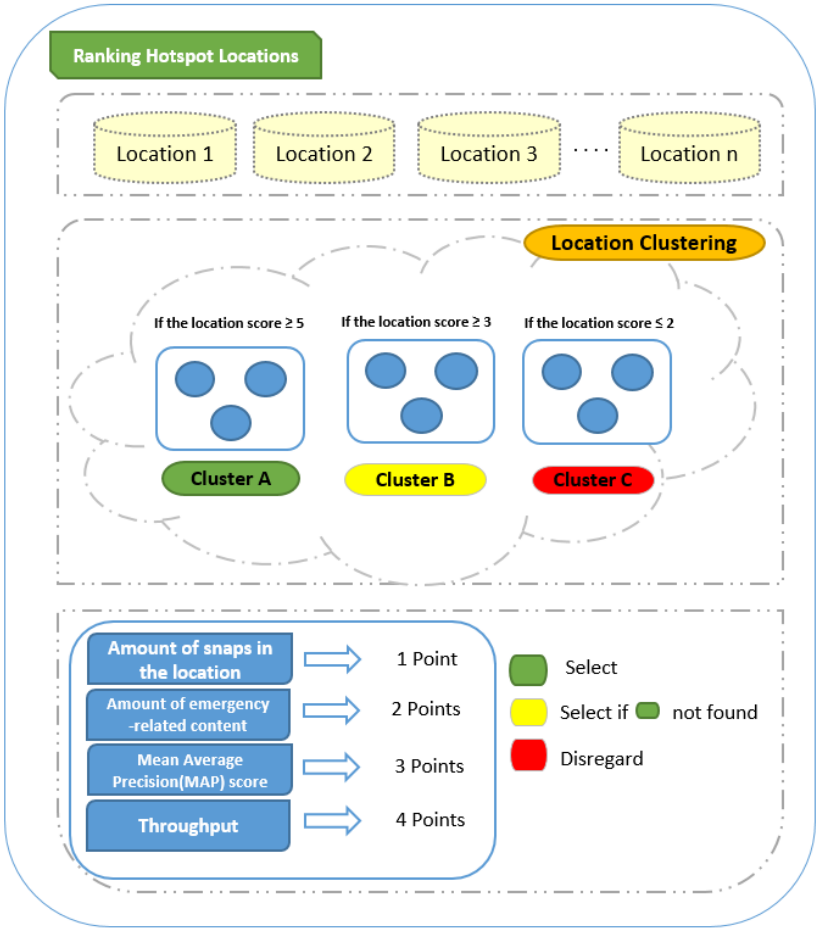


Fig. 4. Proposed location ranking and selection model.

VI. Evaluation Approach

This section describes the specified case studies in depth, including the reasons for their selection and experimental environment construction, evaluation measures, performance analysis for our suggested approach, and discussion of study outcomes.

A. Case Study

During the scan for the hotspot zones in the Snapchat map utilizing the constructed Node.js wrapper, we identified more than five hotspots location between March 2022 and May 2022 .However, according to the preliminary Snapchat data collection ranking criterion, the most hotspot location that contains the most amount of snaps at that moment were India building fire and South Africa flooding. Therefore, after identifying the exact location of these emergency events, We collected the dataset of each incident using Snapchat API. The total number of the collected images and videos are stated in Table IV.

1) India building fire: At least 27 individuals died, and 24 people were injured in a devastating fire in the Indian capital of New Delhi on Friday (May 13), according to rescue teams. The big fire burst in the mid-evening at a four-story residential building in west Delhi, although the cause was not intuitively known. Twenty-seven burnt remains were found from the building, also several of the residents jumped from the building during the fire and were hospitalized [41]. Fig. 5 shows the location of building on fire in India. Fig. 6 shows the detection results of building on fire in India.

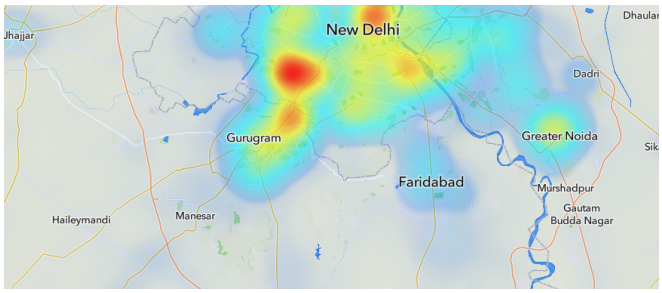


Fig. 5. Location of building on fire in India.

2) South Africa flooding: Heavy rainfall caused catastrophic floods and landslides in southern and south-eastern South Africa, notably in the provinces of KwaZulu-Natal and the Eastern Cape, from April 11 to 13. According to national authorities, 443 people have died in KwaZulu-Natal, while over 40,000 remain missing. Also, over 40,000 people have been evacuated, while 4,000 homes have been demolished or

Algorithm 2 Training and Evaluation

Input: *LocationList*; *Datasetslist*; *ModelsList*;
Output:Trained models list

```
spark ← Spark Connector()
spark.load(Location List)
spark.load(Models List)
spark.load(Datasets List)
for each Dataset in (Datasets List) do
    Trained_models ← Train YOLOv5
    Trained_models ← Train YOLOR
    Trained_models ← Train Faster R-CNN
    Trained_models ← Train Mask R-CNN
    for each model in (Trained_models) do
        Evaluation_score ← Calculate Recall
        Evaluation_score ← Calculate Precision
        Evaluation_score ← Calculate IOU
        Evaluation_score ← Calculate AP
        Evaluation_score ← Calculate mAP
        Evaluation_score ← Calculate Throughput
    return Trained models list ← Trained_models
```

TABLE IV. AMOUNT OF DATASET

Location	All data	Data related to emergency events
India fire	780 images	546
South Africa flooding	759 images	524

Fig. 6. Detection results of building on fire in India.

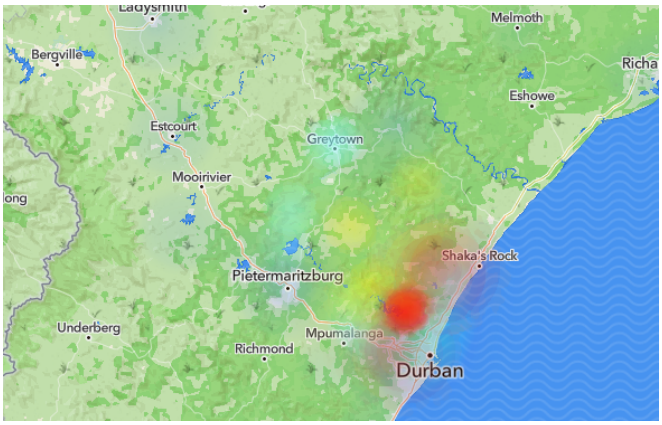


Fig. 7. Location of flooding in South Africa model.

severely damaged, mainly in Durban and its nearby regions. Due to the severe flooding, a National State of Disaster has been announced. Emergency crews have been dispatched to the areas impacted to give immediate aid to people affected by disasters [42]. Fig. 7 shows the location of flooding in South Africa. Fig. 8 shows the detection results of flooding in South Africa.

B. Experimental Setup

We designed our models with different python packages. The core libraries that we used were Pytorch, numpy, opencv, matplotlib, Scikit-Learn, Keras, and native TensorFlow. The hardware used for this training and testing of models was Nvidia Tesla P100 (16 GB) offered by Google Colab Pro. High Ram offered by Google Colab Pro was used to increase the IO speed from google drive datasets.

To evaluate the performance of our approach, we first Identified the hotspot locations from the Snapchat map using the developed Node.js Google function wrapper in our previous paper. Then we collected the dataset from the Snapchat map API in the exact locations of the emergency events that were identified earlier without any prior knowledge.

The total amount of dataset collected from both locations are shown in Table IV. In our first case study (the location of South Africa flooding) we were able to collect 780 images. However after we manually analyzed the dataset, we found out that there were some images that were unrelated to the case

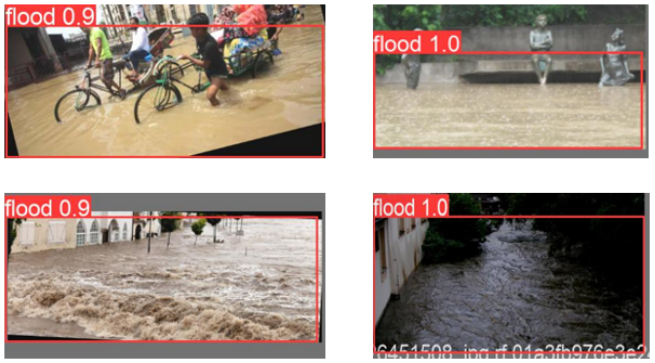


Fig. 8. Detection results of flooding in South Africa.

study and we got rid of them. The total amount of dataset in that location dropped to 546 images. Also, we were able to collect 759 images for the second case study (the location of flooding in South Africa). However, there were some images that unrelated to the case study which we got rid of them. After deleting the noise images, the total number of images became 524 images.

After cleaning our dataset, we used Roboflow online annotation tool that manually labels our dataset, generates boundary boxes, and classifies them. To guarantee that the dataset is spread evenly while considering the correlation between labels and data, the dataset is arbitrarily separated into three sets: training, validation, and test. according to the ratio of %80, %10, and %10. We created three folders to tore the dataset accordingly: training; testing; and validation [43]. Training data are included in the data and enable the deep learning model in producing a prediction. Validation data indicate if the model is competent of accurately detecting new instances or not, and they usually comprise pictures that the model employs to assess and monitor its learning. The images in the testing folder can be used to predict model correctness. The final dataset is saved in the XML dataset format to maintain the same experimental configuration.

Then, we applied the pre-processing steps to the collected images after getting rid of images that were unrelated to the case studies and annotate the related images. After that, due to the fact that object detection models require as much data as possible, our dataset were very low and we should find ways to enlarge them. Therefore, we applied some data augmentation methods to increase the dataset and to ensure the effectiveness of the training. Furthermore, due to the training dataset we used in our research is relatively small even after employing data augmentation methods,We employed transfer learning to overcome this challenge by training our models on the MS-COCO dataset and obtaining a pre-training model as MS-COCO is the baseline standard for validating and testing object detection methods. The pre-trained weights were loaded, and the training began with them as a starting point.

The pre-training algorithm can retrieve the generic characteristics of all objects from the standard datasets. We may employ the matching architecture and weights by leveraging the pre-training framework. Despite the limited dataset, the model may modify the parameters to an optimal form based on the pre-training model. IOU (Intersection over Union) threshold is set to 0.65 in the experiment. After that, we built the object detection models using the trained dataset. Finally, we integrated the trained models into the Apache Spark streaming and used the test dataset to evaluate the performance.

C. Performance Evaluation

During the training phase, the evaluation parameters play an important role in achieving the targeted object detection accuracy. Furthermore, gathering appropriate assessment parameters is a critical component in the differentiation and development of the perfect model. After applying the object detection models to our case studies, We need to test the suggested model's accuracy using several performance assessment indicators such as [44]:

1) Precision (PR): It is utilized to subtract the number of accurately predicted positive patterns from the total number of expected positive patterns in a positive class [45]. Precision can be calculated using the following equation.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

2) Recall (RE): It is used to determine the percentage of correctly categorized positive patterns [45]. Recall can be calculated using the following equation:

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

True Positive (TP): It corresponds to the quantity of cases correctly recognized by the classifier [46].

False Positive (FP) It represents the amount of negative incidents that were incorrectly classified as positive cases [46].

False Negative (FN): It relates to the quantity of positive incidents that were incorrectly labelled as negative cases. [46].

True Negative (TN): The number of negative cases successfully categorised by the model is denoted by the true negative values [46].

3) F-Score (FS): Also recognized as the F1-score, this is a statistic used to assess data correctness. In addition, it's employed to investigate binary classification methods that classify data as "positive" or "negative". [45]. The F1-score can be calculated using the following equation.

$$F1 = 2 \times \frac{(precision \times recall)}{precision + recall} \tag{7}$$

4) Intersection over Union (IOU): IoU is calculated by dividing the overlapping region between detection and ground truth by their union region [47]. Fig. 9 illustrates the concept of IoU.

When IoU is 100%, projection boxes and ground truth boundaries perfectly coincide, and the prediction is the highest. Despite 100% IoU is practically impractical to acquire (because of the constraints of present Convolutional networks),

an IoU score of 50% to 90% is widely employed in numerous computer vision tasks [48].

In this work, a detection is deemed successful if IoU ≥ 65%.

IoU can be calculated using the following equation:

$$IoU(A,B) = \frac{Area\ of\ Overlap}{Area\ of\ Union} = \frac{A \cap B}{A \cup B} \quad (8)$$

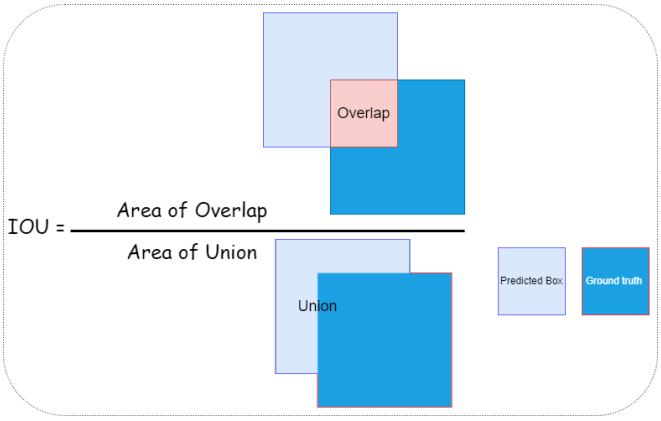


Fig. 9. Intersection over Union (IOU).

5) Average Precision (AP): The Average Precision (AP) is intended to describe the Precision-Recall Curve by summing accuracy over all recall scores ranging from 0 to 1. It's the region underneath the Precision-Recall curve [49]. (AP) can be calculated using the following equation.

$$\text{AP} = \sum_{n} (R_n - R_{n-1}) P_n \quad (9)$$

6) Mean Average Precision (mAP): It is the average of all specified categories of precision. A more excellent mAP implies that an object detection technique performs well in terms of precision and resilience [49]. (mAP) can be calculated using the following equation:

$$\text{mAP} = \frac{1}{N} \sum AP \quad (10)$$

7) Throughput (TP): It refers to the quantity of outcomes delivered in a particular amount of time.

D. Results

We conducted the experiments locally on the environment set up described in the previous subsection to evaluate our approach.

As shown in Table V, which contains six evaluation metrics named Precision(PR), Recall(RE), IOU, F-Score, mAP and Throughput(TP) for the flood case study after applying data augmentation.

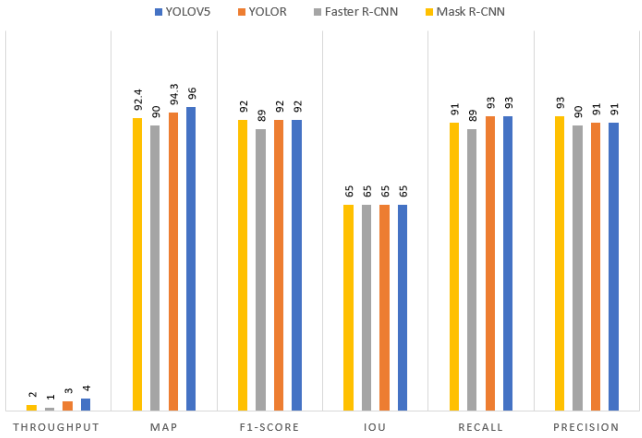


Fig. 10. Results of flood.

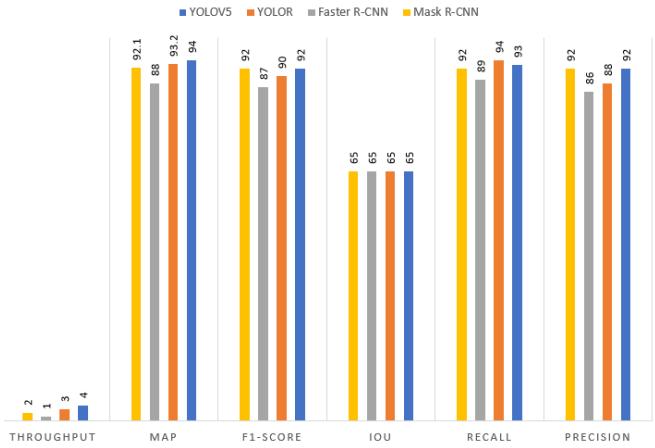


Fig. 11. Results of fire.

Table V illustrates that the results of the one stage models Yolov5 and YoloR are the same for Precision, Recall and F-Measure. However, they vary in terms of mAP score, Yolov5 achieved a very high accuracy of %96 while YoloR achieved %94.3 score. Moreover, the difference in the performance of the two stage models Faster R-CNN and Mask R-CNN is very minimal. However, Mask R-CNN outperform Faster R-CNN in all metrics. Faster R-CNN yields %90 of mAP score while Mask R-CNN yields %92.4.

Table VI shows the performance of the models for the flood case study before applying data augmentation. It can be noted that all the model didn't perform well before applying data augmentation.

Table VII shows the models' performance for the fire case study after applying data augmentation. It can be noted that Yolov5 outperforms all the models in all evaluation metrics except for Recall where YoloR achieved %94 while Yolov5 achieved %93. However, the difference between them is negligible in regards to mAP score, Yolov5 yields %94 while YoloR yields %93.2. On the other hand, the difference between the two stage model Faster R-CNN and Mask R-CNN is noticeable. Mask R-CNN surpass Faster R-CNN in all metrics. Also, Mask R-CNN yields %92.1 of mAP score

TABLE V. RESULTS OF MODELS PERFORMANCE FOR FLOOD CASE STUDY WITH DATA AUGMENTATION

Model	PR	RE	IOU	F-Score	mAP	TP
Yolov5	0.91	0.93	0.65	0.92	0.96	4
YOLOR	0.91	0.93	0.65	0.92	0.943	3
Faster R-CNN	0.90	0.89	0.65	0.89	0.90	1
Mask R-CNN	0.93	0.91	0.65	0.92	0.924	2

TABLE VI. RESULTS OF MODELS PERFORMANCE FOR FLOOD CASE STUDY WITHOUT DATA AUGMENTATION

Model	PR	RE	IOU	F-Score	mAP
Yolov5	0.75	0.77	0.65	0.76	0.82
YOLOR	0.61	0.81	0.65	0.70	0.80
Faster R-CNN	0.66	0.69	0.65	0.67	0.70
Mask R-CNN	0.70	0.72	0.65	0.71	0.79

TABLE VII. RESULTS OF MODELS PERFORMANCE FOR FIRE CASE STUDY WITH DATA AUGMENTATION

Model	PR	RE	IOU	F-Score	mAP	TP
Yolov5	0.92	0.93	0.65	0.92	0.94	4
YOLOR	0.88	0.94	0.65	0.90	0.932	3
Faster R-CNN	0.86	0.89	0.65	0.87	0.88	1
Mask R-CNN	0.92	0.92	0.65	0.92	0.921	2

TABLE VIII. RESULTS OF MODELS PERFORMANCE FOR FIRE CASE STUDY WITHOUT DATA AUGMENTATION

Model	Precision	Recall	IOU	F-Score	mAP
Yolov5	0.74	0.75	0.65	0.74	0.80
YOLOR	0.60	0.80	0.65	0.69	0.78
Faster R-CNN	0.67	0.68	0.65	0.67	0.70
Mask R-CNN	0.68	0.70	0.65	0.69	0.73

which is very minimal to YoloR.

Moreover, as it can be noted from Table VIII, it shows the performance of the models for the fire case study before applying data augmentation. It's obvious that all the model didn't perform well before applying data augmentation.

Fig. 10 shows the performance evaluation of each model for flood dataset model separately while Fig. 11 illustrates the performance evaluation of each model for flood dataset model separately.

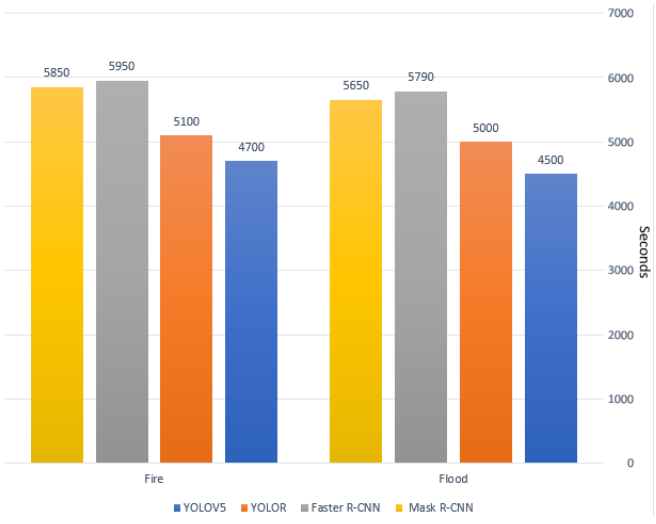


Fig. 12. Models performance using Spark.

Regrading the Throughput score, we applied and run a ranking algorithm to get the scores. After we measured the performance of the models using the test dataset on Spark, we sort out the score results of the models from Spark in ascending order in regards to processing time. Then, the first model (with the least processing time) will get a score of four points which is the highest and the second will get three and the third will get two and the last one will get only one point.

Therefore, Fig. 12 illustrates that Yolov5 model outperform the rest of the models in terms of throughput in both case studies and score four points because it needed less time to process the dataset.

Fig. 12 shows the performance of all the model using Spark.

Table IX shows the results of our location ranking and selection model.

It can be noted that the second case study of South Africa flooding (Location 2) scored seven points in contrast to the first case study building on fire in India (Location 1) which scored only three points. Therefore, the second location was selected as the most urgent location. All Snaps is denoted to all the collected snaps and its score, Emergency is denoted to all the snaps related to emergency cases.

Fig. 13 and 14 depict the the evaluation results of Yolov5 for flood dataset and fire dataset, respectively. In addition, Fig. 15 represents the accuracy score of Mask R-CNN for both datasets. While Fig. 16 shows the accuracy result of YoloR for flood and fire datasets.

TABLE IX. RESULTS OF LOCATION RANKING AND SELECTION ALGORITHM

Location	All Snaps/ Score	Emergency/ Score	mAP/ Score	TP Score	Sum
L1(Fire)	780/ 1	546/ 2	0.918/ 0	0	3
L2(Flood)	759/ 0	524/ 0	0.931/ 3	4	7

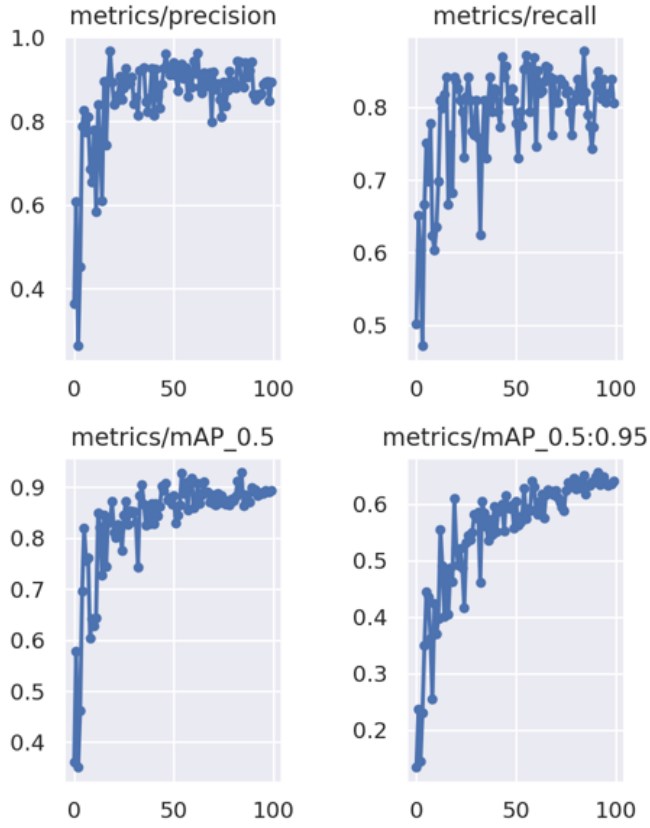


Fig. 13. Yolov5 flood results.

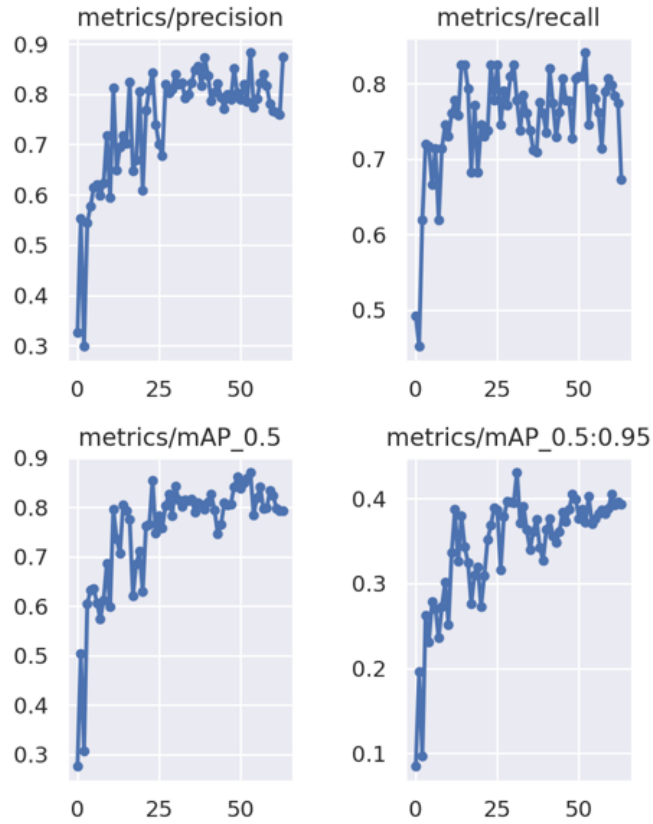


Fig. 14. Yolov5 fire results.

VII. DISCUSSION

The high prevalence of cellphones and social networking platforms such as Snapchat are obviously dissolving traditional barriers between information providers and end-users. It is certainly relevant in emergency events, as individuals on the site produce and exchange real-time information about the event.

However, notwithstanding their demonstrated significance, obtaining event-related information from real-time streams of vast numbers of snaps is a significant challenge.

To address this gap, this paper proposes an automated mapping approach of emergency events and locations based on object detection and social networks. Furthermore, employing object detection methods on social networks to detect emergency events will construct reliable, flexible and speedy approach by utilizing the Snapchat hotspot map as a reliable source to discover the exact location of emergency events.

Moreover, the proposed approach aims to yields high accuracy by employing the state of the arts object detectors to achieve the objectives of this paper. Results show that the proposed approach achieved a very high accuracy of 96% for flood dataset and 94% for fire dataset.

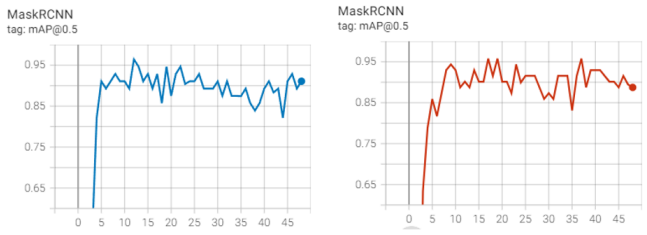


Fig. 15. MaskRCNN.

Among the evaluated models, Yolov5 exhibited the highest performance, proving to be a reliable option for emergency event detection, aligning with previous research indicating that Yolov5 achieves superior accuracy in real-time object detection compared to traditional CNN-based models [50].

Mask R-CNN also demonstrated promising results, particu-

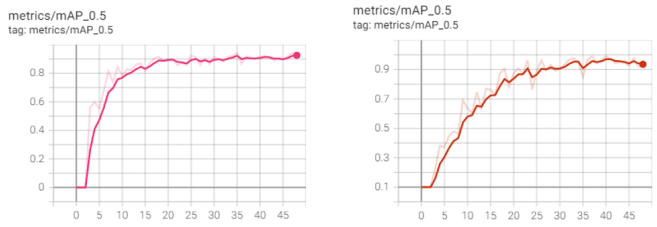


Fig. 16. YOLOR.

larly in terms of learning stability, supporting findings from He et al. [51], which highlight its effectiveness in detecting objects with high precision. However, YoloR struggled with detecting fragmented objects, and Faster R-CNN was the least effective model in this study, consistent with prior work noting Faster R-CNN's computational inefficiency for real-time detection [52] .

VIII. Conclusion

Based on experimentation and testing, the YoloR model did not perform well without applying data augmentation techniques on the fire and flood datasets because the dataset was too small. Without data augmentation, the recall was as high as 0.8, but the precision was merely 0.6.

The model kept predicting numerous small and large boxes, and in doing so, it got many erroneous boxes alongside the ground truth; as a result, it gave many incorrect outputs. Its precision was very low, but since a few of those boxes were the actual boxes, the recall was 0.8. The recall refers to all the actual boxes being detected (irrespective of how many erroneous ones you get in addition). We applied brightness data augmentation where the image was made brighter by 20% or darker by 20% which makes it more stable in different lighting conditions.

By applying data augmentation, the precision increased to 0.88 while recall increased to 0.94, with a total mAP of 0.932 on the fire dataset. A similar augmentation technique was applied to the flood dataset. We noticed that YoloR works better for continuous objects; when we have objects that break or distance between each instance, it does not work very well. Because instead of detecting the whole object, it detects numerous smaller objects. Technically it is correct, but in terms of the mAP score, this could consider a downside.

Moreover, Mask R-CNN performed very well, it was a stable model, and the predictions were similar to other models in terms of accuracy. Mask R-CNN's fascinating side is that its learning curve is better than the other models. If we had more data, Mask R-CNN would do better than other models.

Yolov5 yielded the highest accuracy for both datasets compared to the rest of the models. Even before applying the data augmentation techniques, both datasets' accuracy scores were acceptable for Yolov5. Yolov5 is available in various variants: Yolov5x, Yolov5m, Yolov5l, and Yolov5x, each with deeper layers than the others. We used Yolov5l for both flood and fire datasets in this work.

Faster R-CNN was the least model in terms of performance compared to other models. Before applying data augmentation techniques, its performance was almost the same, and the difference was negligible. However, after applying data augmentation techniques, the performance of Faster R-CNN for the flood dataset was better than that of the fire dataset; it yielded 0.90 and 0.88, respectively.

Future work expected that the performance would improve if more datasets and additional object detection models were applied to the approach. Furthermore, our system shows promising results when combining the Snapchat hotspot map and computer vision approach to detect the location of emergency events.

Authors' Contributions

Alfalqi Khalid: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing original draft preparation, writing review and editing, visualization, Bellaiche Martine: discussion, review, supervision. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Competing Interests

The authors declare no conflict of interest.

References

[1] Z. Hu, G. Wu, H. Wu, and L. Zhang, "Cross-sectoral preparedness and mitigation for networked typhoon disasters with cascading effects," *Urban Climate*, vol. 42, p. 101140, 2022.

[2] M. T. Chaudhary and A. Piracha, "Natural disasters—origins, impacts, management," *Encyclopedia*, vol. 1, no. 4, pp. 1101–1131, 2021.

[3] H. J. Caldera and S. Wirasinghe, "A universal severity classification for natural disasters," *Natural hazards*, vol. 111, no. 2, pp. 1533–1573, 2022.

[4] L. Sun and F. You, "Machine learning and data-driven techniques for the control of smart power generation systems: An uncertainty handling perspective," *Engineering*, vol. 7, no. 9, pp. 1239–1247, 2021.

[5] M. Suriya, V. Chandran, and M. Sumithra, "Enhanced deep convolutional neural network for malarial parasite classification," *International Journal of Computers and Applications*, pp. 1–10, 2019.

[6] J. Zhou, X. Tan, Z. Shao, and L. Ma, "Fvnet: 3d front-view proposal generation for real-time object detection from point clouds," in *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2019, pp. 1–8.

[7] Y. Song, L. Gao, X. Li, and W. Shen, "A novel robotic grasp detection method based on region proposal networks," *Robotics and Computer-Integrated Manufacturing*, vol. 65, p. 101963, 2020.

[8] P. Mittal, R. Singh, and A. Sharma, "Deep learning-based object detection in low-altitude uav datasets: A survey," *Image and Vision Computing*, vol. 104, p. 104046, 2020.

[9] S. Khatoon, A. Asif, M. M. Hasan, and M. Alshamari, "Social media-based intelligence for disaster response and management in smart cities," in *Artificial Intelligence, Machine Learning, and Optimization Tools for Smart Cities*. Springer, 2022, pp. 211–235.

[10] A. Kumar and J. P. Singh, "Location reference identification from tweets during emergencies: A deep learning approach," *International journal of disaster risk reduction*, vol. 33, pp. 365–375, 2019.

[11] A. Fokaefs and K. Sapountzaki, "Crisis communication after earthquakes in greece and japan: Effects on seismic disaster management," *Sustainability*, vol. 13, no. 16, p. 9257, 2021.

[12] C. Xu, A. S. Ding, and K. Zhao, "A novel poi recommendation method based on trust relationship and spatial–temporal factors," *Electronic Commerce Research and Applications*, vol. 48, p. 101060, 2021.

[13] J. P. Singh, Y. K. Dwivedi, N. P. Rana, A. Kumar, and K. K. Kapoor, "Event classification and location prediction from tweets during disasters," *Annals of Operations Research*, vol. 283, no. 1, pp. 737–757, 2019.

[14] E. Rodrigues, R. Assunção, G. L. Pappa, D. Renno, and W. Meira Jr, "Exploring multiple evidence to infer users' location in twitter," *Neurocomputing*, vol. 171, pp. 30–38, 2016.

[15] N. Duong-Trung, N. Schilling, and L. Schmidt-Thieme, "Near real-time geolocation prediction in twitter streams via matrix factorization based regression," in *Proceedings of the 25th ACM international on conference on information and knowledge management*, 2016, pp. 1973–1976.

[16] J. Gelernter and S. Balaji, "An algorithm for local geoparsing of microtext," *GeoInformatica*, vol. 17, no. 4, pp. 635–667, 2013.

[17] D. Rout, K. Bontcheva, D. Preoţiuc-Pietro, and T. Cohn, "Where's@ wally? a classification approach to geolocating users based on their social ties," in *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, 2013, pp. 11–20.

[18] F. Laylavi, A. Rajabifard, and M. Kalantari, "Event relatedness assessment of twitter messages for emergency response," *Information processing & management*, vol. 53, no. 1, pp. 266–280, 2017.

[19] M. Ji, L. Liu, R. Du, and M. F. Buchroithner, "A comparative study of texture and convolutional neural network features for detecting collapsed buildings after earthquakes using pre-and post-event satellite imagery," *Remote Sensing*, vol. 11, no. 10, p. 1202, 2019.

[20] H. Miura, T. Aridome, and M. Matsuoka, "Deep learning-based identification of collapsed, non-collapsed and blue tarp-covered buildings from post-disaster aerial images," *Remote Sensing*, vol. 12, no. 12, p. 1924, 2020.

[21] Q. Zhang, J. Xu, L. Xu, and H. Guo, "Deep convolutional neural networks for forest fire detection," in *2016 International Forum on Management, Education and Information Technology Application*. Atlantis Press, 2016, pp. 568–575.

[22] T. Ci, Z. Liu, and Y. Wang, "Assessment of the degree of building damage caused by disaster using convolutional neural networks in combination with ordinal regression," *Remote Sensing*, vol. 11, no. 23, p. 2858, 2019.

[23] Y. Pi, N. D. Nath, and A. H. Behzadan, "Convolutional neural networks for object detection in aerial imagery for disaster response and recovery," *Advanced Engineering Informatics*, vol. 43, p. 101009, 2020.

[24] W. Zhang, P. Tang, and L. Zhao, "Remote sensing image scene classification using cnn-capsnet," *Remote Sensing*, vol. 11, no. 5, p. 494, 2019.

[25] M.-T. Pham, L. Courtrai, C. Friguet, S. Lefèvre, and A. Baussard, "Yolo-fine: One-stage detector of small objects under various backgrounds in remote sensing images," *Remote Sensing*, vol. 12, no. 15, p. 2501, 2020.

[26] J. J. Yebes, D. Montero, and I. Arriola, "Learning to automatically catch potholes in worldwide road scene images," *IEEE Intelligent Transportation Systems Magazine*, vol. 13, no. 3, pp. 192–205, 2020.

[27] S. N. K. B. Amit, S. Shiraishi, T. Inoshita, and Y. Aoki, "Analysis of satellite images for disaster detection," in *2016 IEEE International geoscience and remote sensing symposium (IGARSS)*. IEEE, 2016, pp. 5189–5192.

[28] U. Nepal and H. Eslamiat, "Comparing yolov3, yolov4 and yolov5 for autonomous landing spot detection in faulty uavs," *Sensors*, vol. 22, no. 2, p. 464, 2022.

[29] E. Kizilay and İ. Aydin, "A yolor based visual detection of amateur drones," in *2022 International Conference on Decision Aid Sciences and Applications (DASA)*. IEEE, 2022, pp. 1446–1449.

[30] B. Liu, J. Luo, and H. Huang, "Toward automatic quantification of knee osteoarthritis severity using improved faster r-cnn," *International journal of computer assisted radiology and surgery*, vol. 15, no. 3, pp. 457–466, 2020.

[31] X. Xu, M. Zhao, P. Shi, R. Ren, X. He, X. Wei, and H. Yang, "Crack detection and comparison study based on faster r-cnn and mask r-cnn," *Sensors*, vol. 22, no. 3, p. 1215, 2022.

[32] M. Liu, J. Dong, X. Dong, H. Yu, and L. Qi, "Segmentation of lung nodule in ct images based on mask r-cnn," in *2018 9th International Conference on Awareness Science and Technology (iCAST)*. IEEE, 2018, pp. 1–6.

[33] J.-H. Lee, D.-h. Kim, S.-N. Jeong, and S.-H. Choi, "Diagnosis and prediction of periodontally compromised teeth using a deep learning-based convolutional neural network algorithm," *Journal of periodontal & implant science*, vol. 48, no. 2, pp. 114–123, 2018.

[34] D. Sarvamangala and R. V. Kulkarni, "Convolutional neural networks in medical image understanding: a survey," *Evolutionary intelligence*, pp. 1–22, 2021.

[35] L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, and Y. Miao, "Review of image classification algorithms based on convolutional neural networks," *Remote Sensing*, vol. 13, no. 22, p. 4712, 2021.

[36] H. Lamba, S. Srikanth, D. R. Pailla, S. Singh, K. S. Juneja, and P. Kumaraguru, "Driving the last mile: Characterizing and understanding distracted driving posts on social networks," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 393–404.

[37] K. Alfalqi and M. Bellaiche, "An emergency event detection ensemble model based on big data," *Big Data and Cognitive Computing*, vol. 6, no. 2, p. 42, 2022.

[38] R. Walambe, A. Marathe, and K. Kotecha, "Multiscale object detection from drone imagery using ensemble transfer learning," *Drones*, vol. 5, no. 3, p. 66, 2021.

[39] N. E. Khalifa, M. Loey, and S. Mirjalili, "A comprehensive survey of recent trends in deep learning for digital images augmentation," *Artificial Intelligence Review*, pp. 1–27, 2021.

[40] J. Lee, W. Wang, F. Harrou, and Y. Sun, "Reliable solar irradiance prediction using ensemble learning-based models: A comparative study," *Energy Conversion and Management*, vol. 208, p. 112582, 2020.

[41] https://www.reuters.com/world/india/building-fire-kills-27-new-delhi\protect\penalty-\@M-police-arrest-company-owners-2022-05-14/, accessed: 2025-2-2.

[42] D. floods, "Durban floods: South Africa floods kill more than 300," *BBC News*, Apr. 2022. [Online]. Available: https://www.bbc.com/news/world-africa-61092334

[43] O. L. F. De Carvalho, O. A. de Carvalho Junior, C. R. e. Silva, A. O. de Albuquerque, N. C. Santana, D. L. Borges, R. A. T. Gomes, and R. F. Guimarães, "Panoptic segmentation meets remote sensing," vol. 14, no. 4, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/4/965

[44] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.

[45] J. Miguel, S. Caballé, F. Xhafa, and J. Prieto, "A massive data processing approach for effective trustworthiness in online learning groups," *Concurrency and Computation: Practice and Experience*, vol. 27, no. 8, pp. 1988–2003, 2015.

[46] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview. arxiv preprint. 2020: 1–17," 2022.

[47] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. da Silva, "A comparative analysis of object detection metrics with a companion open-source toolkit," *Electronics*, vol. 10, no. 3, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/3/279

[48] Y. Pi, N. D. Nath, and A. H. Behzadan, "Convolutional neural networks for object detection in aerial imagery for disaster response and recovery," *Advanced Engineering Informatics*, vol. 43, p. 101009, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1474034619305828

[49] N. Ottakath, O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Mohamed, T. Khattab, and K. Abualsaud, "Vidmask dataset for face mask detection with social distance measurement," *Displays*, vol. 73, p. 102235, 2022.

[50] R. Khanam and M. Hussain, "What is yolov5: A deep look into the internal features of the popular object detector," 2024. [Online]. Available: https://arxiv.org/abs/2407.20892

[51] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[52] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2016. [Online]. Available: https://arxiv.org/abs/1506.01497

Resampling Imbalanced Healthcare Data for Predictive Modelling

Manoj Yadav Mamilla, Ronak Al-Haddad, Stiphen Chowdhury
Computing and Information Science
Anglia Ruskin University, East Rd, Cambridge CB1 1PT

***Abstract*—Imbalanced datasets pose significant challenges in healthcare for developing accurate predictive models in medical diagnostics. In this work, we explore the effectiveness of combining resampling methods with machine learning algorithms to enhance prediction accuracy for imbalanced heart and lung disease datasets. Specifically, we integrate undersampling techniques such as Edited Nearest Neighbours (ENN) and Instance Hardness Threshold (IHT) with oversampling methods like Random Oversampling (RO), Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN). These resampling strategies are paired with classifiers including Decision Trees (DT), Random Forests (RF), K-Nearest Neighbours (KNN), and Support Vector Machines (SVM). Model performance is evaluated using accuracy, precision, recall, F1 score, and the Area Under the Curve (AUC). Our results show that tailored resampling significantly boosts machine learning model performance in healthcare settings. Notably, SVM with ENN undersampling markedly improves accuracy for lung cancer predictions, while SVM and RF with IHT achieve higher validation accuracies for both diseases. Random oversampling shows variable effectiveness across datasets, whereas SMOTE and ADASYN consistently enhance accuracy. This study underscores the value of integrating strategic resampling with machine learning to improve predictive reliability for imbalanced healthcare data.**

***Keywords*—Imbalanced data; resampling; machine learning; healthcare**

I. Introduction

The accurate prediction of medical conditions can play a key role in improving patient outcomes and healthcare systems. Early detection of disease is essential for anticipating its clinical progression and developing effective treatments. However, accurate predictions remain a significant challenge [1].

Such predictions enable for well-informed decisions on diagnosis, intervention, and treatment by both patients and healthcare practitioners. Early detection of potential health issues enables timely, potentially life-saving interventions. These interventions significantly improve patient survival and quality of life. Furthermore, more accurate disease forecasts can help reduce healthcare costs by minimising unnecessary diagnostic tests and treatments. Hence, robust predictive models are clearly needed to enhance forecasting accuracy in clinical settings [2], [3].

Imbalanced datasets, however, present a significant challenge to achieving high predictive performance. This issue arises when one class, often a crucial medical condition, is significantly under-represented compared to the majority class. As a result, models trained on such datasets may become biased towards the majority class. This bias reduces the accuracy of minority-class predictions [4].

Precise identification of the minority class is often crucial in medical diagnostics, especially in screening programs where identifying disease-positive patients is paramount. The consequences of misclassification in these cases can be severe. The dominance of the majority class in imbalanced datasets can compromise a model's ability to detect rare but critical cases. This can lead to delayed or missed diagnoses, suboptimal treatment, and potential healthcare disparities [5].

In this paper, we propose a novel approach to predictive modelling for imbalanced healthcare data, specifically targeting heart and lung disease datasets. Our primary contribution is the integration of advanced resampling techniques with established machine learning algorithms to address data imbalance effectively. This study is driven by the following research question:

How can resampling strategies and machine learning algorithms be combined to improve predictive accuracy on imbalanced healthcare datasets?

To answer this question, we pursue the following research objectives:

- Evaluate the effectiveness of various resampling techniques (ENN, IHT, RO, SMOTE, ADASYN) in mitigating class imbalance in heart and lung disease datasets.
- Compare the performance of multiple machine learning algorithms (DT, RF, KNN, SVM) when paired with different resampling strategies.
- Determine the optimal combination of resampling methods and classifiers for improving disease prediction accuracy while maintaining computational efficiency.
- Assess the generalisability of resampling strategies across distinct medical conditions (heart disease and lung cancer) and provide insights into their applicability in real-world healthcare settings.

Various resampling strategies have been developed to ameliorate these challenges. Oversampling approaches, such as Random Oversampling, SMOTE, and ADASYN, increase the representation of the minority class by either replicating existing minority instances or generating new synthetic samples. Conversely, undersampling methods reduce the size

of the majority class by removing selected instances, which allows algorithms greater opportunity to learn minority-class patterns [6], [7]. Despite their promise, each technique has limitations. For instance, oversampling methods risk overfitting by replicating minority instances or generating low-variance synthetic points. Undersampling, on the other hand, may discard potentially valuable information.

Recent studies confirm that such resampling methods can significantly enhance the analysis and interpretation of medical data, which enables more accurate and reliable predictions [8]–[10]. However, questions remain about how best to integrate these techniques within different disease contexts. This is particularly true for conditions like heart and lung diseases, which have distinct patterns of risk factors, symptomatology, and prevalence rates.

Our experiments demonstrate that strategic combinations of sampling methods and machine learning models significantly improve both accuracy and reliability, providing a robust framework for addressing the challenges posed by imbalanced datasets in healthcare. Specifically, we find that Support Vector Machines (SVM) coupled with ENN significantly enhance lung cancer prediction accuracy. Moreover, SVM and Random Forest models utilising IHT achieve high validation accuracies for heart and lung disease data. These findings surpass traditional approaches in performance and illustrate a robust framework for effectively addressing the challenges posed by imbalanced medical datasets.

By systematically evaluating the impact of multiple resampling techniques on predictive performance, this study provides a comprehensive understanding of how to improve medical data analysis in the presence of imbalance. Our findings offer healthcare researchers and practitioners enhanced tools for early disease detection and intervention, contributing to improved patient care and outcomes.

II. Related Work

A growing body of research has sought to refine predictive modelling in healthcare by addressing the persistent challenge of data imbalance. Although these works collectively demonstrate the efficacy of outlier detection, resampling methods, and advanced machine learning strategies, key uncertainties persist regarding the transferability and consistency of these approaches across different disease domains. This section examines the existing literature, highlighting both substantial progress and the outstanding questions that motivate our study.

A. Foundational Approaches to Imbalanced Data

Fitriyani et al. [11] developed a heart disease prediction model integrating Density-based Spatial Clustering of Applications with Noise (DBSCAN) for outlier detection, hybrid SMOTE-ENN for balancing training data, and XGBoost for classification. Their results on the Statlog and Cleveland datasets emphasise the gains possible when combining sophisticated outlier removal with carefully chosen resampling techniques. However, their investigation focuses on a single disease domain and two datasets, leaving open the question of whether such hybrid pipelines would maintain similar levels of performance when confronting data from different medical conditions or with varying degrees of imbalance.

Khushi et al. [12] shifted attention to lung cancer prediction, systematically comparing traditional classifiers and imbalance strategies (including oversampling) on the PLCO and NLST datasets. Their findings reinforce the significance of well-chosen resampling techniques for boosting model performance in highly skewed datasets. Yet, the study's scope largely confines itself to lung cancer, with minimal discussion of how these strategies might generalise to other diseases - particularly those with distinct clinical signatures, feature sets, and imbalance ratios.

B. Feature Selection and Ensemble Classifiers

Ishaq et al. [13] proposed a feature-selection-based strategy for predicting heart-failure survival, demonstrating how SMOTE, combined with a Random Forest approach for feature importance, can markedly improve predictive accuracy. This underscores that well-targeted feature engineering, used in tandem with resampling, can mitigate data imbalance. However, while they demonstrate promising gains, their approach again centres on a single disease (heart failure) and does not address whether similar techniques would be equally beneficial for other conditions particularly where key feature sets differ significantly (e.g. in lung disease).

Ghorbani et al. [14] examined multiple SMOTE variants for educational data, illustrating how subtle algorithmic differences in resampling methods can translate into pronounced differences in classifier performance. Although their work concerns non-medical data, it offers a compelling reminder that each dataset may respond uniquely to different sampling methods. The question remains how to identify which undersampling or oversampling strategies e.g. ENN, IHT, SMOTE, ADASYN are most compatible with specific disease datasets that frequently exhibit high dimensionality, noisy features, or overlapping classes.

C. Advanced Architectures and Cross-Domain Insights

Li et al. [15] demonstrated how complex, dual-stage attention-based models with CBAM modules can handle data imbalance in engineering fault diagnosis. While fault-diagnosis data differ from clinical datasets, their positive results point toward the adaptability of advanced architectures for skewed classification problems. Nonetheless, applying such intricate methods to medical data raises additional concerns of interpretability and domain relevance, as well as the significant computational overhead often associated with deep learning in real-world clinical settings.

D. Summary of Gaps and Motivation

Despite these advancements, three primary gaps emerge:

1. Current studies typically focus on a single disease (e.g. heart disease or lung cancer), limiting our understanding of whether and how resampling methods generalise across conditions with distinct pathophysiology, class ratios, and feature distributions.

2. While various works explore SMOTE, ENN, or other techniques, they frequently implement a single approach or compare only a narrow set of sampling techniques. Similarly, they often limit themselves to a small subset of classification

algorithms, resulting in an incomplete picture of how different combinations of sampling algorithms and machine learning models might perform under different clinical constraints.

3. Many studies highlight accuracy gains, but relatively few provide detailed insights into the computational trade-offs such as training and inference times, resource consumption, or real-time feasibility that are crucial for deployment in actual healthcare settings, for instance, Fitriyani et al. [11] and Khushi et al. [12] focus on performance without efficiency analysis. Moreover, interpretability and domain-specific constraints have not always been sufficiently addressed when employing more advanced modelling frameworks.

In light of these gaps, our study aims to:

1) *Evaluate multiple resampling methods (ENN, IHT, RO, SMOTE, and ADASYN) across two distinct diseases*: heart disease and lung cancer. By spanning two clinically significant but different pathologies, we shed light on how imbalanced data strategies may generalise or require adaptation across disease contexts.
2) *Compare four diverse machine learning algorithms* (Decision Trees, Random Forests, K-Nearest Neighbours, and Support Vector Machines) under each resampling technique, thus providing a broader evidence base on which model–sampling pairs work best for specific disease datasets.
3) *Examine computational efficiency and practicality*, reporting on training/validation/testing times and potential interpretability issues, thereby informing real-world clinical adoption.

By addressing these specific gaps, our research contributes a deeper and more generalisable understanding of how to tailor resampling approaches and classifier choices in the face of disease-specific imbalances. This work ultimately aspires to promote more robust, reliable, and domain-aware predictive models in healthcare.

III. Methodology

We employ a straightforward yet comprehensive methodology (see Fig. 1) to address the challenges posed by data imbalance in medical predictive modelling. Our process encompasses data preprocessing, various resampling strategies, machine learning algorithm selection, and performance evaluation. The choice of each algorithm and sampling method is guided by the characteristics of the classification tasks (i.e. heart disease vs. lung disease) and by the need to mitigate imbalance effectively.

A. Sampling Techniques

1) Undersampling Techniques: Edited Nearest Neighbours (ENN) and Instance Hardness Threshold (IHT) are employed to mitigate the effect of imbalanced datasets. ENN removes misclassified or noisy observations by comparing each sample to its k nearest neighbours, thereby cleansing borderline cases in the dataset. In contrast, IHT identifies and removes "hard-to-classify" instances based on their proximity to the decision boundary, helping models concentrate on more informative examples [16], [17]. By discarding problematic samples, both methods aim to create a dataset that better represents minority-class patterns without overwhelming the model.

2) Oversampling Techniques: Random Oversampling duplicates minority-class instances to balance the dataset, albeit with a heightened risk of overfitting. To address this limitation, we also use the Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic (ADASYN) sampling, both of which generate synthetic examples of the minority class [18], [19]. These strategies introduce diversity into training data and reduce the bias towards the majority class, enabling models to learn more nuanced decision boundaries.

B. Datasets Utilisation

Two publicly available datasets are utilised in this study. The Kaggle Lung Cancer dataset [20] comprises 163,763 records with detailed patient information (including binary labels denoting the presence of lung cancer). The heart disease dataset [21] from the UCI repository combines multiple sources and focuses on 14 key attributes (out of an original 76) to predict the presence of cardiovascular disease. These datasets were chosen for their relevance to clinical practice and their relatively high degree of imbalance, offering an appropriate testbed for evaluating sampling methods.

C. Machine Learning Algorithms

Four machine learning algorithms are employed to assess the impact of each sampling strategy:

- Decision Trees (DT): An interpretable, rule-based approach that partitions data into increasingly homogeneous subsets.
- Random Forests (RF): An ensemble of multiple decision trees, aggregating predictions to reduce overfitting and enhance generalisability.
- K-Nearest Neighbours (KNN): A distance-based classifier capable of handling non-linear decision boundaries, sensitive to local data structure.
- Support Vector Machines (SVM): A robust method for binary classification, especially effective in high-dimensional spaces, using kernel functions to separate classes with maximal margins.

Each algorithm is trained and evaluated on both datasets with the aforementioned sampling techniques applied. This setup allows for a direct comparison of how undersampling and oversampling methods impact model accuracy, precision, recall, and F1-score.

D. Methodological Flow and Evaluation

As summarised in Fig. 1, data preprocessing and resampling are carried out first to correct for imbalance. The selected models (DT, RF, KNN, and SVM) are then trained on these preprocessed datasets. Finally, we measure predictive performance using accuracy, precision, recall, and F1-score to gauge the effectiveness of each sampling approach.

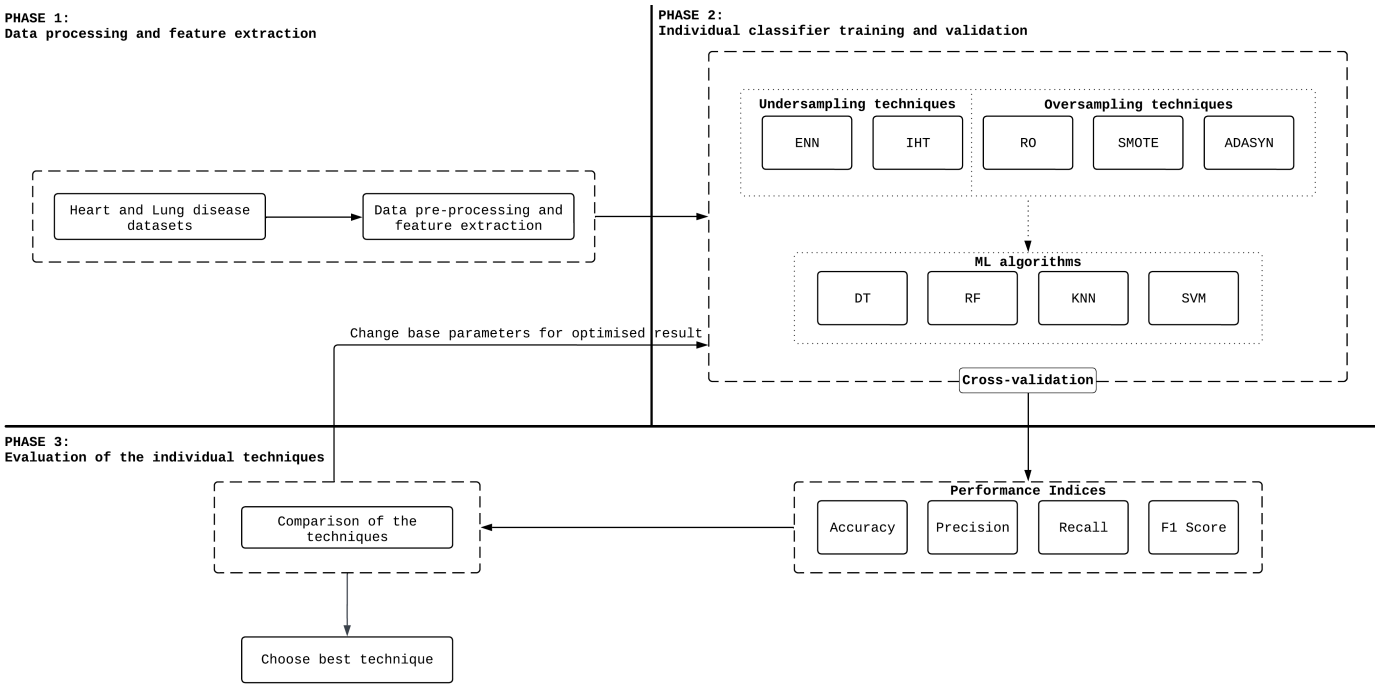


Fig. 1. A three-phase approach to heart and lung disease classification, including data preprocessing, resampling techniques, machine learning model selection, and performance evaluation.

IV. Experimental Setup

This study investigates the effect of various resampling strategies on the predictive accuracy of machine learning models when dealing with imbalanced healthcare data from Kaggle (lung cancer) and UCI Machine Learning Repository (heart disease).

A. Dataset Description

1) Kaggle Lung Cancer Dataset: This dataset comprises patient-level data intended for predicting the occurrence of lung cancer. It includes features such as age, gender, smoking history, and various diagnostic measurements. The primary challenge stems from its imbalanced distribution, with a significantly lower proportion of positive lung cancer instances compared to negative ones.

2) UCI Heart Disease Dataset: This collection of patient records is aimed at predicting the presence of heart disease. It includes demographic information, symptomatology, and diagnostic test results. Similar to the lung cancer dataset, it is highly imbalance, with heart disease instances occurring less frequently than negative cases, which makes it an excellent testbed for evaluating the effectiveness of resampling techniques.

B. Data Collection and Preprocessing

Data were collected and rigorously preprocessed to ensure quality and consistency. The preprocessing phase addressed missing values, outliers, and normalisation:

- Mean Imputation for Missing Values: For each feature containing missing entries, we replaced missing values with the mean value of that feature across all available data points:
$$\mu_{\text{missing}} = \frac{\sum_{i=1}^{n} x_i}{n}.$$
- Normalisation (Min–Max Scaling): We rescaled each feature to a common range of $[0, 1]$ using Min–Max scaling:
$$x' = \frac{x - \min(x)}{\max(x) - \min(x)},$$
ensuring that differences in the original data ranges are preserved without distortion.

C. Sampling Techniques

Given the imbalanced nature of these datasets, we applied five resampling methods: two undersampling techniques - Edited Nearest Neighbour (ENN) and Instance Hardness Threshold (IHT) - and three oversampling techniques - Random Oversampling (RO), Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic (ADASYN).

- Edited Nearest Neighbour (ENN): ENN refines the dataset by removing samples that are misclassified by their k-nearest neighbours, thereby enhancing the purity of the minority class:
$$ENN(S) = \{\, x \in S \mid \text{class}(x) = \text{class}(\text{kNN}(x)) \}.$$
- Instance Hardness Threshold (IHT): IHT iteratively applies oversampling approaches and evaluates classifier performance on the modified dataset, aiming to attain a balance that optimises the model's sensitivity to the minority class.

- Random Oversampling (RO): RO duplicates instances of the minority class to rebalance the dataset, albeit at an increased risk of overfitting.
- SMOTE and ADASYN: Both SMOTE and ADASYN generate synthetic minority examples, helping to balance class distribution and introduce diversity to the training data. This prepares the model to generalise more effectively from underrepresented classes.

D. Model Selection and Training

We selected four machine learning algorithms for evaluation: Decision Trees, Random Forests, K-Nearest Neighbours, and Support Vector Machines. These models represent a spectrum of complexity: from the relatively interpretable structure of Decision Trees to the more sophisticated nature of SVMs. The training process involved optimising hyperparameters for each model using techniques such as grid search and cross-validation to maximise overall performance.

E. Evaluation Metrics

We employed accuracy, precision, recall, and the F1-score to evaluate model performance comprehensively. These metrics were chosen not only to gauge overall accuracy but also to assess the models' ability to identify positive instances (recall) and the reliability of those predictions (precision).

a) Precision:

$$\text{Precision} = \frac{TP}{TP + FP},$$

where TP (true positives) are correctly predicted positive instances, and FP (false positives) are incorrectly predicted as positive. High precision in this work indicates that when the model predicts a disease, it is likely correct.

b) Recall (Sensitivity):

$$\text{Recall} = \frac{TP}{TP + FN},$$

where FN (false negatives) are positive instances incorrectly predicted as negative. Recall here reflects the model's capacity to detect all relevant cases.

c) F1-Score:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

This harmonic mean of precision and recall is especially relevant when misclassifications either false positives or false negatives carry significant risks, as is common in healthcare.

Finally, multiple runs were conducted for each model and sampling combination to ensure the reliability of findings and to capture performance variance across different trials. This robust experimental structure enables a comprehensive assessment of how well different resampling strategies address data imbalance in healthcare predictive modelling.

V. Results

A. Experiment with the Heart Dataset

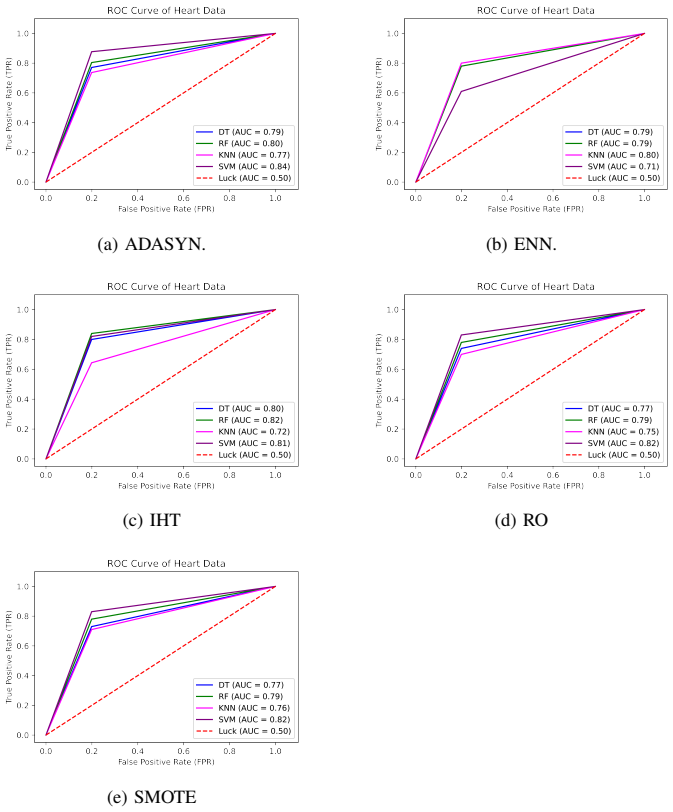
Table I presents the results for the heart dataset. Overall, Support Vector Machines (SVM) excelled in precision, particularly when paired with Edited Nearest Neighbours (ENN), underlining its effectiveness in handling imbalanced data. SVM thus appears well-suited to heart disease prediction, consistently achieving high scores in precision, F1, and AUC. These results are especially significant in medical diagnostics, where balancing sensitivity and specificity can be challenging. Notably, the AUC values were strong for Random Forest (RF) with Instance Hardness Threshold (IHT) and SVM with Adaptive Synthetic Sampling (ADASYN), further highlighting SVM's capacity to reduce false positives while maintaining a high detection rate.

TABLE I. Performance Metrics for Various Classifiers on the Heart Dataset Across Different Resampling Techniques, Demonstrating Superior Precision and Recall by SVM and RF

		ENN	IHT	RANDOM	SMOTE	ADASYN
DT	Precision	0.90	0.82	0.88	0.82	0.81
	Recall	0.90	0.79	0.83	0.82	0.76
	F1 Score	0.90	0.81	0.85	0.82	0.79
	AUC	0.781	0.803	0.747	0.7361	0.7712
RF	Precision	0.90	0.89	0.90	0.87	0.88
	Recall	**0.93**	0.89	0.77	0.77	0.75
	F1 Score	0.91	**0.89**	0.83	0.82	0.81
	AUC	0.781	**0.843**	0.78	0.7806	0.804
KNN	Precision	0.92	0.68	0.71	0.73	0.73
	Recall	0.83	0.59	0.71	0.63	**0.88**
	F1 Score	0.87	0.63	0.71	0.68	**0.80**
	AUC	**0.802**	0.6449	0.705	0.719	0.7374
SVM	Precision	**0.96**	**0.95**	**0.93**	0.74	**0.91**
	Recall	0.93	0.75	0.74	**0.94**	0.90
	F1 Score	**0.95**	0.84	0.83	0.83	0.90
	AUC	0.618	0.8224	0.8305	0.8306	**0.8779**

Fig. 2 illustrates ROC curves for various classifiers on the Heart Dataset across different resampling techniques, further confirming that SVM and RF generally outperform the other classifiers. Particularly noteworthy is ENN's significant enhancement of SVM's performance, as reflected in the high AUC scores. Likewise, ADASYN combined with SVM reached the highest AUC among the tested setups, indicating its efficacy in tackling data imbalance. By comparison, the RANDOM and SMOTE approaches yielded moderately lower AUCs, underscoring the importance of carefully aligning sampling methods with the dataset's characteristics.

Table II indicates that computational efficiency varies with the choice of resampling method. Decision Tree (DT) and KNN demonstrate notably faster training times, especially under the RANDOM and ADASYN schemes, suggesting relatively low computational overhead. By contrast, SVM exhibits longer training times particularly with RANDOM highlighting a potential trade-off between speed and predictive performance.



(a) ADASYN. (b) ENN.

(c) IHT (d) RO

(e) SMOTE

Fig. 2. ROC curves for various classifiers on the Heart Dataset across different resampling techniques. The curves demonstrate the superior performance of SVM and RF, particularly with ENN and ADASYN. These methods enhance the classifiers' ability to handle imbalanced data, as evidenced by high AUC values.

TABLE II. COMPARISON OF TRAINING, VALIDATION, AND TESTING TIMES FOR DIFFERENT CLASSIFIERS ON THE HEART DATASET USING VARIOUS RESAMPLING METHODS. NOTABLE DIFFERENCES IN COMPUTATIONAL EFFICIENCY ARE OBSERVED, PARTICULARLY FOR SVM AND RF

		ENN	IHT	RANDOM	SMOTE	ADASYN
DT	Training	0.0031	0.002	**0.0001**	0.0038	**0.0001**
	Validation	0.013	0.010	**0.0084**	**0.0084**	0.0093
	Testing	0.0151	0.0136	0.0050	0.0090	**0.0006**
RF	Training	0.2885	0.0172	0.0802	0.0745	**0.0438**
	Validation	0.0280	**0.0080**	0.0082	0.0170	0.0083
	Testing	0.0280	**0.0063**	0.0104	0.0161	0.0133
KNN	Training	0.0051	**0.0001**	**0.0001**	0.0035	**0.0001**
	Validation	0.0160	**0.0033**	0.0111	0.0102	0.0111
	Testing	0.0125	0.0099	0.0063	0.0128	**0.0040**
SVM	Training	0.2191	0.0780	**0.0577**	0.0919	0.0825
	Validation	0.0143	0.0114	**0.0071**	0.0088	0.0103
	Testing	0.0121	**0.0020**	0.0077	0.0101	0.0056

B. Experiment with the Lung Dataset

Table III details the results for the lung dataset. Both Decision Tree (DT) and Random Forest (RF) achieved better precision when paired with ENN, IHT, and RANDOM sampling, underlining their capacity for highly accurate lung disease classification. However, DT's slightly lower AUC under RANDOM sampling and SMOTE suggests minor compromises in balancing sensitivity and specificity. While DT occasionally overlooks some true cases, as evidenced by its recall, RF's flawless recall with ENN and IHT demonstrates that it seldom misses positive instances. As a result, both DT and RF prove highly dependable, with RF offering a small edge in comprehensive patient identification.

TABLE III. PERFORMANCE METRICS FOR VARIOUS CLASSIFIERS ON THE LUNG DATASET ACROSS DIFFERENT RESAMPLING TECHNIQUES. PHENOMENAL PRECISION AND RECALL ARE ACHIEVED BY DT AND RF WITH ENN AND IHT, WHILE SVM SHOWS CONSISTENTLY HIGH PERFORMANCE

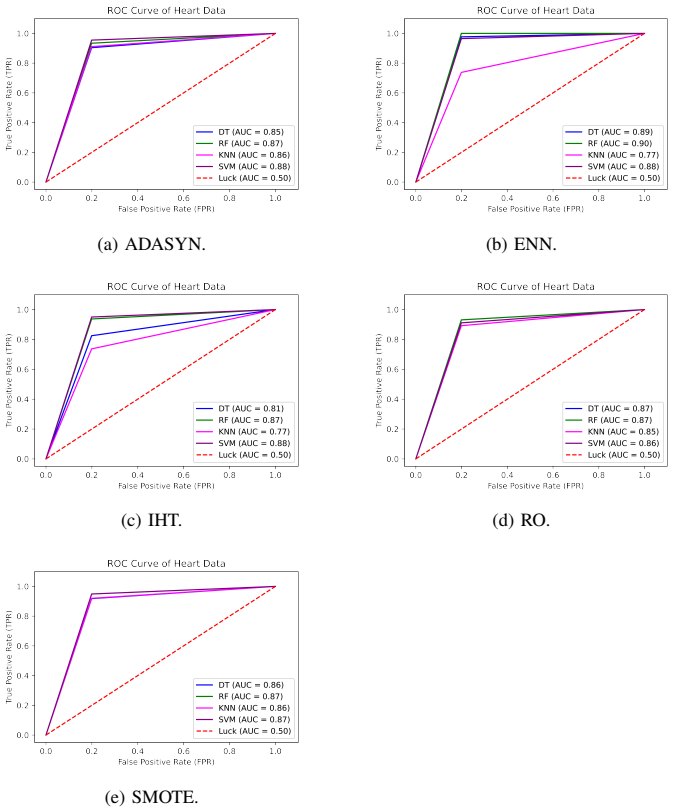
		ENN	IHT	RANDOM	SMOTE	ADASYN
DT	Precision	**1**	**1**	**1**	0.96	0.94
	Recall	0.95	**1**	0.88	0.88	0.91
	F1 Score	0.98	**1**	0.93	0.92	0.93
	AUC	0.977	0.825	0.931	0.918	0.9041
RF	Precision	**1**	**1**	**1**	0.92	0.95
	Recall	**1**	**1**	0.88	0.98	0.93
	F1 Score	**1**	**1**	0.94	0.95	0.94
	AUC	**1**	0.93	0.931	0.94	0.934
KNN	Precision	0.98	0.75	**1**	0.93	0.92
	Recall	0.98	0.90	0.73	0.93	0.91
	F1 Score	0.98	0.82	0.84	0.93	0.91
	AUC	0.738	0.7375	0.8921	0.92	0.911
SVM	Precision	**1**	**1**	**1**	0.95	0.96
	Recall	**1**	**1**	0.90	0.95	0.96
	F1 Score	**1**	**1**	0.95	0.95	0.96
	AUC	0.9659	**0.95**	0.911	0.949	**0.955**

While KNN shows some variability particularly in precision with IHT it still maintains solid results in precision and recall, indicating its ability to identify lung disease patients accurately. By contrast, SVM consistently delivers near-perfect performance across all evaluated metrics and resampling configurations, highlighting its powerful capacity to discriminate between diseased and healthy individuals. Its balanced high precision and recall minimise both misdiagnoses and missed diagnoses.

TABLE IV. TRAINING, VALIDATION, AND TESTING TIMES FOR DIFFERENT CLASSIFIERS ON THE LUNG DATASET USING VARIOUS RESAMPLING METHODS. DT AND KNN OFTEN REQUIRE MINIMAL COMPUTATION WITH RANDOM AND ADASYN

		ENN	IHT	RANDOM	SMOTE	ADASYN
DT	Training	**0.0001**	0.0077	**0.0001**	**0.0001**	0.0081
	Validation	**0.0046**	0.0098	0.0083	0.0102	0.0131
	Testing	0.0085	**0.0030**	0.0085	0.0098	0.0202
RF	Training	**0.0193**	0.0250	0.0999	0.0833	0.0329
	Validation	0.0122	**0.0107**	0.0134	0.0169	0.0203
	Testing	0.0110	0.0081	0.0110	0.0087	0.0123
KNN	Training	0.0065	0.0017	**0.0001**	0.0015	0.0020
	Validation	0.0081	0.0082	0.0100	**0.0081**	0.0142
	Testing	0.0100	0.0067	**0.0076**	0.0146	0.0166
SVM	Training	0.0080	**0.0001**	0.8198	0.0121	0.0166
	Validation	0.0062	0.0085	0.0084	**0.0082**	0.0145
	Testing	0.0084	0.0086	0.0101	0.0112	**0.0101**

The F1 and AUC values reaching 100% for RF with ENN and IHT underscore the potency of these resampling methods in boosting accuracy for markedly skewed datasets. Interestingly, the lung dataset seems to respond more strongly to these techniques than the heart dataset, potentially reflecting



(a) ADASYN.

(b) ENN.

(c) IHT.

(d) RO.

(e) SMOTE.

Fig. 3. ROC curves for various classifiers on the Lung Dataset across different resampling techniques. The curves highlight the exceptional precision and recall achieved by RF and SVM, especially with ENN and IHT, confirming their robustness in identifying lung disease despite data imbalance.

inherent differences in class distribution or unique dataset attributes.

Fig. 3 shows ROC curves for various classifiers on the Lung Dataset across different resampling techniques. The curves highlight the exceptional precision and recall achieved by RF and SVM, especially with ENN and IHT, confirming their robustness in identifying lung disease despite data imbalance.

Finally, Table IV summarises the training, validation, and testing times for different classifiers on the Lung Dataset. Decision Trees (DT) and KNN demonstrate relatively low computational demands, most notably under the RANDOM and ADASYN resampling methods, making them viable options where computational resources are limited. Random Forest (RF) and SVM, in contrast, show varying training times, with RF proving efficient under ENN and IHT, while SVM attains rapid validation times with ENN. These results suggest that DT and KNN may suffice where speed is paramount, whereas RF and SVM, coupled with certain resampling strategies, strike a more favourable balance between computational overhead and predictive performance.

VI. Discussion

The experimental results confirm that pairing tailored resampling methods with appropriate classifiers can greatly improve predictive performance in imbalanced healthcare datasets. The results show that combining advanced resampling techniques with machine learning models improves predictive accuracy for imbalanced healthcare data. However, performance differences across datasets highlight the need to select resampling strategies and classifiers based on dataset characteristics.

A. Effectiveness of Resampling Methods

Resampling techniques play a crucial role in addressing class imbalance in heart disease and lung cancer predictions. Instance Hardness Threshold (IHT) and Edited Nearest Neighbours (ENN) performed best for undersampling, while ADASYN and SMOTE increased minority-class representation through synthetic data generation. Their effectiveness varied based on the dataset and classifier.

IHT improved precision and recall in heart disease predictions because it removed ambiguous majority-class instances. This effect was most pronounced when combined with Random Forest and Support Vector Machines. ENN produced the best results for lung cancer predictions because it refined the decision boundary and enhanced SVM's ability to distinguish between classes. ADASYN and SMOTE produced general improvements, but they worked best for heart disease predictions where a moderate imbalance allowed for better synthetic sample generation. These findings confirm that resampling techniques must be chosen based on the dataset rather than applied universally.

B. Classifier Performance Across Datasets

Classifier performance depended on disease type and resampling strategy. Support Vector Machines performed best for lung cancer prediction, especially when combined with ENN. This combination removed noisy samples and allowed SVM to establish a clearer decision boundary. Random Forest and Decision Trees outperformed other models for heart disease predictions because they captured feature interactions effectively.

SVM provided the highest accuracy and recall for lung cancer predictions, making it the best choice for highly imbalanced, binary-structured datasets. Random Forest and Decision Trees worked better for heart disease because they handled mixed categorical and numerical data more effectively. K-Nearest Neighbours showed inconsistent results across both datasets because it was highly sensitive to imbalance and feature distribution.

C. Dataset-Specific Performance

Model performance varied between the heart disease and lung cancer datasets due to differences in class imbalance severity, feature types, and data distribution. The heart disease dataset, with moderate imbalance and mixed categorical-numerical features, favored Random Forest with IHT and ADASYN, which excel at capturing complex feature interactions. Conversely, the lung cancer dataset, with extreme imbalance and mostly binary features, suited SVM with ENN, as this combination effectively refines sparse decision boundaries. These results suggest our algorithms are better suited to specific data types: RF thrives with heterogeneous features,

while SVM excels with binary, highly skewed data. This adaptability underscores the importance of matching resampling and classifiers to dataset properties.

D. Computational Considerations

The study also highlights key computational trade-offs. SVM and Random Forest provided the highest accuracy, but they required longer training times. These models are suitable for offline model development rather than real-time applications. Decision Trees and K-Nearest Neighbours trained and tested faster but produced less consistent results for imbalanced data. Selecting a model for deployment depends on balancing efficiency, interpretability, and predictive performance.

E. Real-World Implications

These findings have direct applications in predictive healthcare. SVM with ENN delivers the most accurate classification for lung cancer screening, reducing false negatives. Random Forest with IHT works better for heart disease prediction because it maintains interpretability while achieving strong performance. These findings guide healthcare practitioners in selecting predictive models that improve early diagnosis and patient outcomes.

F. Advantages Over Existing Methods

Our approach outperforms traditional methods by integrating resampling with classifier optimisation tailored to dataset characteristics. Unlike studies like Fitriyani et al. [11], which focus on single-disease hybrid pipelines (e.g. SMOTE-ENN with XGBoost), our framework evaluates multiple resampling techniques (ENN, IHT, RO, SMOTE, ADASYN) across two diseases, offering broader applicability. Compared to Khushi et al. [12], which limits comparisons to lung cancer with basic oversampling, our SVM+ENN and RF+IHT combinations achieve higher precision and recall (e.g. 100% F1 for lung cancer with RF+ENN) while addressing computational trade-offs. This adaptability and performance edge make our method a versatile tool for imbalanced healthcare data beyond single-context solutions.

G. Validation and Comparative Significance

Our use of comprehensive validation measures such as accuracy, precision, recall, F1 score, and AUC ensures robust evaluation of model performance, prioritising both correctness and sensitivity critical in healthcare. Unlike prior works like Ishaq et al. [13], which focus on accuracy alone, our multi-metric approach highlights trade-offs (e.g. SVM+ENN's perfect recall for lung cancer). Thorough comparisons with related studies, for instance, outperforming Khushi et al.'s [12] lung cancer precision with RF+IHT, validate our method's superiority. This rigorous validation confirms its reliability and generalisability across imbalanced medical datasets.

VII. Conclusion

This study demonstrates that advanced resampling techniques, when integrated with optimised machine learning models, can significantly enhance classification performance for imbalanced healthcare data. Specifically, SVM with ENN excelled in lung cancer prediction, while RF with IHT or ADASYN consistently performed well for heart disease prediction. These findings provide a roadmap for choosing optimal resampling - classifier pairs based on dataset characteristics, which is crucial for early diagnosis and resource allocation in clinical practice. Our results emphasise the importance of robust methods to handle skewed medical datasets, leading to more reliable healthcare predictions.

VIII. Future Direction

Despite these advances, much remains to be done to fully realise the potential of predictive models in healthcare. The following key challenges merit attention:

1) Adaptation to novel diseases: Models trained on current datasets may underperform when confronted with emerging diseases, such as those appearing during pandemics. Differences in symptoms, transmission mechanisms, or patient demographics can impede model adaptability. Future studies should explore adaptive learning frameworks - particularly online learning techniques that enable continuous learning from new data streams. Incorporating concept-drift detection can help identify performance degradation due to shifts in data distribution, while transfer learning offers a means of rapidly adapting pre-trained models to smaller, disease-specific datasets.

2) Scalability: The exponential growth of healthcare data in both volume and complexity necessitates scalable predictive models. Managing large datasets requires substantial computational resources and careful maintenance of model efficiency. Distributed computing platforms (e.g., Apache Spark or Dask) can facilitate the processing of extensive datasets across clusters. Additionally, algorithmic efficiency may be improved through dimensionality reduction and the deployment of more efficient neural architectures, such as EfficientNets, to achieve real-time processing capabilities.

3) Multimodal integration: Healthcare data are inherently multimodal, encompassing structured electronic health records, unstructured clinical notes, medical imaging, and genomic information. Current predictive models often fail to fully exploit these diverse data sources. Developing multimodal learning methods that accommodate varying data formats can provide a more comprehensive view of patient health. Techniques such as Canonical Correlation Analysis (CCA) and cross-modal neural networks can learn joint representations, while strategic fusion strategies help integrate disparate data modalities effectively.

4) Handling rare diseases: Rare diseases pose particular challenges due to significant class imbalance and very limited positive cases. Insufficient data hampers the training of robust predictive models, thereby limiting opportunities for early intervention. Future work should focus on few-shot learning approaches, for example, prototypical networks and Model-Agnostic Meta-Learning (MAML) that enable models to learn from minimal examples. Transfer learning can also leverage existing biomedical knowledge by fine-tuning models trained on more common diseases for rare disease datasets.

5) Bias and fairness: Predictive models can inherit biases from training data, potentially reinforcing inequities in healthcare access and outcomes. Without appropriate safeguards,

such models may over-predict or under-predict disease risk in certain demographic groups. Fairness-aware machine learning strategies such as re-weighting datasets for balanced representation, introducing fairness constraints during training, and performing post-hoc adjustments are therefore essential. Rigorous bias audits, which scrutinise performance across diverse patient populations, are vital for identifying and mitigating disparities.

6) Interpretability and trust: The "black box" nature of many advanced machine learning models is a major obstacle to clinical adoption, as healthcare professionals are often reluctant to rely on systems whose inner workings they cannot examine. Enhancing interpretability is crucial for clinician trust and seamless adoption within medical practice. Explainable AI (XAI) methods such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) can clarify model reasoning. User-friendly interfaces that integrate with existing healthcare IT systems can further bolster clinician acceptance and support practical workflows.

Beyond these methodological and computational considerations, practical issues such as data-sharing restrictions, privacy laws, and regulatory compliance (e.g. with the MHRA or NICE guidelines) will also shape the real-world deployment of AI-driven healthcare models. Although our results are encouraging, implementing these solutions at scale poses substantial challenges. Overcoming these hurdles will be essential for fully leveraging the potential of predictive modelling to improve patient outcomes and transform healthcare delivery.

REFERENCES

[1] S. Emmons-Bell, C. Johnson, and G. Roth, "Prevalence, incidence and survival of heart failure: a systematic review," *Heart*, vol. 108, no. 17, pp. 1351–1360, 2022.

[2] F. Alahmari, "A comparison of resampling techniques for medical data using machine learning," *Journal of Information & Knowledge Management*, vol. 19, no. 01, p. 2040016, 2020.

[3] J. Thompson Burdine, S. Thorne, and G. Sandhu, "Interpretive description: a flexible qualitative methodology for medical education research," *Medical education*, vol. 55, no. 3, pp. 336–343, 2021.

[4] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *Proceedings of the National Academy of Sciences*, vol. 117, no. 23, pp. 12 592–12 594, 2020.

[5] S. Hong, Y. Zhou, J. Shang, C. Xiao, and J. Sun, "Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review," *Computers in biology and medicine*, vol. 122, p. 103801, 2020.

[6] R. Kumar, W. Wang, J. Kumar, T. Yang, A. Khan, W. Ali, and I. Ali, "An integration of blockchain and ai for secure data sharing and detection of ct images for the hospitals," *Computerized Medical Imaging and Graphics*, vol. 87, p. 101812, 2021.

[7] H. Matsuo, M. Nishio, T. Kanda, Y. Kojita, A. K. Kono, M. Hori, M. Teshima, N. Otsuki, K.-i. Nibu, and T. Murakami, "Diagnostic accuracy of deep-learning with anomaly detection for a small amount of imbalanced data: discriminating malignant parotid tumors in mri," *Scientific Reports*, vol. 10, no. 1, p. 19388, 2020.

[8] S. J. Stratton, "Population research: convenience sampling strategies," *Prehospital and disaster Medicine*, vol. 36, no. 4, pp. 373–374, 2021.

[9] Y. K. Cherniavskyi, A. Fathizadeh, R. Elber, and D. P. Tieleman, "Computer simulations of a heterogeneous membrane with enhanced sampling techniques," *The Journal of Chemical Physics*, vol. 153, no. 14, 2020.

[10] R. Evans, L. Hovan, G. A. Tribello, B. P. Cossins, C. Estarellas, and F. L. Gervasio, "Combining machine learning and enhanced sampling techniques for efficient and accurate calculation of absolute binding free energies," *Journal of chemical theory and computation*, vol. 16, no. 7, pp. 4641–4654, 2020.

[11] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "Hdpm: an effective heart disease prediction model for a clinical decision support system," *IEEE Access*, vol. 8, pp. 133 034–133 050, 2020.

[12] M. Khushi, K. Shaukat, T. M. Alam, I. A. Hameed, S. Uddin, S. Luo, X. Yang, and M. C. Reyes, "A comparative performance analysis of data resampling methods on imbalance medical data," *IEEE Access*, vol. 9, pp. 109 960–109 975, 2021.

[13] A. Ishaq, S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, "Improving the prediction of heart failure patients' survival using smote and effective data mining techniques," *IEEE access*, vol. 9, pp. 39 707–39 716, 2021.

[14] R. Ghorbani and R. Ghousi, "Comparing different resampling methods in predicting students' performance using machine learning techniques," *IEEE Access*, vol. 8, pp. 67 899–67 911, 2020.

[15] J. Li, Y. Liu, and Q. Li, "Intelligent fault diagnosis of rolling bearings under imbalanced data conditions using attention-based deep learning method," *Measurement*, vol. 189, p. 110500, 2022.

[16] Y. Zhu, C. Jia, F. Li, and J. Song, "Inspector: a lysine succinylation predictor based on edited nearest-neighbor undersampling and adaptive synthetic oversampling," *Analytical biochemistry*, vol. 593, p. 113592, 2020.

[17] J. L. Arruda, R. B. Prudêncio, and A. C. Lorena, "Measuring instance hardness using data complexity measures," in *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part II 9*. Springer, 2020, pp. 483–497.

[18] A. Glazkova, "A comparison of synthetic oversampling methods for multi-class text classification," *arXiv preprint arXiv:2008.04636*, 2020.

[19] Z. Chen, L. Zhou, and W. Yu, "Adasyn- random forest based intrusion detection model," in *2021 4th International Conference on Signal Processing and Machine Learning*, 2021, pp. 152–159.

[20] Z. Hong and J. Yang, "Lung Cancer," UCI Machine Learning Repository, 1992, DOI: https://doi.org/10.24432/C57596.

[21] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart disease data set," 1988. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Exploiting Ray Tracing Technology Through OptiX to Compute Particle Interactions with Cutoff in a 3D Environment on GPU

David Algis[1], Bérenger Bramas[2]
University of Poitiers, XLIM, France[1]
Studio Nyx, France[1]
Inria Nancy, France[2]
ICube Laboratory, France[2]
University of Strasbourg, France[2]

***Abstract*—Particle interaction simulation is a fundamental method of scientific computing that require high-performance solutions. In this context, computing on graphics processing units (GPUs) has become standard due to the significant performance gains over conventional CPUs. However, since GPUs were originally designed for 3D rendering, they still retain several features that are not fully exploited in scientific computing. One such feature is ray tracing, a powerful technique for rendering 3D scenes. In this paper, we propose exploiting ray tracing technology via OptiX and CUDA to compute particle interactions with a cutoff distance in a 3D environment on GPUs. To this end, we describe algorithmic techniques and geometric patterns for efficiently determining the interaction lists for each particle. Our approach enables the computation of interactions with quasi-linear complexity in the number of particles, eliminating the need to construct a grid of cells or an explicit kd-tree. We compare the performance of our method to a classical grid-based approach and demonstrate that our approach is faster in most cases with non-uniform particle distributions.**

Keywords*—CUDA; graphics processing unit; high-performance computing; OptiX; particle interactions; ray tracing; scientific computing

I. Introduction

High-performance computing (HPC) is a key technology in scientific computing. Since the early 2000s, the use of graphics processing units (GPUs) has become standard in HPC, and they now equip many of the fastest supercomputers.[1] GPUs are massively parallel processors that allow computations to be performed on thousands of cores, making them perfectly suited for inherently parallel problems. The use of GPUs in scientific computing has led to incredible performance gains in many fields, such as molecular dynamics [1], fluid dynamics [2], astrophysics [3], and machine learning [4].

Most scientific computing applications that use GPUs are based on the CUDA[2] programming model, and to a lesser extent, the OpenCL programming model. They do not exploit all the features of GPUs, particularly those dedicated to 3D rendering. Among these features, ray tracing is a powerful technology used to render 3D scenes. In this method, rays are cast from the camera to the scene, and the intersections of the rays with the objects in the scene are computed to generate the colors of the pixels in the image [5]. For example, the NVIDIA GeForce RTX 4090, as a single consumer-grade GPU, demonstrates the raw power of this technology by rendering complex 3D scenes at 87 frames per second in 4K resolution (3840 x 2160 pixels), handling millions of triangles per frame [6].

In this paper, we are interested in evaluating how ray tracing technology could be used to compute particle interactions in a 3D environment. Our motivation is twofold: We want to evaluate if it is possible to use ray tracing technology, and we want to create the algorithmic patterns needed for that purpose.

With these aims, we focus on the computation of particle interactions in a 3D environment, which is a common problem in scientific computing. When the potential of the interaction kernel decreases exponentially with distance, the interactions can be computed with a cutoff distance, i.e. the interactions are only computed between particles that are closer than a given distance, achieving less but still satisfactory accuracy. This allows the complexity of the interactions to drop from $O(N^2)$ to $O(N)$, where N is the number of particles if the cutoff distance is small enough. Classical methods to compute such interactions are based on the use of a grid of cells or an explicit kd-tree to quickly find the interaction lists for each particle. In this paper, we aim to avoid using such data structures and instead exploit ray tracing technology.

The contributions of this paper are as follows:

- We propose a method to compute particle interactions with a cutoff distance in a 3D environment based on ray tracing technology.
- We describe two algorithmic techniques based on geometric patterns to find the interaction lists for each particle using real intersections.
- We compare the performance of our approach with a classical approach based on a grid of cells and with an existing method that relies on ray tracing.

The rest of the paper is organized as follows: In Section II, we present the prerequisites. In Section III, we review the related work. In Section IV-A, we present our proposed

[1]https://top500.org/
[2]https://developer.nvidia.com/cuda-toolkit

approach. In Section V, we present the performance study. Finally, we conclude in Section VI.

II. Prerequisites

A. Particle Interactions

Computing the interactions between N particles in a 3D environment is a common problem in scientific computing. For example, this is essential in fluid simulations using smoothed particle hydrodynamics [7] and in molecular dynamics for simulating forces between atoms [8]. These interactions are usually modeled by a potential function that depends on the distance between the particles. A straightforward way to compute the interactions is to evaluate the potential function for all pairs of particles. However, the potential function can be short-range or long-range. When the potential function is short-range, the interactions can be computed with a cutoff distance, i.e. the interactions are only computed between particles that are closer than a given distance. This reduces the complexity of the interactions from $O(N^2)$ to $O(N)$, where N is the number of particles, but this is only possible if we have an efficient way to find the particles' neighbors. Moreover, the positions of the particles are usually updated after each computation step. Consequently, the system used to find the interactions between the particles should be updated at each iteration of the simulation.

A possible solution to get the interaction list is to build a grid of cells mapped over the simulation space, where each cell contains the particles that are inside. The cells have a width equal to the cutoff distance C. Then, for each particle, the interactions are computed with the particles that are in the same cell and in the neighboring cells. Building the grid of cells and finding the interaction lists for each particle can be done in $O(N)$: we start by computing the cell index for each particle, then we order the particles in a new array by assigning a unique index per particle using atomic operations, and finally, we move the particles to a new array [9]. Each of these three operations can be implemented with a parallel loop over the N particles.

B. Graphics Processing Units

Graphics processing units (GPUs) are massively parallel processors that allow computations to be performed on thousands of cores. The hardware design of GPUs has been optimized for graphics rendering, particularly for the rendering of 3D scenes. To this end, GPUs have features dedicated to 3D rendering, such as texture mapping, rasterization, and ray tracing. Internally, GPUs are organized in a hierarchy of processing units, including streaming multiprocessors (SMs), warp schedulers, and execution units.

NVIDIA has proposed the CUDA programming model to develop parallel applications for GPUs. CUDA is designed to express algorithms in a way that can be mapped to the GPUs' hardware organization. For instance, thread blocks are distributed across SMs, and threads are executed in warps. There are also keywords and functions to use shared memory, constant memory, and texture memory. Thus, creating optimized applications for GPUs requires an understanding of the hardware architecture of GPUs and potentially adapting algorithms to their specificities.

C. Ray Tracing

Among the many features of GPUs dedicated to 3D rendering, ray tracing stands out as a powerful technology widely used in video games for rendering realistic 3D scenes [10]. It is a hardware-accelerated technique that computes the interactions of rays with objects in a scene. Ray tracing generates pixel colors in an image by casting rays from the camera into the scene. For each intersection of a ray with an object, the pixel's color is determined based on the object's material properties and lighting conditions. Rays can also be reflected or refracted by objects, or continue through non-opaque surfaces, enabling recursive computation of interactions for enhanced realism.

Typically, ray tracing kernels are implemented within the shaders of the graphics pipeline. Shaders are small programs executed on the GPU for each pixel in an image, often written in specialized languages such as GLSL (for OpenGL) or HLSL (for DirectX). These shaders run in parallel on the GPU, allowing for concurrent computation of ray-object interactions. NVIDIA introduced a method to utilize ray tracing within the CUDA programming model through its OptiX library [11]. With OptiX, developers retain the flexibility of CUDA programming while leveraging ray tracing technology, albeit with some constraints on kernel implementation.

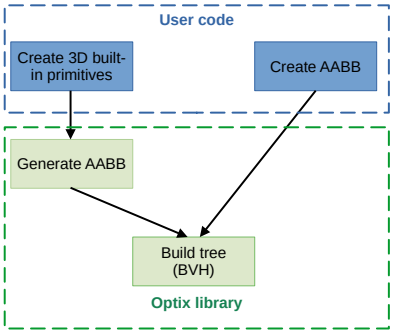


Fig. 1. Schematic view of BVH tree creation. The OptiX library requires a list of encompassing axis-aligned bounding boxes (AABBs), which can either be directly provided or generated from a list of basic primitives.

In ray tracing, scenes are represented using geometric primitives such as triangles or spheres, which are encapsulated within bounding volumes to enhance computational efficiency [12]. A common bounding volume is the Axis-Aligned Bounding Box (AABB), a rectangular box aligned with the coordinate axes and defined by its minimum and maximum bounds $(x_{min}, y_{min}, z_{min})$ and $(x_{max}, y_{max}, z_{max})$. AABB-ray intersection tests are conducted by calculating t_{min} and t_{max} for each axis and ensuring that overlaps exist across all axes, efficiently determining whether the ray intersects the box.

To further optimize performance, AABBs are organized into a Bounding Volume Hierarchy (BVH), a tree-like data structure where each node represents an AABB. Internal nodes group child AABBs, while leaf nodes encapsulate the actual primitives. In OptiX, users can either provide a list of basic primitives or directly supply a list of AABBs, as illustrated in Fig. 1.

During ray traversal, the algorithm performs intersection tests at each BVH node. If a ray misses an AABB, the entire

subtree beneath that node is skipped, reducing unnecessary computations. When a ray reaches a leaf node, precise intersection tests are conducted with the enclosed primitives, and the closest intersection is recorded. This process reduces the complexity of ray-scene intersections from linear to logarithmic by pruning large portions of the scene, ensuring only relevant branches of the BVH are explored.

The final output of the ray tracing algorithm, typically the closest hit point, is used for shading or further processing to determine the visual appearance of the scene. This hierarchical approach, combined with the use of AABBs and BVH, ensures that ray tracing remains computationally feasible even for complex scenes containing millions of primitives.

Fig. 2 provides a summary of the different operations in ray tracing using OptiX. In the first step, the user casts rays in the desired direction, and OptiX manages the BVH tree traversal and potential interactions. When a ray enters an enclosing bounding box (AABB), a callback function is invoked to determine whether the ray actually intersects the corresponding primitive, based on a built-in or user-defined Intersection Shader (IS), referred to as the intersection strategy. For custom primitives, the user must implement the IS. If the ray does not intersect or is marked to continue, the process is repeated for other primitives. The traversal stops when a final intersection is found or no more primitives can be intersected, triggering a call to the miss callback.

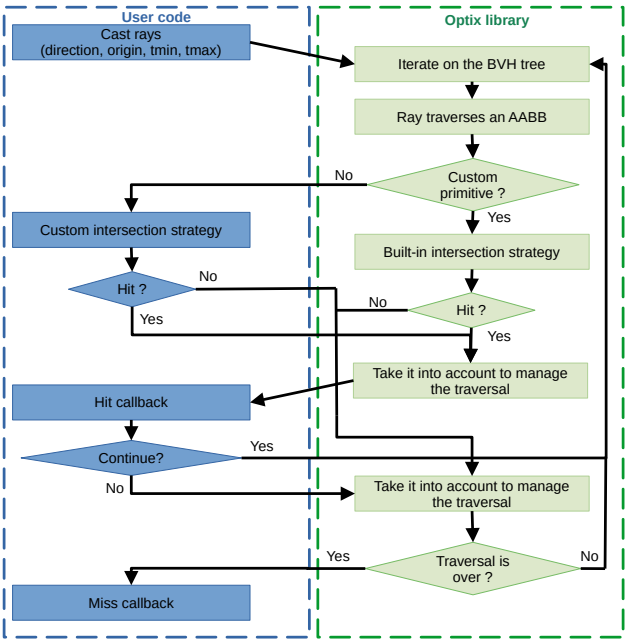


Fig. 2. Schematic view of the OptiX ray tracing workflow.

III. Related Work

A. Neighbor Search on GPU

The main work on physical simulation of particle interactions on GPU has been proposed by [13], to compute the gravitational potential. They described an efficient implementation using shared memory that became a standard implementation on GPU. As mentioned in Section II-A, grid-based and kd-tree structures are widely utilized for neighbor search operations on GPUs. For grid-based approaches, a detailed description can be found in our previous study [9], which focuses on scenarios with few particles per cell and demonstrates that utilizing shared memory often does not yield significant benefits. Regarding kd-tree implementations, [14] present a method for fast k-nearest neighbor searches using GPUs, highlighting the efficiency of kd-trees in high-dimensional spaces.

B. Ray Tracing in Computer Graphics

Ray tracing has been widely applied in particle-based representations, primarily in the domain of computer graphics and rendering. The author in [15] propose a method for efficiently ray tracing Gaussian particles to enable advanced rendering effects such as shadows, reflections, and depth of field in dense particle scenes, with applications in novel-view synthesis and visual realism. Similarly, [16] explores hardware-accelerated ray tracing for rendering particles that cast shadows, focusing on evaluating the performance of a prototype system. While these works demonstrate the utility of ray tracing in particle-based scenes, their focus lies in rendering and visualization, contrasting with our application of ray tracing for neighbor search in physical simulations.

C. OptiX in Scientific Computing

The NVIDIA OptiX framework has shown potential for diverse applications in scientific computing. Blyth et al. utilized OptiX to enable high-performance optical photon simulations in particle physics. This approach reduced memory and computation overheads by using GPU-based culling of photon hits, handling millions of photons in complex geometries [17], [18]. OptiX has been utilized as a flexible and high-performing tool for optical 3D modeling, enabling virtual measurements of sample surfaces by tracing over 1 billion light rays per image and comparing simulated results with those from physical measuring devices [19].

While these applications highlight OptiX's potential in handling computationally intensive tasks, they primarily focus on modeling physical processes rather than using ray tracing for spatial queries or particle interactions. Our work bridges this gap by leveraging OptiX for neighbor detection and particle interaction calculations, building on its demonstrated strengths in scientific modeling to expand its applicability into the domain of particle-based simulations.

D. OptiX for Neighbor Search

Using OptiX to find neighbors between elements has already been explored in several works. I. Evangelou et al. [20] introduced a novel approach to spatial queries, particularly radius search, by leveraging GPU-accelerated ray-tracing frameworks with OptiX. Instead of traditional spatial data structures like kd-trees, the authors proposed mapping the radius-search task to the ray-tracing paradigm by treating query points as primitives within a bounding volume hierarchy (BVH). In the proposed method, a ray used in the radius-search operation is essentially infinitesimal in extent, and its purpose is not to compute a traditional intersection but to test whether the origin of the ray (the query point) is within the bounding volume of a "sphere" surrounding each sample point. For that purpose, the

authors reimplemented the function that tests if a ray intersects a sphere to control how the tests are performed. This approach enabled significant performance gains in dynamically updated datasets.

Yuhao Zhu [21] further advanced the application of ray tracing for neighbor search by introducing optimizations for mapping the neighbor search problem onto the ray-tracing hardware available in modern GPUs. The author identified two key performance bottlenecks: unmanaged query-to-ray mapping, which led to control-flow divergences, and excessive tree traversals stemming from monolithic BVH construction. To address these issues, they proposed query scheduling and partitioning strategies that exploit spatial coherence and reduce BVH traversal time. Their experiments demonstrated substantial speedups ranging from 2.2 to 65.0 over existing GPU neighbor search libraries. Their work highlights the potential of using ray-tracing hardware not only for rendering but also for efficient spatial queries. The source code for their work is available on GitHub; however, the implementation is not directly applicable to our purpose as it primarily focuses on radius search and constructing interaction lists. In contrast, our interest lies in directly computing the interactions. Furthermore, for large test cases involving particles distributed on a sphere, the performance of their method appears to be of the same order of magnitude as ours.

Shiwei Zhao et al. [22] extended the use of ray tracing for particle-based simulations by converting neighbor search into a ray tracing problem. Each particle was represented as a bounding box, similar to previous works, with tiny rays emitted to detect intersections and identify neighboring particles. By leveraging NVIDIA's RT cores alongside CUDA cores, they demonstrated 10% to 60% performance improvements over traditional cell-based methods in various particle-based simulations, including discrete element methods and smooth particle hydrodynamics. Their approach underscores the versatility of ray-tracing cores for accelerating computationally intensive neighbor search tasks across different domains.

These works collectively demonstrate the potential of using OptiX and ray-tracing hardware for efficient spatial queries, providing a foundation for further exploration in GPU-accelerated computational methods. However, these approaches lack a direct integration of physical interaction computations within the ray-tracing framework. Our approach addresses this gap by embedding cutoff distances directly into the intersection tests. Additionally, we introduce two novel methods for neighbor detection that utilize actual intersection computations, accommodating scenarios where custom IS are unsupported.

IV. Proposed Solutions

A. Overview

The core idea of our approach is to represent particles using geometric primitives and to use ray tracing to find the neighbors of each particle by detecting intersections with these primitives.

With this aim, we investigate three possibilities:

1) Custom AABB: in this case, we directly provide the list of englobing bounding boxes. Therefore, we also have to provide our own IS in which we do not compute real intersections but only check if a ray is inside an AABB. This strategy is the closest one to the state of the art.

2) Spheres: in this case, we use built-in sphere primitives, which allows us to use the built-in IS that check if a ray really intersect with the surface of a sphere. However, we had to create our own geometric algorithm to make it work.

3) Triangles: in this case, we use built-in triangle primitives to create squares, which allows to use the built-in IS that check if a ray really intersect with the surface of a triangles. As for the spheres, we had to create our own geomatric algorithm to make it work.

The spherical approach is simpler and more intuitive than using squares, but we are interested in evaluating if the squares made of triangles is more efficient as it is the most used primitive in 3D rendering. We also want to evaluate if the built-in IS is more efficient than our custom IS.

In the cases that rely on built-in IS, we use the following algorithmic pattern:

1) We build a geometric representation for each particle.
2) We cast rays from each particle in specific directions to find the neighbors, depending on the representation.
3) We filter the intersections to avoid computing the same interaction multiple times.
4) We compute the interactions between the particles that are closer than C.

To reduce overhead, we aim to use as few rays as possible and ensure they do not intersect with too many particles that are not within the distance C.

In terms of implementation, we use the OptiX library to develop the ray tracing kernels, which can be used in conjunction with the CUDA programming model. Specifically, in the OptiX API, we create a scene by providing a list of geometric primitives. We then provide a CUDA kernel that launches the rays, where each ray has an origin, a direction, a starting point, and an endpoint. Usually, one CUDA thread is used to launch one ray. Finally, a callback is invoked by OptiX when a ray intersects with a primitive or if no intersection is found between the starting and ending points (in 3D rendering, this usually means that the background color should be used).

The data accessible from the callback is limited. OptiX built-in IS can provide information about the intersection, such as the intersection point, the normal on the surface, the distance from the ray's origin, and the index of the primitive that was hit. Additionally, the user can pass information from the CUDA kernel that launches the rays to the callback using a payload. A payload is a user-defined data structure that is passed along with a ray as it traverses the scene. It allows the ray to carry information that can be read or modified. The number of payload variables is limited (usually 16 32-bit integers in recent versions).

When the hit callback is invoked, and we want the ray to continue, there are two possibilities. The first is to launch a new ray from the intersection point in a new direction. This is

done by storing the intersection distance in a payload variable, returning from the callback, and then launching a new ray from the intersection point using a loop in the CUDA kernel to reach the desired distance. The second possibility is to inform OptiX that we want to continue the traversal by calling a corresponding function in the intersection callback. This second approach is generally more efficient, as it avoids the overhead associated with launching new rays and and should be favored in practice.

Additionally, we cannot allocate memory in the callbacks, so we cannot build complex data structures, such as lists, to store intersection lists. Consequently, if we want to filter the intersections, we cannot fill an array with indices and check if an index exists in the array to ensure uniqueness; instead, we must use geometric properties to filter the intersections.

In the remainder of the section, we consider that the target particle is the particle for which we want to find the neighbors.

B. Spherical Representation

In this section, we consider the case where the particles are represented by spheres and we use the built-in IS. In OptiX, a sphere is a geometric primitive defined by its center and radius, and multiple spheres of the same radius can be instantiated in the scene, which is the approach we use.

We have the following objectives:

1) Expressing the radius of the spheres depending on the cutoff distance;
2) Defining the origins and directions of the rays;
3) Providing a mechanism to filter the intesection when the rays intersect multiple times with the same sphere.

In our model, we will use three rays for each particle, one in each direction of the coordinate system. Consider a sphere of radius C centered at the origin in a three-dimensional space. The points on this sphere that are at the farthest distance from the three coordinate axes are located in the corners of a cube inscribed within the sphere. For instance, one such point at distance C from the origin lies in the direction $(1, 1, 1)$. These 8 points, corresponding to the vertices of the cube, are all at a distance C from the origin, and we want to know how far they are from the coordinate axes. This can be calculated as follows: Since the points have coordinates where $|x| = |y| = |z|$, we use the equation of the sphere $x^2 + y^2 + z^2 = C^2$. If we take the point for which $x = y = z$, it gives $3x^2 = C^2$, resulting in $x = \frac{C}{\sqrt{3}}$. Therefore, each of these points is at a distance of $\frac{C\times\sqrt{2}}{\sqrt{3}}$ from any of the three coordinate axes.

We use this information to define the radius of the spheres and the length of our rays. The radius is set to $r = \frac{C\times\sqrt{2}}{\sqrt{3}}$. In this scenario, it is sufficient that the rays go up to $l = \frac{C}{\sqrt{3}}$ in each direction relatively to the particle's position (so a single ray goes from $-l$ to l). We provide a simplified 3D rendering of the spheres in Fig. 3 that illustrate our model.

However, if $l < r$, there are positions where the sphere could simply englobe the rays, and we would miss some intersections (when the source and target are closer than $r - l$ in the three dimensions). Therefore, we set $l = r$ and add an ϵ to the radius of the sphere to ensure that the rays intersect with the sphere in all cases, obtaining $r = \frac{C\times\sqrt{2}}{\sqrt{3}} + \epsilon$ and $l = \frac{C\times\sqrt{2}}{\sqrt{3}}$. The ϵ is a small value such that it must be impossible that two particles are closer than ϵ, or some intersections will be missed (see Appendix A for more details).

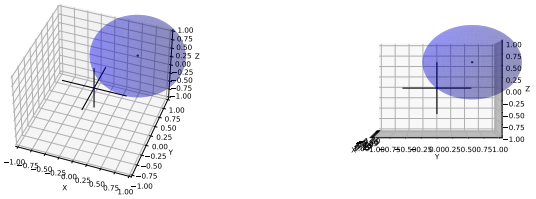


Fig. 3. 3D Spherical representation of a source and target particles distance from $C = 1$. The source sphere has a radius of $\frac{\sqrt{2}}{\sqrt{3}}$ and the rays, represented by segments, are of length $\frac{1}{\sqrt{3}}$ in each direction. In this case, the sphere could englobe the rays.

When the source and target particles are perfectly aligned on one axis, the ray will intersect for particles distant from the extremity by $r + \epsilon$. Therefore, two particle distant from $l + r + \epsilon$ can interact. This case and any intermediate situation where the source/target are actually too far can easily be filtered by checking the distance. In Fig. 4, we provide a 2D representation of the particles using spheres on intricate cases.

However, each ray can potentially intersect with a sphere several times, and the same sphere can be intersected by several rays of the same target particle, so we need to filter them. It is impossible to maintain a global list that all the rays can access to check if an intersection has already been found, or even a single list per ray to ensure that it does not intersect with the same sphere. Therefore, we must do this based on geometric rules as described in Algorithm 1. When we detect an intersection, we get the position of the source and target particles and first check they are within the cutoff distance. Then, we check if the closest ray to the source particle is the current one, and if yes, we can proceed with the computation (see Appendix A for more details).

C. Double Squares Representation

Most 3D rendering applications use triangles to represent the objects in the scene, which motivated us to create a second model that relies on triangles instead of spheres. Of course, building the AABB representations and the HBV tree is not expected to be faster than the sphere, but the hardware modules and built-in could be more optimize for the triangles, as it is more common primitives. In our model, we use four triangles to draw two squares, which are positioned opposite each other. Each square has a width of $C+\epsilon$ and is positioned at a distance of $C/2$ along the X-axis relative to the particle, one in each direction. We provide Code 1 in Appendix B, which shows how we generate the triangles from the particles' positions.

We then launch four rays, all with the same length and direction, but positioned at the corners of the squares. Each ray is positioned at $-C/2$ from the center of the square and has a length of $C + \epsilon$. The ϵ is used to ensure that when the source and target particles have the same x coordinate and their squares overlap, the rays cross the triangles. We provide

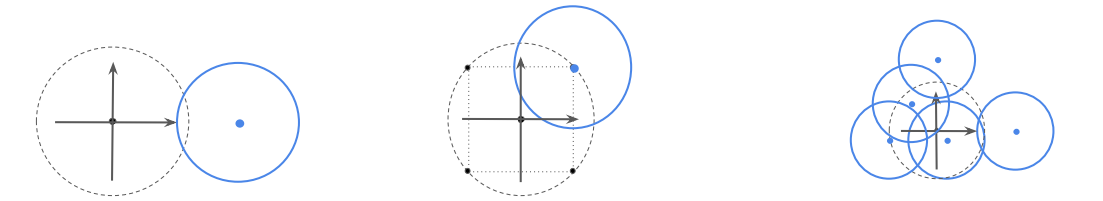


Fig. 4. 2D Spherical representation of the particles in three different cases. In the first one (left), what will be the position of the farthest source particle and how it will be detected by the ray. In the second one (center), we show the case where the source particle is the farthest from the rays (it also shows that in 2D the sphere radius could be smaller). In the last one (right), we show different source particles with their spheres and the rays that will intersect with them.

Algorithm 1: Sphere intersection callback

Data: Optix variables
Result: Callback when a ray intersect with a sphere

```
Function callback()
begin
    /* Get the center of the sphere (source position) */
    q ← optixGetSphereData ()
    /* Current particle position (target position) */
    point ← getPayloadPartPos()
    /* Compute differences and distances */
    diff_pos ← {abs(point.x − q.x), abs(point.y − q.y), abs(point.z − q.z)}
    diff_pos_squared ← {diff_pos.x², diff_pos.y², diff_pos.z²}
    dist_squared ← diff_pos_squared.x + diff_pos_squared.y + diff_pos_squared.z
    /* Get the cutoff distance from payload */
    c ← getPayloadC()
    /* Ensure it is in the cutoff distance */
    if dist_squared < c² then
        dist_axis_squared ← {diff_pos_squared.y + diff_pos_squared.z, diff_pos_squared.x+ diff_pos_squared.z, diff_pos_squared.x+ diff_pos_squared.y}
        ray_dir ← optixGetWorldRayDirection()
        closest_axis ← getClosestAxis(dist_axis_squared)
        closest_axis_is_ray_dir ← (ray_dir == closest_axis)
        /* Ensure this ray and this intersection are the good one */
        if closest_axis_is_ray_dir then
            /* Call computation kernel */
```

in Fig. 5 the 2D representation of the particles using squares (which are lines in 2D).

From this description, one particle can be seen as a box of sides $(C + \epsilon, C + \epsilon, C)$ (see Fig. 6). The rays can be seen as the four edges of the box in the x direction, and the triangles composed the front and back faces. If any two boxes have an intersection, we will detect it as shown in Fig. 5.

Potentially, the ray will intersect with the squares of the target particle, but this can easily be filtered by checking either the coordinates or the index of the geometric elements. Additionally, if the target and source particles are aligned on the y or z axis, two rays will intersect with the source's squares. To filter these intersections, we proceed as shown in Algorithm 2. We compare the coordinates between the source and the target and proceed as follows: If y and z are different, we perform the computation (only the current ray will intersect). If y is equal, we use the ray of index 0 if z is smaller, and the ray of index 2 if z is greater. If z is equal, we use the ray of index 0 if y is smaller, and the ray of index 1 if y is greater. Otherwise, only the ray of index 0 will be used for computation (all four rays will intersect).

D. Custom AABB

In this strategy, a bounding box with a width equal to $2 \times C$ is created around each particle. OptiX treats these boxes as custom primitives, requiring us to implement a custom IS that is invoked when a ray enters an AABB. We do not perform any intersection tests within the IS, as our goal is not to compute ray-AABB or ray-surface intersections but to identify pairs of particles that are within a distance C of each other. To achieve this, rays are cast from the particles' positions with an infinitesimally short length, as illustrated in Fig. 7.

The interaction between the source and target particles is computed directly within the IS callback. Consequently, the hit and miss callbacks are left empty, as there is no need to filter intersections. In the IS callback, OptiX provides information about the ray (specifically, the starting point, which corresponds to the position of the source particle) and the index of the AABB being tested. To obtain geometric information about the AABB, one could register a Primitive Geometry Acceleration Structure (PGAS) for each AABB during the OptiX scene build stage and retrieve this data in the IS callback. However, we observed that this approach significantly increased the scene build cost. Instead, we opted to access the global memory directly to retrieve the position of the target particle, which proved to be a more efficient solution.

V. Performance Study

This section presents a performance evaluation of the different methods under various configurations. The analysis begins with a description of the experimental setup, including

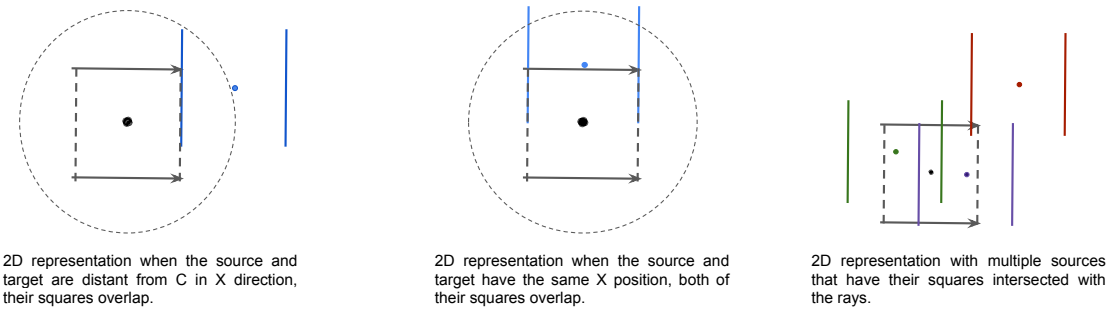


Fig. 5. Double squares (line) representation of the particles. On the left, we show how the squares can overlap for particles that are too far, but which can be easily filtered with the distance. In the middle, we show how particles that have the same x coordinate can have their squares that overlap. On the right, we show different source particles with their squares and the rays that will intersect with them.

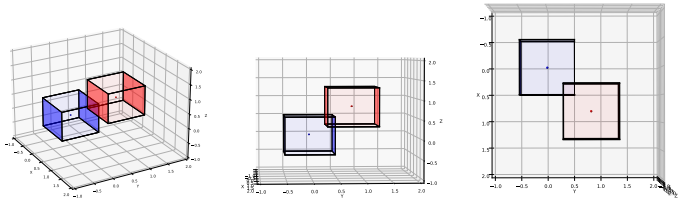


Fig. 6. 3D Rectangular representation of a source and target particles. Some rays (thick lines) intersect with the squares.

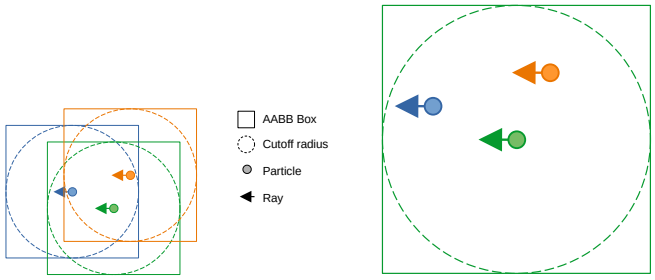


Fig. 7. Custom AABB representation of the particles. The rays cast from the particles' positions with a infinitesimal length. In the IS callback invoked if a ray enters an AABB, the distance is checked and the interaction between the source and target particles is computed. Self interactions are filtered by ensuring that the source and target are different.

the hardware and software used for implementation. Performance is then examined for two types of particle distributions: uniform and non-uniform. Differentiating these distributions allows for assessing the efficiency of grid-based methods like CUDA, which are optimized for uniform distributions but can encounter inefficiencies with sparse or highly localized data in non-uniform scenarios.

A. Experimental Setup

1) Hardware: We have used two NVidia GPUs:

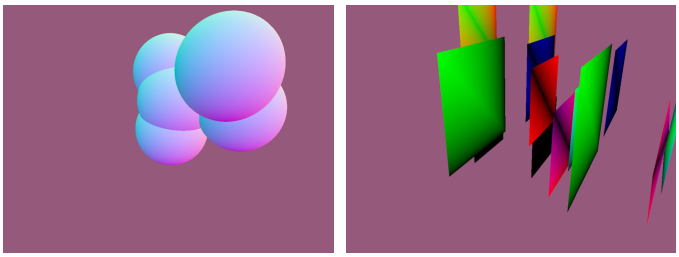
- A100[3] with 40GB hBM2, 48KB of shared-memory, 108 multi-processors, zero RT Cores, 8192 CUDA Cores max single-precision performance 19.5TFLOPS, and max tensor performance 311.84TFLOPS.
- RTX8000[4] with 48GB GDDR6, 48KB of shared-memory, 72 multi-processors, 72 RT Cores, 4608 CUDA Cores, maximum Ray casting of 10Giga Rays/sec, max single-precision performance 16.3TFLOPS, and max tensor performance 130.5TFLOPS.

Despite the lack of RT cores, the A100 is capable of executing ray tracing kernels, the GPU then use its other units to behave similarly.

2) Software: We have implemented the proposed approach in OptiX 8.[5] We use the GNU compiler 11.2.0 and the NVidia CUDA compiler 12.3. The source code is available online.[6]

The code was compiled with the following flags: `-arch=sm_75` for the RTX8000, `-arch=sm_80` for the A100 (and `-O3 -DNDEBUG` on both). We execute each kernel five times and take the average as reference.

We provide in Fig. 8 the 3D rendering of the primitives using ray tracing: in Fig. 8a for the spheres and in Fig. 8b for the squares. These figures were drawn using the OptiX API and ray tracing from camera to the scene.



(a) 3D Rendering of sphere primitives. (b) 3D Rendering of square primitives.

Fig. 8. 3D Rendering of the primitives using conventional ray tracing.

B. Uniform distribution

In this test case, the particles are distributed uniformly within a unit box. Consequently, there are no (or very few)

[3]https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-nvidia-us-2188504-web.pdf

[4]https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/quadro-product-literature/quadro-rtx-8000-us-nvidia-946977-r1-web.pdf

[5]https://developer.nvidia.com/rtx/ray-tracing/optix

[6]https://gitlab.inria.fr/bramas/particle-interaction-with-optix

Algorithm 2: Triangle intersection callback

Data: Optix variables
Result: Callback when a ray intersect with a triangle

Function callback()
begin
  /* Get information on the intersected triangle */
  $vertices \leftarrow$ `optixGetTriangleVertexData` $(gas, prim\_idx, sbtGASIndex, 0.f, vertices)$
  /* Retreive the source position from the triangle vertices */
  Declare q as float3
  $q.y \leftarrow (\max(vertices[0].y, vertices[1].y, vertices[2].y) + \min(vertices[0].y, vertices[1].y, vertices[2].y))/2$
  $q.z \leftarrow (\max(vertices[0].z, vertices[1].z, vertices[2].z) + \min(vertices[0].z, vertices[1].z, vertices[2].z))/2$
  $c \leftarrow$ `getPayloadC()`
  if $(prim\_idx \bmod 4) < 2$ **then**
    $q.x \leftarrow vertices[0].x + c/2$
  else
    $q.x \leftarrow vertices[0].x - c/2$
  /* Get target particle position */
  $point \leftarrow$ `getPayloadPartPos()`
  /* Compute distance */
  $dist\_p1\_p2 \leftarrow$ `distance`$(point, q)$
  /* Ensure it is not a self intersection and it is in the cutoff distance */
  if $dist\_p1\_p2 < cANDdist\_p1\_p2 > \epsilon$ **then**
    $ray\_idx \leftarrow$ `getPayloadRayidx()`
    $is\_ray\_for\_compute \leftarrow (point.y \neq q.y$ AND $point.z \neq q.z)$
    OR $((point.z < q.z$ AND $ray\_idx == 0)$ OR $(point.z > q.z$ AND $ray\_idx == 2))$
    OR $((point.y < q.y$ AND $ray\_idx == 0)$ OR $(point.y > q.y$ AND $ray\_idx == 1))$
    OR $ray\_idx == 0$
    if $is\_ray\_for\_compute$ **then**
      /* Perform the computation */

empty cells in the grid of the CUDA version. The cutoff distance is defined as $1/\beta$, where β can be 2, 4, 8, 16, or 32. For a given β, the simulation grid consists of 3^β cells. The number of particles N is then calculated as $N = p \times 3^\beta$, where p represents the average number of particles per cell, taking values of 1, 2, 4, 8, 16, or 32. All computations are performed in single-precision floating point.

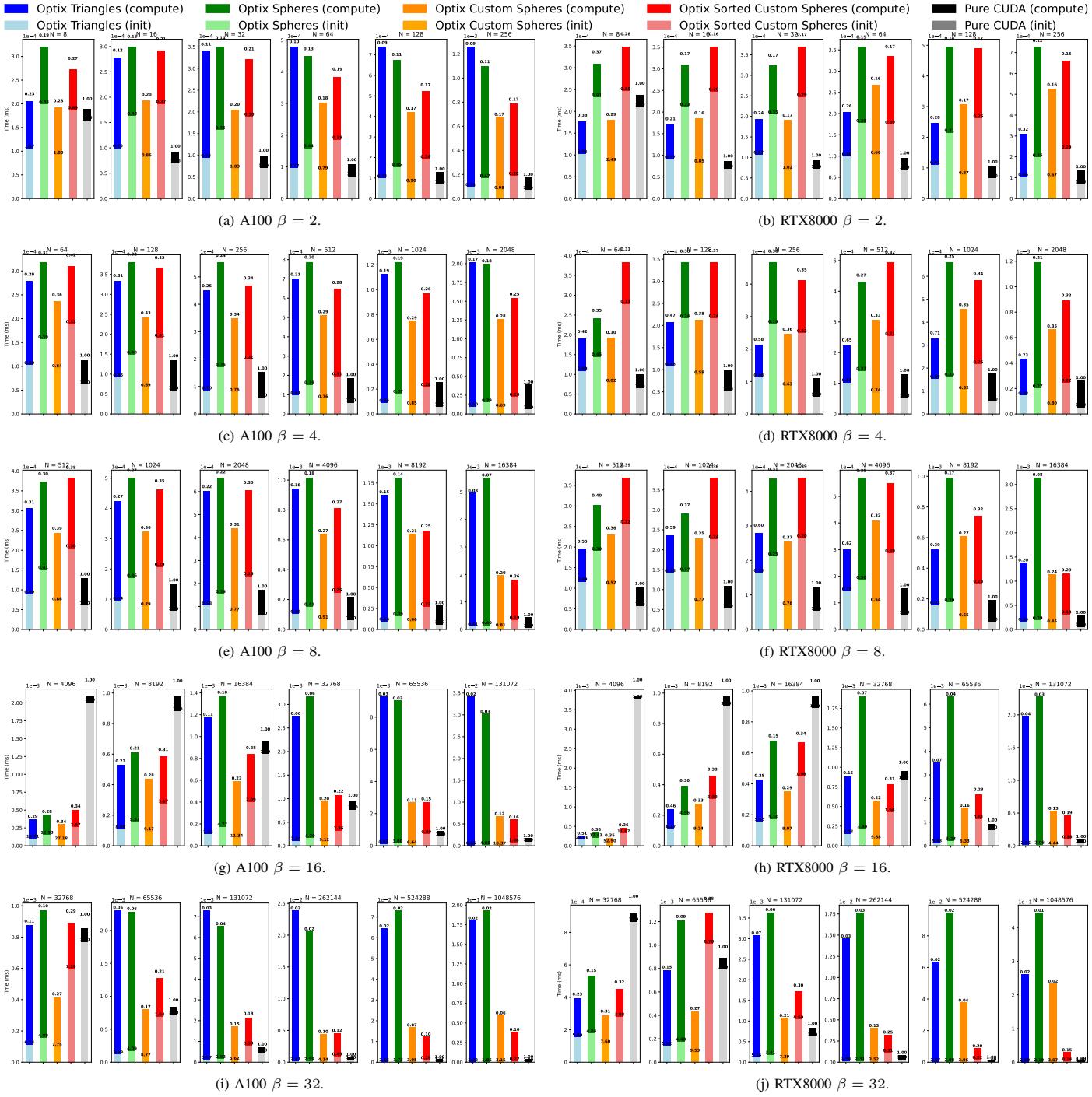
We present the results in Fig. 9. For all configurations, we measured the initialization step (light color) and the computation step (dark color).

For the OptiX-based implementation, the initialization step involves building the scene by invoking the OptiX API to create the primitives. For the CUDA version, the initialization step involves constructing the grid of cells. Consequently, in the OptiX-based version, the computation step includes the time spent launching the rays, executing the callback functions, and performing the interactions. In the CUDA version, the computation step corresponds to the kernel time required to compute the interactions.

1) Comparison between triangles and spheres: First, we analyze the performance difference between triangles (blue) and spheres (green). On the A100 GPU, both models deliver similar performance, but the ratio of initialization time (light color) to computation time (dark color) is higher for spheres. This indicates that the computation step is more efficient for spheres than for triangles. This suggests that in scenarios where the initialization step is performed only once (e.g. static elements), spheres might be a better choice. However, on the RTX8000, triangles are faster than spheres across all configurations. Although the initialization step for spheres is quicker than for triangles in cases with fewer particles, this trend does not hold for larger configurations.

2) Comparison between custom AABB and built-in primitives: Next, we compare the Custom AABB method (orange/red) with the built-in primitives (triangles in blue and spheres in green). On the A100, the Custom AABB consistently outperforms the built-in primitives, and its advantage grows as the number of particles or cells increases. For $\beta = 32$ (Fig. 9i), the Custom AABB is up to four times faster than the triangle- or sphere-based methods. On the RTX8000, the Custom AABB performs similarly to triangles for small particle counts but becomes faster as the number of particles increases. Overall, the Custom AABB is the fastest method. This demonstrates that creating a custom Intersection Shader (IS) that is significantly simpler and lighter than the built-in ones, which compute actual intersections, can accelerate execution.

3) Benefit of sorting the particles in the custom AABB methods: Sorting the particles to position them closer in memory based on their spatial proximity in the simulation is expected to optimize memory accesses. However, sorting the particles also incurs a computational cost. We observe that the strategy without sorting (orange) is faster than the strategy with sorting (red) when there are few particles. However, as the number of particles increases, the benefits of sorting become significant. In our test case, for a given β, increasing the number of particles leads to an increase in neighboring particles and, consequently, the number of interactions. In such cases, optimizing memory access becomes critical.



(a) A100 $\beta = 2$.

(b) RTX8000 $\beta = 2$.

(c) A100 $\beta = 4$.

(d) RTX8000 $\beta = 4$.

(e) A100 $\beta = 8$.

(f) RTX8000 $\beta = 8$.

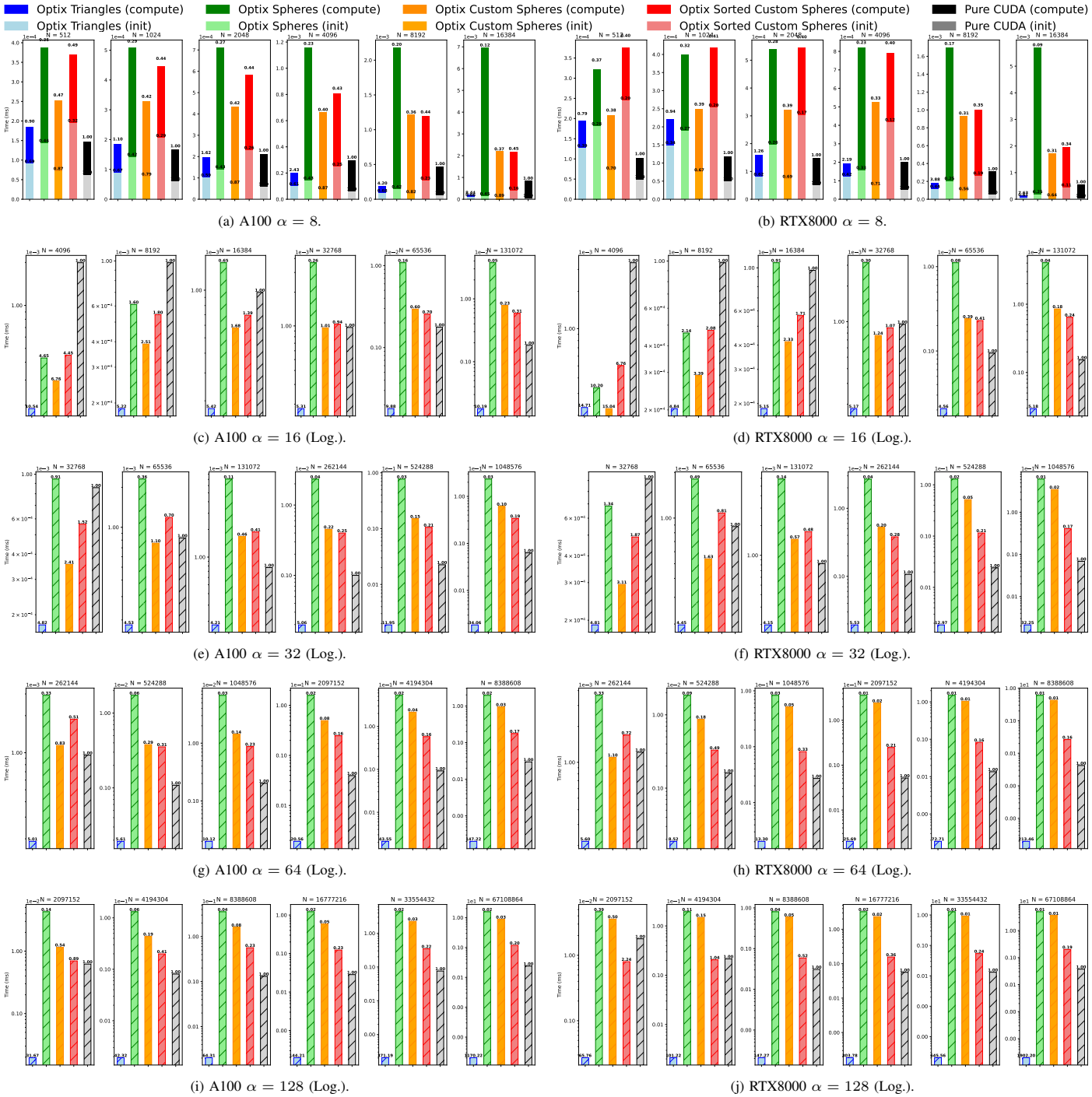
(g) A100 $\beta = 16$.

(h) RTX8000 $\beta = 16$.

(i) A100 $\beta = 32$.

(j) RTX8000 $\beta = 32$.

Fig. 9. Performance results for the two GPUs for our approaches (OptiX spheres, OptiX triangles, and Custom AABB without and with sort) and the pure CUDA implementation that use a grid of cells (CUDA). β is the divisor coefficient of the simulation box. The cutoff distance is $1/\beta$ and there are β^3 cells in the grid in the Cuda version. The speedup against the CUDA version is shown above the bars for both the build and compute steps.



(a) A100 $\alpha = 8$.

(b) RTX8000 $\alpha = 8$.

(c) A100 $\alpha = 16$ (Log.).

(d) RTX8000 $\alpha = 16$ (Log.).

(e) A100 $\alpha = 32$ (Log.).

(f) RTX8000 $\alpha = 32$ (Log.).

(g) A100 $\alpha = 64$ (Log.).

(h) RTX8000 $\alpha = 64$ (Log.).

(i) A100 $\alpha = 128$ (Log.).

(j) RTX8000 $\alpha = 128$ (Log.).

Fig. 10. Performance results for the two GPUs for our approaches (OptiX spheres, OptiX triangles, and Custom AABB without and with sort) and the pure CUDA implementation that use a grid of cells (CUDA). α is the divisor coefficient of the simulation box. The cutoff distance is lower than $1/\beta$ and there are around α^3 non-empty cells in the grid in the Cuda version. The speedup against the CUDA version is shown above the bars for both the build and compute steps, or for the complete run.

4) GPU Performance comparison: We then compare the performance of the two GPUs. Results for the A100 are presented in Fig. 9a, 10c, 9e, 9g, and 9i, while results for the RTX8000 are shown in Fig. 9b, 9d, 9f, 9h, and 9j. Both GPUs exhibit comparable performance and a similar initialization-to-computation time ratio. Although the A100 is expected to offer higher raw performance, the OptiX implementation is not fully optimized for this GPU, while the RTX8000 benefits from having more RT cores to accelerate computation.

Additionally, we do not utilize the tensor cores of the A100, which represent a key differentiator between the two GPUs and could help approach theoretical performance.

5) Comparison between OptiX and CUDA implementations: Finally, we compare the OptiX triangles (blue), OptiX spheres (green), Custom AABB (orange), Custom sorted AABB (red), and CUDA implementation (green). The OptiX-based implementation is generally slower than the CUDA implementation for β values from 2 to 8, except in a few cases (e.g. $\beta = 2$ and $N = 8$). In these scenarios, the initialization step in the CUDA implementation is negligible, and the computation step dominates. Consequently, avoiding the grid of cells does not yield a significant advantage. For $\beta = 8$ (Fig. 9e and 9f), the CUDA computation step is shorter because the OptiX implementations include not only the interaction computation but also the intersection list computation using rays. Additionally, the memory access pattern in OptiX-based implementations differs, as spatially close particles in memory may not be close in space and may have varying neighbors.

For $\beta = 16$ and $\beta = 32$, the Custom AABB performs better for configurations with few particles per cell (e.g. one or two particles per cell on average in the CUDA version). In these cases, the computation cost is low, and the CUDA version spends most of its time in the initialization step. Furthermore, the CUDA version incurs overhead for empty cells because it constructs a dense grid. For such configurations, the OptiX-based implementations allocate more time to the computation step, making the Custom AABB a suitable choice when the number of particles per cell is low, particularly in scenarios where moving particles require frequent grid rebuilding.

C. Non Uniform Distribution

In this test case, we distribute the particles on the surface of a unit sphere. Consequently, most of the cells on the grid in the CUDA version are empty. With this aim, we use a coefficient α, which can be 8, 16, 32, 64 or 128. For a given α, the simulation grid consists of approximately 3^α non-empty cells. The number of particles N is then calculated as $N = p \times 3^\alpha$, where p represents the average number of particles per cell, taking values of 1, 2, 4, 8, 16, or 32. The cutoff radius is set such that each particle should have approximately $9 \times p$ neighbor particles (it will be in general smaller than $2/\alpha$). In order to facilitate the reproducibility of our test case, we provide in Appendix, the Code 2 that generates the different configurations. All computations are performed in single-precision floating point.

We present the results in Fig. 10. For all configurations, we measured both the initialization step (light color) and the computation step (dark color) for the first configuration. However, for the others, we directly present the total execution time using a logarithmic scale on the y-axis due to extreme differences.

For the OptiX-based implementation, the initialization step involves building the scene by invoking the OptiX API to create the primitives. For the CUDA version, the initialization step involves constructing the grid of cells. Consequently, in the OptiX-based version, the computation step includes the time spent launching the rays, executing the callback functions, and performing the interactions. In the CUDA version, the computation step corresponds to the kernel time required to compute the interactions.

1) OptiX spheres: The OptiX spheres method is consistently the slowest (except for $\alpha = 2$ and a small number of particles). It has a more expensive initialization step compared to the Custom AABB methods and a significantly larger computation step. This is because the built-in IS kernel for spheres is computationally intensive, as it calculates several values, such as the norm of the intersection, that we do not use since we are not rendering an image.

2) Custom AABB without and with sorting: For low α, sorting does not provide an advantage; it increases the cost of the initialization step too much relative to the gains in the computation step. However, for high α and a large number of particles, sorting the particles results in a performance improvement. This demonstrates that random access to global memory is so costly that the overhead of sorting the particles is worthwhile. The Custom AABB method with sorting performs comparably to the CUDA version. However, the Custom AABB methods are generally slower than both the CUDA version and the OptiX triangles method.

3) OptiX triangles: This strategy is the most efficient. For $\alpha = 8$, its initialization step is comparable in cost to the OptiX spheres and Custom AABB methods. However, its computation step is significantly faster. We attribute this to two primary reasons. First, the built-in IS kernel for triangles is likely much simpler than that for spheres and is probably heavily optimized by NVIDIA, as triangles are the most commonly used primitive in 3D rendering. Second, the AABB bounding boxes around triangles are much smaller than those around spheres or the custom AABB primitives. This enables faster tree traversal and fewer false positives (i.e. cases where a ray intersects a bounding box but not the primitive inside it).

Finally, the CUDA version includes several components in its implementation that rely on algorithms with complexity linear in the number of cells. Since the vast majority of these cells are empty, these steps become highly inefficient. This is particularly evident for $\alpha = 8$, where the initialization step is notably prominent. Moreover, if we were to create a test case with $\alpha = 256$ (not included in the current study), the GPU would run out of memory for the the CUDA version allocating an excessively large grid. This limitation makes alternative approaches based on a tree structure — such as our OptiX-based methods — the only viable options.

VI. Conclusion

In this paper, we proposed leveraging ray tracing technology to compute particle interactions within a cutoff distance in a 3D environment. We introduced one method that uses

custom primitives and a custom Intersection Shader (IS), similar to the state-of-the-art, and two methods that use built-in primitives and actual intersection computations. For the latter, we described geometric algorithms to build the interaction list based on ray intersections with spheres or triangles.

Our results indicate that our approach provides a modest advantage in the preprocessing stage by avoiding the construction of a grid of cells. In addition, it is slower than the classical approach during the computation step when dealing with large numbers of uniformly distributed particles. However, when most cells are empty, our approach can provide a significant speedup. Therefore, we believe these methods have the potential to deliver better performance in the future or for specific applications, and we hope our work inspires the community to explore this direction further.

As GPU architectures continue to evolve, with advancements such as NVIDIA's Ada Lovelace architecture featuring third-generation RT cores and AMD's RDNA 3 architecture incorporating second-generation ray-tracing accelerators, we anticipate that algorithms leveraging these capabilities will become increasingly effective and relevant. Consequently, our approach is prepared to benefit from these hardware improvements, enhancing its performance and applicability in future computational scenarios and we hope our work inspires the community to explore this direction further.

We plan to evaluate our approach on other GPUs, such as AMD Radeon, and aim to port our method even in cases where the IS cannot be customized, thanks to our geometric algorithms. In a second step, we will take our best strategy and compare it with state-of-the-art particle interaction solvers that support GPUs.

ACKNOWLEDGMENT

Experiments presented in this paper were carried out using the PlaFRIM experimental test-bed.

REFERENCES

[1] C. Kutzner, S. Páll, M. Fechner, A. Esztermann, B. L. de Groot, and H. Grubmüller, "More bang for your buck: Improved use of gpu nodes for gromacs 2018," *Journal of Computational Chemistry*, vol. 40, no. 27, pp. 2418–2431, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.26011

[2] M. Zubair, D. Ranjan, A. Walden, G. Nastac, E. Nielsen, B. Diskin, M. Paterno, S. Jung, and J. H. Davis, "Efficient gpu implementation of automatic differentiation for computational fluid dynamics," in *2023 IEEE 30th International Conference on High Performance Computing, Data, and Analytics (HiPC)*. IEEE, 2023, pp. 377–386.

[3] B. Rybakin and V. Goryachev, "Parallel algorithms for astrophysics problems," *Lobachevskii Journal of Mathematics*, vol. 39, pp. 562–570, 2018.

[4] A. Ilievski, V. Zdraveski, and M. Gusev, "How cuda powers the machine learning revolution," in *2018 26th Telecommunications Forum (TELFOR)*. IEEE, 2018, pp. 420–425.

[5] H. Friedrich, J. Günther, A. Dietrich, M. Scherbaum, H.-P. Seidel, and P. Slusallek, "Exploring the use of ray tracing for future games," in *Proceedings of the 2006 ACM SIGGRAPH Symposium on Videogames*, 2006, pp. 41–50.

[6] NVIDIA, "Nvidia ada gpu architecture," 2023, version 2.1 https://images.nvidia.com/aem-dam/Solutions/Data-Center/l4/nvidia-ada-gpu-architecture-whitepaper-v2.1.pdf.

[7] D. Koschier, J. Bender, B. Solenthaler, and M. Teschner, "Smoothed particle hydrodynamics techniques for the physics based simulation of fluids and solids," *arXiv preprint arXiv:2009.06944*, 2020.

[8] M. S. Badar, S. Shamsi, J. Ahmed, and M. A. Alam, "Molecular dynamics simulations: concept, methods, and applications," in *Transdisciplinarity*. Springer, 2022, pp. 131–151.

[9] D. Algis, B. Bramas, E. Darles, and L. Aveneau, "Efficient GPU Implementation of Particle Interactions with Cutoff Radius and Few Particles per Cell," in *International Symposium on Parallel Computing and Distributed Systems (PCDS2024)*. Singapore, Singapore: IEEE, Sep. 2024. [Online]. Available: https://inria.hal.science/hal-04621128

[10] G. IGN, "Top 5 Games That Make the Best Use of NVIDIA's RTX Technologies," https://www.ign.com/articles/top-5-games-that-make-the-best-use-of-nvidias-rtx-technologies, Dec. 2023.

[11] S. G. Parker, J. Bigler, A. Dietrich, H. Friedrich, J. Hoberock, D. Luebke, D. McAllister, M. McGuire, K. Morley, A. Robison, and M. Stich, "Optix: a general purpose ray tracing engine," *ACM Trans. Graph.*, vol. 29, no. 4, jul 2010. [Online]. Available: https://doi.org/10.1145/1778765.1778803

[12] P. Shirley, I. Wald, T. Akenine-Möller, and E. Haines, "Ray tracing gems: High-quality and real-time rendering with dxr and other apis." Berkeley, CA: Apress, 2019, ch. What is a Ray?, pp. 15–19. [Online]. Available: https://doi.org/10.1007/978-1-4842-4427-2_2

[13] L. Nylons, M. Harris, and J. Prins, "Fast n-body simulation with cuda," *GPU gems*, vol. 3, pp. 62–66, 2007.

[14] V. Garcia, E. Debreuve, and M. Barlaud, "Fast k nearest neighbor search using gpu," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2008, pp. 1–6.

[15] N. Moenne-Loccoz, A. Mirzaei, O. Perel, R. de Lutio, J. Martinez Esturo, G. State, S. Fidler, N. Sharp, and Z. Gojcic, "3d gaussian ray tracing: Fast tracing of particle scenes," vol. 43, no. 6, Nov. 2024. [Online]. Available: https://doi.org/10.1145/3687934

[16] L. Lindau, "Hardware accelerated ray tracing of particle systems," 2020.

[17] S. Blyth, "Opticks: Gpu optical photon simulation for particle physics using nvidia® optixtm," in *EPJ Web of Conferences*, vol. 214. EDP Sciences, 2019, p. 02027.

[18] ——, "Integration of juno simulation framework with opticks: Gpu accelerated optical propagation via nvidia® optix™," in *EPJ Web of Conferences*, vol. 251. EDP Sciences, 2021, p. 03009.

[19] A. Keksel, S. Schmidt, D. Beck, and J. Seewig, "Scientific modeling of optical 3d measuring devices based on gpu-accelerated ray tracing using the nvidia optix engine," in *Modeling Aspects in Optical Metrology IX*, vol. 12619. SPIE, 2023, pp. 117–125.

[20] I. Evangelou, G. Papaioannou, K. Vardis, and A. A. Vasilakis, "Fast radius search exploiting ray tracing frameworks," *Journal of Computer Graphics Techniques (JCGT)*, vol. 10, no. 1, pp. 25–48, February 2021. [Online]. Available: http://jcgt.org/published/0010/01/02/

[21] Y. Zhu, "Rtnn: accelerating neighbor search using hardware ray tracing," in *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, ser. PPoPP '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 76–89. [Online]. Available: https://doi.org/10.1145/3503221.3508409

[22] S. Zhao, Z. Lai, and J. Zhao, "Leveraging ray tracing cores for particle-based simulations on gpus," *International Journal for Numerical Methods in Engineering*, vol. 124, no. 3, pp. 696–713, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/nme.7139

APPENDICES

A. DISCUSSION ON THE SPHERE MODEL

The filtering algorithm presented in Section IV-B is based on the assumption that there is always at least one ray that is intersected once by the sphere and that it is the closest one to the center of the sphere. The following section is dedicated to demonstrating this hypothesis.

In Fig. 11, we show the different possibilities depending on the radius r and ray's length l. As it can be seen, even if the sources located at a distance of C from the target could have their spheres that intersect with the rays when $r > l$, we must set $l = r$ to ensure that the rays will intersect with the

sphere in all cases, especially when the source and target are close.

For the clarity of the proof, a demonstration is first provided in the plane, followed by a generalization to $\mathbb{R}^3$.

1) 2D Case: Let us consider a cercle of radius r centered at (x_c, y_c), and a cross[7] centered at the origin with a length of r. We consider the cases where the cercle is located in the first quarter, i.e., $0 \leq x_s \leq r$ and $0 \leq y_s \leq r$, but the demonstration remains valid for other quarters.

The equation of a cercle is given by:

$$(x - x_c)^2 + (y - y_c)^2 = r^2. \tag{1}$$

The coordinates of the intersection points of the cercle with the axis are given by:

$$\begin{aligned} x_0 &= x_c - \sqrt{r^2 - y_c^2} \quad \text{and} \quad x_1 = x_c + \sqrt{r^2 - y_c^2} \\ y_0 &= y_c - \sqrt{r^2 - x_c^2}. \quad \text{and} \quad y_1 = y_c + \sqrt{r^2 - x_c^2}. \end{aligned} \tag{2}$$

We provide the Fig. 12 that shows where these points are located on the sphere and the cross.

x_0 and y_0 are the coordinates of the intersection points of the cercle that remain on the cross as the cercle get away from the origin. On the other hand, x_1 and y_1 are the coordinates of the intersection points of the cercle that are the farthest from the origin of the cross and that can potentially be too far to remain on the cross (they will be on the corresponding axis but behind l).

Lemma 1. *Given a circle $\mathcal{C}$ and let e_x and e_y be the two segments of the cross of length less than r, then the number of intersections of $\mathcal{C}$ with e_x or with e_y is strictly less than 2.*

Proof: As we consider that it is impossible that both x_1 and y_1 exist at the same time, we consider that these two equations cannot be true at the same time

$$x_c + \sqrt{r^2 - y_c^2} \leq r \quad \text{and} \quad y_c + \sqrt{r^2 - x_c^2} \leq r. \tag{3}$$

Which can be simplified as

$$\begin{aligned} x_c < r - \sqrt{r^2 - y_c^2} \quad \text{and} \quad & \sqrt{r^2 - x_c^2} \leq r - y_c \\ & r^2 - x_c^2 \leq r^2 - 2ry_c + y_c^2 \\ & x_c^2 > 2ry_c - y_c^2 \\ & x_c > \sqrt{2ry_c - y_c^2}, \end{aligned} \tag{4}$$

ending up with

$$\sqrt{2ry_c - y_c^2} < r - \sqrt{r^2 - y_c^2}. \tag{5}$$

For our definition range of $y_c \in [0, r]$, this equation has no solution (see Fig. 13), which confirms that $\mathcal{C}$ will always intersect with the cross at most once for e_x or e_y.

[7]That is, two orthogonal axis-aligned segments of the same size intersecting at their respective centers.

The second statement is to show that the ray that is intersected once is the closest to the center of the sphere.

Lemma 2. *Given a circle $\mathcal{C}$ and e one of the two segments of the cross of length less than r, such as the number of intersection of $\mathcal{C}$ with e is equal to 1, then the distance of e with $\mathcal{C}$ is smaller than the distance of the other bar to $\mathcal{C}$.*

Proof: Consider that $\mathcal{C}$ intersects with e_y twice and e_x once, it means that $y_1 < r$ and $x_1 > r$, i.e. $y_c + \sqrt{r^2 - x_c^2} < r$ and $x_c + \sqrt{r^2 - y_c^2} > r$.

We aim to demonstrate that the inequality

$$y_c + \sqrt{r^2 - y_c^2} < x_c + \sqrt{r^2 - x_c^2} \tag{6}$$

holds if and only if $x_c > y_c$.

We start with the inequality:

$$y_c + \sqrt{r^2 - y_c^2} < x_c + \sqrt{r^2 - x_c^2}. \tag{7}$$

Subtracting y_c from both sides, we obtain:

$$\sqrt{r^2 - y_c^2} < x_c - y_c + \sqrt{r^2 - x_c^2}. \tag{8}$$

We can further simplify this to:

$$\sqrt{r^2 - y_c^2} - \sqrt{r^2 - x_c^2} < x_c - y_c. \tag{9}$$

The inequality now compares two quantities: $\sqrt{r^2 - y_c^2} - \sqrt{r^2 - x_c^2}$ and $x_c - y_c$.

- The term $\sqrt{r^2 - y_c^2}$ represents the horizontal distance from the point (x_c, y_c) to the vertical axis.
- The term $\sqrt{r^2 - x_c^2}$ represents the vertical distance from the point (x_c, y_c) to the horizontal axis.

Let's consider the case where $x_c > y_c$:

- If $x_c > y_c$, then $x_c - y_c > 0$.
- Additionally, $\sqrt{r^2 - y_c^2} > \sqrt{r^2 - x_c^2}$ because $y_c < x_c$.

This implies that the term $\sqrt{r^2 - y_c^2} - \sqrt{r^2 - x_c^2}$ is positive, and it is less than $x_c - y_c$, proving that the inequality holds under this condition.

Thus, for the inequality $y_c + \sqrt{r^2 - y_c^2} < x_c + \sqrt{r^2 - x_c^2}$ to hold, it is necessary that $x_c > y_c$. So, the x axis is the closest axis to the center of the sphere and is intersected once.

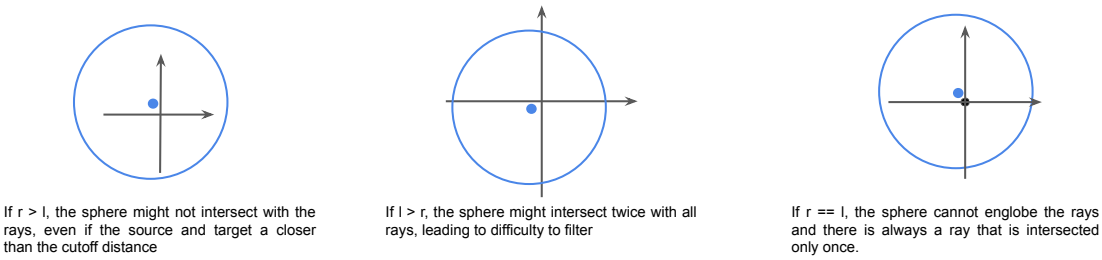


Fig. 11. 2D Spherical representation of the particles in three different cases with $r > l$, $l > r$ and $r == l$.

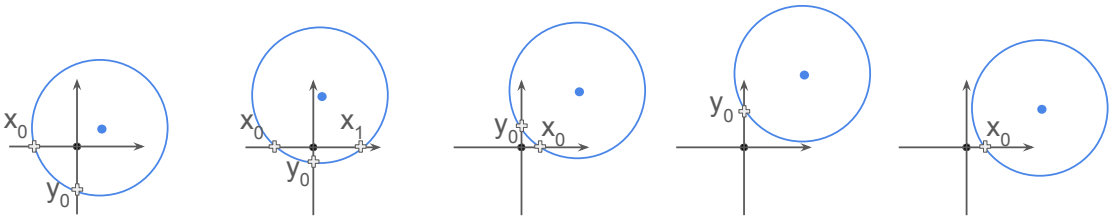


Fig. 12. 2D spherical representation illustrating the classification of intersection points between the sphere and the cross.

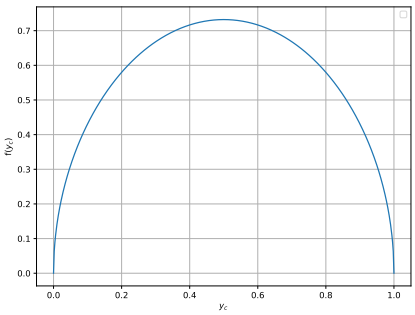


Fig. 13. Plot of the equation $\sqrt{2ry_c - y_c^2} - r + \sqrt{r^2 - y_c^2}$, for r = 1.

2) 3D Case: To convert this proof from 2D to 3D, we need to extend the concepts from the circle and cross to a sphere and a three dimensional cross.[8] Consider a sphere with radius r centered at (x_c, y_c, z_c) in 3D space, and a cross (or coordinate axes) centered at the origin with each axis extending from $-l$ to l. We are interested in analyzing the intersection of the sphere with the axes, focusing particularly on the first octant where $0 \leq x_c \leq l$, $0 \leq y_c \leq l$, and $0 \leq z_c \leq l$.

The equation of the sphere is given by:

$$(x - x_c)^2 + (y - y_c)^2 + (z - z_c)^2 = r^2. \tag{10}$$

The coordinates of the intersection points of the sphere with the axes are found by setting two of the coordinates to zero in the sphere's equation:

- Intersection with the x-axis (set $y = 0$ and $z = 0$):

$$x = x_c \pm \sqrt{r^2 - y_c^2 - z_c^2} \tag{11}$$

- Intersection with the y-axis (set $x = 0$ and $z = 0$):

$$y = y_c \pm \sqrt{r^2 - x_c^2 - z_c^2} \tag{12}$$

- Intersection with the z-axis (set $x = 0$ and $y = 0$):

$$z = z_c \pm \sqrt{r^2 - x_c^2 - y_c^2} \tag{13}$$

Let's denote the intersection points on the positive half of the axes as:

- $x_1 = x_c + \sqrt{r^2 - y_c^2 - z_c^2}$
- $y_1 = y_c + \sqrt{r^2 - x_c^2 - z_c^2}$
- $z_1 = z_c + \sqrt{r^2 - x_c^2 - y_c^2}$

We need to analyze whether these points lie within the bounds of the cross, i.e. whether $x_1 \leq l$, $y_1 \leq l$, and $z_1 \leq l$.

Assume that one of these coordinates, say x_1, exceeds l. This would mean that the intersection does not lie on the cross, i.e. $x_c + \sqrt{r^2 - y_c^2 - z_c^2} > l$.

Similarly, for y_1 and z_1, we require that:

$$y_c + \sqrt{r^2 - x_c^2 - z_c^2} > l \quad \text{or} \quad z_c + \sqrt{r^2 - x_c^2 - y_c^2} > l. \tag{14}$$

These conditions cannot all be true simultaneously for x_c, y_c, and z_c within the defined range, similar to the 2D case. Thus, a sphere will intersect the cross at most once per axis.

Next, we determine the axis closest to the sphere's center. If $x_c > y_c > z_c$, we aim to prove that the intersection on the x-axis occurs first (i.e. is the smallest).

Starting with:

$$\begin{aligned} &x_c + \sqrt{r^2 - y_c^2 - z_c^2} < y_c + \sqrt{r^2 - x_c^2 - z_c^2} \\ &\text{and} \quad x_c + \sqrt{r^2 - y_c^2 - z_c^2} < z_c + \sqrt{r^2 - x_c^2 - y_c^2}. \end{aligned} \tag{15}$$

[8] That is, three axis aligned orthogonal segments of the same size intersecting at their respective centers.

These can be simplified, following similar steps as in the 2D case:

$$\begin{aligned} &\sqrt{r^2 - y_c^2 - z_c^2} - \sqrt{r^2 - x_c^2 - z_c^2} < y_c - x_c \\ &\text{and} \quad \sqrt{r^2 - y_c^2 - z_c^2} - \sqrt{r^2 - x_c^2 - y_c^2} < z_c - x_c. \end{aligned} \tag{16}$$

The argument follows that since $x_c > y_c > z_c$, the inequalities hold true, confirming that the x-axis is the closest, and thus it is intersected first.

The 3D extension of the proof shows that a sphere intersects each axis of a coordinate cross at most once, and the axis closest to the sphere's center (in the order of $x_c > y_c > z_c$) will have the intersection point closest to the origin.

B. Conversion from Particles' Positions to Triangles

```
for(int i = 0; i < nbPoints; i++)
{
    const float3 point = points[i];
    std::array<float3, 8> corners;
    for(int idxCorner = 0 ; idxCorner < 8 ; ++idxCorner){
        corners[idxCorner].z = point.z + (idxCorner&1 ? ...
 (cutoffRadius/2)+epsilon : (-cutoffRadius/2)-epsilon );
        corners[idxCorner].y = point.y + (idxCorner&2 ? ...
 (cutoffRadius/2)+epsilon : (-cutoffRadius/2)-epsilon );
        corners[idxCorner].x = point.x + (idxCorner&4 ? ...
 cutoffRadius/2 : -cutoffRadius/2 );
    }
    vertices.push_back(corners[0]);
    vertices.push_back(corners[1]);
    vertices.push_back(corners[3]);

    vertices.push_back(corners[0]);
    vertices.push_back(corners[2]);
    vertices.push_back(corners[3]);

    vertices.push_back(corners[4]);
    vertices.push_back(corners[5]);
    vertices.push_back(corners[7]);

    vertices.push_back(corners[4]);
    vertices.push_back(corners[6]);
    vertices.push_back(corners[7]);
}
```

Code 1: Triangles generation from particles's positions.

C. Generation of Particles on a Sphere

```
const int MaxParticlesPerCell = 32;
const int MaxBoxDiv = 128;
for(int boxDiv = 2 ; boxDiv <= MaxBoxDiv ; boxDiv *= 2){
    const int nbBoxes = boxDiv*boxDiv*boxDiv;
    for(int nbParticles = nbBoxes ; nbParticles <= ...
 nbBoxes*MaxParticlesPerCell ; nbParticles *= 2){
        const double particlePerCell = ...
 double(nbParticles)/double(nbBoxes);
        const double expectedNbNeighbors = 9*particlePerCell;
        const double coef = 1. - ...
 ((2*expectedNbNeighbors)/nbParticles);
        const double validCoef = std::min(1.0, std::max(-1.0, coef));
        const double sphereRadius = acos(validCoef);

        // The following is only used if we need to build a grid ...
 of cell
        const float boxWidth = std::ceil(2.0 / sphereRadius) * ...
 sphereRadius;
        const int gridDim = boxWidth/sphereRadius;
        const float cellWidth = boxWidth/gridDim;
.....

auto generateRandomParticle() {
    double theta = 2.0 * M_PI * ((double)rand() / RAND_MAX); // ...
 Random angle between 0 and 2PI
    double phi = acos(1.0 - 2.0 * ((double)rand() / RAND_MAX)); ...
 // Random angle between 0 and PI

    // Convert spherical coordinates to Cartesian coordinates
    double x = sin(phi) * cos(theta);
    double y = sin(phi) * sin(theta);
    double z = cos(phi);

    return Particle{x, y, z};
}
```

Code 2: Non-uniform test case generation.

RSCHED: An Effective Heterogeneous Resource Management for Simultaneous Execution of Task-Based Applications

Etienne Ndamlabin, Bérenger Bramas
Inria Nancy – Grand Est, CAMUS Team,
Villers-lès-Nancy, France
ICPS Team, ICube, Illkirch, France

***Abstract*—Modern parallel architectures have heterogeneous processors and complex memory hierarchies, offering up to billion-way parallelism at multiple hierarchical levels. Their exploitation by HPC applications greatly boosts scientific discoveries and advances, but they are still not fully utilized, leading to proportionally high energy consumption. The task-based programming model has demonstrated promising potential in developing scientific applications on modern high-performance platforms. This work introduces a new framework for managing the concurrent execution of task-based applications, RSCHED. The framework aims to minimize the overall time spent executing a set of applications and maximize resource utilization. RSCHED is a two-level resources management framework: resource distribution and task scheduling, with sharable and reusable resources on the fly. A new model of Gradient Descent has been proposed, among other strategies for resource distribution, due to its well-known speedy convergence event in fast-growing systems. We implemented our proposal on StarPU and evaluated it on real applications. RSCHED demonstrated the potential to speed up the overall makespan of executed applications compared to consecutive execution with an average factor of 10x and the potential to increase resource utilization.**

***Keywords*—Heterogeneous resource management; scheduling; task-based applications; gradient descent; StarPU**

I. Introduction

High-performance computing (HPC) is crucial to making discoveries and advances in several scientific domains (astrophysics, climatology, epidemiology, biology, geology, etc.). HPC offers the ability to perform complex calculations and massive data processing at very high speed by aggregating the power of several thousand processing units, called supercomputers. Supercomputers rely on a complex, heterogeneous, and hierarchical hardware organization. The largest supercomputers are mostly composed of central processing units (CPUs) and graphical processing units (GPUs)[1], or even Field Programmable Gate Arrays (FPGAs). As parallel systems, they can process several jobs at the same time by scheduling their execution on the available resources.

In the current HPC paradigm, there are schedulers at multiple levels, that all have the same aim: distributing the workload over hardware resources. At the higher level, the batch-scheduler, like Slurm [2], OAR [3], or OpenPBS [4], manages the hardware resources of an entire supercomputer by deciding the order of execution of the jobs submitted by the users. Submitted jobs are treated by the batch scheduler as black boxes. This is advantageous because the batch scheduler can run applications implemented with any technology, giving freedom to the programmers. However, this approach might lead to resource wastage and increasing energy consumption. Such a situation can happen when an application inefficiently uses the allocated resources, when an adjustment of resources is required during different phases of execution, or when the resource manager cannot adapt the resources to the workload of newly submitted jobs. Especially since the dominant scheme for scheduling parallel jobs on parallel computers is known as variable partitioning [1], in which scheduled jobs have partitioned assigned processors they keep and use throughout their lifetime. However, executing one job after another is likely counterproductive, since HPC applications are often composed of interdependent executing kernels, and therefore cannot fully use resources due to their precedence constraints. In the current study, we aim to improve the batch scheduler by using a dynamic resource allocation strategy. Our objective is to improve the executions at the scale of the supercomputer, i.e. to reduce the overall makespan of an application set, and not to focus on a single application only. In addition, our work is tied to the task-based method, as we consider that each application is composed of tasks and that some of them can be executed on CPU, GPU or both.

In this paper, we present a resource manager for heterogeneous environments considering task-based model applications called RSCHED, aiming at optimizing their usage. In RSCHED, we propose strategies for resource distribution between concurrent task-based applications while orchestrating their execution. We have implemented our proposal within StarPU [2] and analyzed the performance of our proposal via diverse experiments. Our contributions can be summarized as follows:

- We propose a framework for managing the execution of concurrent task-based applications.
- We propose strategies for dynamically distributing resources between concurrent task-based applications.

[1]https://www.top500.org/

[2]https://www.schedmd.com/

[3]https://oar.imag.fr/

[4]https://www.openpbs.org/

- We propose a new model of Gradient Descent in a (three-dimensional) discrete space.
- We propose a strategy to automatically create and configure a scheduling context for a given task-based application.
- We present performance analysis and results showing the effectiveness of our proposals.

The remaining sections of the paper are organized as follows. In Section II, we introduce the notion of task-based application and present the state-of-the-art concerning the scheduling of task-based application under StarPU. Then, in Section III, we present our proposed resource management for simultaneous execution of task-based applications. Finally in Section IV, we evaluate the performance of our proposals.

II. Background

A. Task-Based Application

Several strategies to parallelize applications on heterogeneous computing nodes aim at maximizing resource usage. The task-based model has demonstrated high potential in various fields [3], [4], [5]. This method allows obtaining hardware-independent algorithm descriptions while developing efficient HPC applications. The level of abstraction and encapsulation relieves the users by shifting the complexity to the runtime systems, where researchers can invest the effort to create generic and efficient optimization solutions. The HPC community has at its disposal highly documented and maintained runtime systems supporting the task-based model, such as Parsec [6], and StarPU [2] a runtime system library developed at Inria Bordeaux.

Among other runtime systems supporting the task-based model, StarPU has a plus in that it has a component – *hypervisor* – allowing concurrent execution of task-based applications with minimal interference [7]. StarPU hypervisor provides confined execution environments – *scheduling contexts* – which can be used to partition computing resources. StarPU scheduling contexts can be dynamically resized and linked to a well-designed scheduler to optimize the allocation of computing resources among concurrent task-based applications/libraries. A scheduler can be chosen for each application via its linked scheduling context. Since task-based applications have various types and structures, a scheduler can be effective only for some applications, or a specific type of machine architecture [8], [9], [10], [11], [12]. StarPU also proposes basic strategies for resizing scheduling contexts and a platform for implementing additional custom ones.

Different task-based frameworks have been used to develop efficient HPC applications, such as the Lattice-Boltzmann method [13], [14], the fast-multipole method (FMM) [3], [4], [15], N-body simulations [16], linear algebra solvers [17], [18], [19], H-matrix solvers [20], the particle-in-cell method [21], the polar decomposition method [22], seismic imaging [23], [24], [25], Galerkin solver [26], to mention a few. This demonstrates that at least at a moderate scale and when used by experts, the existing task-based runtime systems can be efficient for various classes of algorithms.

1) Task-based parallelization: The task-based method divides an application into interdependent sections, called tasks. The dependencies between the tasks ensure valid parallel executions and task execution orders without race conditions. This can be likened to a graph, where the nodes represent the tasks and the edges represent the dependencies. We consider a task-based application as a Directed Acyclic Graph (DAG) $G(V, E)$ where $V = \{t_1, t_2, ..., t_n\}$ is the set of nodes and $E = \{e_{i,j} = (t_i, t_j) | 1 \leq i, j \leq n, i \neq j\}$ the set of edges representing the existing data dependencies between tasks. An edge $(t_i, t_j) \in E$ if there is a precedence constraint between t_i and $t_j \in V$, such that t_j can be executed only after the task t_i is over, and the data made available.

A task t_i is a computational element executable on one or (potentially) several types of hardware and incorporates different interchangeable kernels, each targeting a specific architecture. For instance, a matrix-matrix multiplication task in linear algebra could be either a call to cuBLAS and executed on a GPU, or a call to Intel MKL and executed on a CPU, but both kernels return equivalent results.

B. Task Scheduling and Related Work

The scheduling problem on heterogeneous computing systems has been proven NP-complete [27], whether in static or dynamic situations. Dealing with the first situation requires prediction models which are not always accurate, and a knowledge of the complete view of the task graph [28], which need expensive analysis mechanisms and incur significant overhead. The latter one is the most used [29], [30], [31], [25], [32], [22], [4] and has demonstrated its ability to deliver high performance with reduced overhead.

The two main steps of a scheduler are task selection or prioritization, and resource selection. This action can be static or dynamic according to the scheduling situation.

Due to the evolution of computing architectures, task scheduling in heterogeneous computing is an aged but hot topic. Various strategies for task scheduling have been proposed, with Heterogeneous Earliest Finish Time (HEFT) being one of the most widely used. [33] In HEFT, tasks are prioritized using a heuristic based on a prediction of the processing length of the tasks and the data transfer time between them. Whereas, resource selection is based on a heuristic that determines the resource providing the best finish time for the tasks according to the scheduling decision of previous tasks. Several variances of this approach with more advanced ranking and resource selection models have been proposed [34], [35], [36], [37]. These schedulers have in common the limitations of static schedulers previously argued, and therefore rely on greedy algorithms. In a changing environment, re-prioritizing the tasks could be necessary, which can add more overhead. A larger spectrum of task schedulers can be found in the literature [38], [39].

1) Task scheduling in StarPU: Our scheduling process follows the terminology of StarPU. In StarPU, the user first splits the problem into smaller computational tasks. Afterwards, the tasks are implemented into codelets, which are simple C functions. One task can be implemented differently in several codelets according to the targeted hardware, allowing the user to harness special accelerators, such as vectorial CPU cores

or OpenCL devices. In StarPU terminology, these devices are called workers. For each task, the user also has to describe precisely the input data, in read mode, and the output data, in write or read/write mode. StarPU considers that a scheduler has an entry point where the ready tasks are pushed, and it provides a request method where workers pop the tasks to execute, as depicted in Fig. 1.

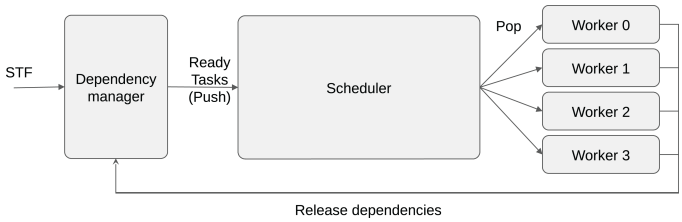


Fig. 1. Schematic view of task-based runtime system organization. A program as a sequential task flow (STF) model and converted into tasks/dependencies by the RS. New ready tasks pushed on resolved dependencies. Any idle worker calls the scheduler pop function to request a task to execute.

In StarPU, both pop/push methods are directly called by the workers that either release the dependencies or ask for a task. Consequently, assigning a task to a given worker means returning this task when the worker calls the pop method. During the execution of a StarPU program, it is possible to choose among several schedulers. The DMDA (deque model data-aware) scheduler is one of the most famous and sophisticated. It uses a HEFT-like strategy and tries to minimize the makespan by using a look-ahead strategy and data transfer costs. Another effective StarPU scheduler is Heteroprio [40], [4], a semi-automatic scheduler designed for heterogeneous machines where users must provide task priorities. A fully automatic version of the Heteroprio scheduler that computes efficient priorities is proposed [8]. Another extension of Heteroprio is the MulTreePrio [9] scheduler based on a set of balanced trees data structure, in which assignment of tasks to available resources is done according to priority scores per task for each type of processing unit. MulTreePrio makes overall good scheduling results thanks to its fast and efficient heuristics, despite the considerable variety of DAG structures from one application to another.

In all of the above, one task-based application is considered for execution/scheduling.

2) Scheduling concurrent task-based applications: In general, there is contention for the usage of resources in an HPC environment. Each user's application requests a number of processing resources (CPU/GPU for instance), an entire node or part of it. However, in both cases, it is up to the user to determine the number and type of resources, therefore this might often lead to resource wastage.

The problem presented here has a completely distinct parallelization and resources management approaches in cloud computing and Big-data frameworks, such as Spark[5] or Apache Hadoop[6]. In cloud computing, the scheduler orchestrates different executions from different types of applications, such as Big data programs, over given hardware resources. It has a view on the different operations that compose the executions, hence it can schedule and interleave them finely. However, there is a gap between the programming model and resource management.

In the basic HPC context, there is not yet a dynamic solution for concurrent job execution on the same resource as it is done in a cloud environment. In Slurm for instance, there is the notion of Job Array which consists of submitting and managing collections of similar jobs such that they may run in parallel with different input parameters or data on a node. However, it is the responsibility of the programmer to orchestrate the execution of the tasks over the resources. The same problem is encountered when several jobs are submitted separately, but now at the level of the scheduler.

The RECIPE project [41], [42] attempts to control the resources more efficiently without bridging the gap with the applications, which will end as being oriented to cloud computing instead of HPC.

Most recently, other researchers proposed a method to build scalable containerized HPC clusters in the Cloud [43], [44], by containerizing three batch schedulers, namely SLURM, OpenPBS, and OAR. They attempt to solve the problem of scaling, dynamically adding or removing containerized HPC nodes, without altering neither the Cloud orchestrator nor the HPC scheduler. This works presented promising preliminary results in that direction, scaling jobs do not impact running or pending jobs. However, it is still in the direction of Cloud Computing, where some researchers are trying to answer the question "Is the Cloud able to encompass all the categories of scientific issues in a unified way?". In addition, this does not consider task-based applications, but MPI-based applications only.

In most of the works presented above, it is either cloud oriented, or based on basic batch scheduling-like approaches using the variable partitioning scheme [1]. In both case, we can have more idle resources, which could have been used by starved applications, or could have helped in fastening the overall completion of applications and therefore in optimizing energy consumption. However, StarPU [7] offers a good platform to dynamically partition computing resources into contexts that can be used to execute several applications, and with the possibility to share resource between them. Nevertheless, there is no means to launch several applications nor dynamically distribute the workers among them.

III. Presentation of RSCHED

In this section, we present our concurrent execution approach, named RSCHED. Let us consider that we have n concurrent users' applications requesting resources for execution in a node with p workers (nb_cpus the number of CPUS and nb_gpus the number of GPUS). The main objective of RSCHED is to minimize the overall makespan for all the n applications. A secondary objective of RSCHED is maximizing resource utilization, which is a well-known energy consumption reduction approach (Fig. 2).

The RSCHED framework has a two-level scheduling model: resource distribution and task scheduling. The scheduling at each level assumes the resources are sharable between

[5]https://spark.apache.org/
[6]https://hadoop.apache.org/

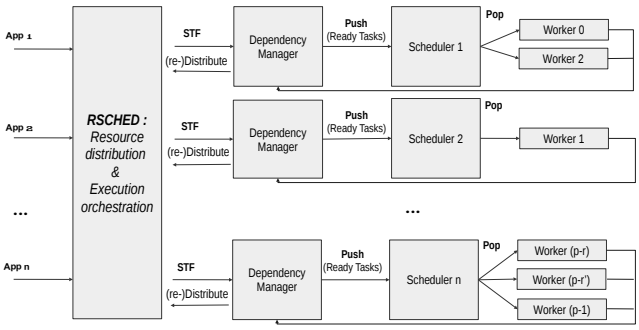


Fig. 2. Schematic view runtime system organization with n concurrent task-based applications and p workers.

the applications, and reusable on the fly (for instance when an application ends), even without a new resource subscription. The resource distribution aims at distributing the available resource among the application, and task scheduling to effectively map each task unto a given compute resource.

That said, our model is intended to be more flexible for efficient resource usage than basic batch scheduling-like approaches. Idle resources can be reused by starved applications (even before the completion of the application to which they have been assigned). For shared resources, the end of one application can help in fastening the completion of others.

Before the execution of an application, a task scheduler is associated with a separate context linked to it. The required resource distribution can be done before any task is pushed, or after all tasks have been pushed but no task has been executed. In the beginning, the lack of information on the applications makes it difficult to process an advanced distribution strategy. That is why in the first case, a naive strategy can be used to distribute the workers between the contexts, and afterward a more advanced distribution strategy. In the second case, we can have sufficient information on the applications to process an advanced distribution strategy before any task is executed.

A compromise between having all information before starting and starting sooner is to process an advanced distribution at an arbitrary time after the tasks have started to be pushed and executed. That time could be for instance when any first application completes its execution, or after all tasks have been pushed. The execution of tasks starts as soon as possible, and afterward, we re-distribute workers to the rest of the applications to balance the load and therefore minimize the overall makespan. While distributing resources, two contexts may share some workers, but with a possible performance penalty, due to context switches.

We assume all the n applications were submitted before the distribution. However, we propose early investigations of continuous application arrivals (mimicked by the proposed resizing options) that will be properly and intensively investigated in our future work.

A. RSCHED API

To distribute resources among task-based applications, the RSCHED's API requires their performances on the targeted hardware. We assume in this work that we have two types of workers, CPUs and GPUs. For each application, the required information are the following:

- CPU_W: Total (sequential) CPU workload of the graph on the targeted CPU.
- GPU_W: Total (sequential) GPU workload of the graph on the targeted GPU.
- CPU_{PW}: Total (sequential) pure CPU workload of the graph on the targeted CPU.
- GPU_{PW}: Total (sequential) pure GPU workload of the graph on the targeted GPU.

By pure CPU (or GPU) workload, we mean the workload of tasks that can only be executed by a CPU (or GPU). Providing those pieces of information implies knowing (an approximation of) the processing time of each task in the application per type of worker. There are several means of obtaining such information, like Machine Learning, or history-based performance models as it exists in StarPU.

The constraints over this information are given by the Rule 1.

Rule 1. *Given the required information as defined above, either the total workload is equal to the pure one for all the types of workers, or strictly greater than the pure one for all (see Eq. 1 and 2).*

$$CPU_W \geq CPU_{PW} \ \boldsymbol{AND} \ GPU_W \geq GPU_{PW} \tag{1}$$

$$(CPU_W = CPU_{PW} \ \boldsymbol{AND} \ GPU_W = GPU_{PW}) \ \boldsymbol{OR} \\ (CPU_W > CPU_{PW} \ \boldsymbol{AND} \ GPU_W > GPU_{PW}) \tag{2}$$

Proof: The proof of Eq. 1 is obvious. For Eq. 2, there are two cases to have equality: all the tasks are either pure CPU or pure GPU. In both cases, the other type of worker will have zero workload. ■

1) RSCHED resource distribution: Given the n applications with the required information described in the last section, a distribution strategy should produce the following information for each graph:

- L_{CPUS}: List of CPUs assigned.
- L_{GPUS}: List of GPUs assigned.
- $\#CPUS$: Number of distinct CPUs assigned ($\#CPUS = |L_{CPUS}|$).
- $\#GPUS$: Number of distinct GPUs assigned ($\#GPUS = |L_{GPUS}|$).
- $CPUS_{PR}$: The power rate of the assigned CPUs ($0 \leq CPUS_{PR} \leq \#CPUS$).
- $GPUS_{PR}$: The power rate of the assigned GPUs ($0 \leq GPUS_{PR} \leq \#GPUS$).

Rule 2. *Each application must receive at least one worker, and applications can share all the workers.*

$$(\#CPUS + \#GPUS > 0.0) \ \boldsymbol{AND} \\ (0.0 \leq \textstyle\sum_{0 \leq r \leq n} \#CPUS_r \leq n \times nb\_cpus) \ \boldsymbol{AND} \\ (0.0 \leq \textstyle\sum_{0 \leq r \leq n} \#GPUS_r \leq n \times nb\_gpus) \tag{3}$$

Rule 3. *Each application must receive at least one worker per type of pure workload.*

$$(CPU_{PW} = 0.0 \; \textbf{\textit{OR}} \; (CPU_{PW} \neq 0.0 \; AND \; \#CPUS > 0.0)) \; \textbf{\textit{AND}} \; (GPU_{PW} = 0.0 \; \textbf{\textit{OR}} \; (GPU_{PW} \neq 0.0 \; AND \; \#GPUS > 0.0)) \tag{4}$$

An estimation of the makespan of each application is used as a building block of our strategies, given a set of workers (CPUs/GPUs) assigned to the applications. We proposed an estimation called "Ideal Makespan", and more details are given in Appendix (see Algorithm 1). The following metrics are used in the evaluation of our "Ideal Makespan".

When an application has a pure workload for a given type of worker, there is a minimum length constraint over its makespan. We denote them as $tgpu_{minM}$ and $tcpu_{minM}$ (see Eq. 5 and 6).

$$tgpu_{minM} = \frac{GPU_{PW} \times coef\_par\_eff^{\#GPUS+\#CPUS-1}}{\#GPUS} \tag{5}$$

$$tcpu_{minM} = \frac{CPU_{PW} \times coef\_par\_eff^{\#GPUS+\#CPUS-1}}{\#CPUS} \tag{6}$$

In the case no CPUs (or GPUs) are assigned unto the application, $tcpu_{minM}$ (or $tgpu_{minM}$) is equal to zero.

The general formulation of our makespan estimation is given by Eq. 7.

$$\begin{aligned} ideal\_makespan \; &= MAX(tcpu_{minM}, tgpu_{minM}) \\ &+ \frac{cpu\_rem\_wl}{\#CPUS + \#GPUS \times \frac{cpu\_rem\_wl}{gpu\_rem\_wl}} \end{aligned} \tag{7}$$

Where cpu_rem_wl (resp. gpu_rem_wl) is the exceeding CPU (resp. GPU) workload compared to the gap between $tgpu_{minM}$ and $tgpu_{minM}$, and the CPU/GPU (resp. GPU/CPU) speedup.

In this work, we present four distribution strategies: Lp-Solve, MinMaxWL (Min-Max Workload balancing), DSR-CLUS (Dedicated plus Shared Resource with Clustering) and DSR-GD (Dedicated plus Shared Resource with Gradient Descent).

a) LpSolve: In this strategy, we rely on the linear programming model presented in a previous study [4]. Originally, this model was employed to compute an ideal makespan (a theoretical lower bound) for tasks executed on heterogeneous architectures. The model is given by:

$$\begin{cases} Objective\,function : min(T) \\ \sum\limits_{\omega \; in \; \Omega} \alpha_1^{\omega} \, t_1^{\omega} = t_1 \leq T \\ \sum\limits_{\omega \; in \; \Omega} \alpha_2^{\omega} \, t_2^{\omega} = t_2 \leq T \\ \quad ... \\ \sum\limits_{\omega \; in \; \Omega} \alpha_P^{\omega} \, t_P^{\omega} = t_P \leq T \end{cases} \tag{8}$$

$$\begin{cases} \sum\limits_{p=1}^{P} \alpha_p^1 = 1 \\ \sum\limits_{p=1}^{P} \alpha_p^2 = 1 \\ \quad ... \\ \sum\limits_{p=1}^{P} \alpha_p^{|\Omega|} = 1 \end{cases} \tag{9}$$

Here, P denotes the number of processing units and $|\Omega|$ is the total number of tasks. The coefficient α_p^{ω} indicates the proportion of task ω processed by unit p, and t_p^{ω} represents the time taken to complete task ω on unit p, given that this duration varies based on the type of the processing unit. Accordingly, the first system determines the computation duration for each unit, with T being the longest duration. The second part ensures that each task is computed at 100%.

While this model provides an upper bound for ideal performance, it doesn't account for the dependencies between the task order and consider that tasks can be divided among various processing units.

In our adaptation, we assume that a single application consists of three tasks: one for exclusive CPU work, another for exclusive GPU work, and the last one for work that can be executed on either CPU or GPU. Given this characterization, the aforementioned LP model remains applicable.

However, our focus isn't on the ideal makespan T, but on the α coefficients as we aim to find the most efficient way to distribute the application across processing units. Although the LP provides an optimal distribution for an ideal system, it also guides us on the proportion of each application that should be allotted to each processing unit type. However, this distribution might be inefficient in real-world scenarios, leading to a task being fragmented across all units or a skewed allocation, like 99% on one unit and 1% on another, which may not always be practical. Consequently, it's essential to transform these coefficients into practical distribution values to derive a feasible scheduling strategy.

To achieve this, we use a two-step method. First, we sum the distribution coefficients per unit type to determine the fraction of each application designated for every processing unit type. For instance, if the LP solution suggests distributing an application as 0.1 and 0.4 across two CPUs, and 0.2, 0.2, and 0.1 across three GPUs, we infer it should be equally split (0.5 for CPU and 0.5 for GPU). Subsequently, we decide on the application distribution based on these values. In the next step, we compute the processing time for each unit type by multiplying the number of a given unit type with T. For instance, with two CPUs and a makespan of 10s, we have 20s of total CPU time to distribute. Each application is then assumed to use a fraction of this time proportional to its distribution coefficient. Our greedy algorithm identifies the application with the highest use proportion and allocates it to the processing unit with the least utilization, continuing until every application has been entirely mapped to processing units.

b) DSR (Dedicated plus Shared Resource) strategies: For illustrations, let us consider for instance that we have three applications, four CPUs, and two GPUs. The DSR strategies

proceed in two steps to distribute the workers among the applications:

The first step consists in assigning dedicated (unshared) workers (GPUs or CPUs) to each application, proportionally to their GPU/CPU workload compared to the sum of all the applications' workloads. Supposed the proportions of CPU workloads (cpu_pwl) are 0.31, 0.56, and 0.13, the numbers of dedicated CPUs (given by $\lfloor nb\_cpus \times cpu\_pwl \rfloor$) will be respectively $\lfloor 4 \times 0.31 \rfloor = \lfloor 1.24 \rfloor = 1$, $\lfloor 4 \times 0.56 \rfloor = \lfloor 2.24 \rfloor = 2$, and $\lfloor 4 \times 0.13 \rfloor = \lfloor 0.52 \rfloor = 0$. In that case, the number of remaining CPUs is 1. The same is similarly done for GPUs. In the case of hybrid workloads, and if the standard deviation between the GPU/CPU speedups is above a certain threshold, we proceed to the barter which consists of exchanging GPU against CPUs to accelerate the most GPU-optimized applications.

The second step involves sharing the remaining workers to the applications, using a given technique. Here we proposed two DSR strategies, based on two different techniques for workers sharing, the clustering and the gradient descent: DRS-CLUS (Dedicated plus Shared Resource with Clustering) and DSR-GD (Dedicated plus Shared Resource with Gradient Descent). The sharing process is done based on the remaining cpu_pwl (here in the case of CPUs, we have: 0.24, 0.24, and 0.52).

As for the DRS-CLUS strategy, resource sharing is done as follows. The number of clusters is equal to the number of remaining workers. The workers/applications mapping is done to balance the load over the workers as much as possible.

In the DSR-GD strategy, the second step is done as follows (see Algorithm 2 in Appendix, from lines 18 to 24). The remaining workers are shared between applications using a new model of gradient descent (GD) in a (three-dimensional) discrete space.

The Gradient Descent is an efficient strategy well-known for its speedy convergence in convex and smooth optimization problems (if well-tuned), even in low memory, and computational loading environments [45], [46].

Our GD strategy is modelled as follows:

Given the CPUs and GPUs IDs, {0,1,2,3} and {0,1} respectively, our research space is modelled as a three-dimensional discrete space (X, Y, Z):

- X: The numbers of assigned CPUs ($\#CPUS$) per graph. Each application can have from zero to the number of CPUs (Ex. {1,2,1}, the first app has 1 CPU, the second one 2 CPUs, and the third one has 1 CPU).
- Y: The numbers of assigned GPUs ($\#GPUS$) per graph. Each application can have from zero to the number of GPUs (Ex. {1,1,1}, the first app has 1 GPU, the second one 1 GPU, and the third one has 1 GPU).
- Z: The possible (graph-to-gpu/cpu) mappings given X and Y. An example of mapping related to the Y and Y ones above is $\{\{0\},\{0\}\}$, $\{\{1,2\},\{0\}\}$, $\{\{3\},\{1\}\}$. In this example, the first app has the CPU id=0 and the GPU id=0, app 2 has the CPU ids=1,2 and the GPU id=0 and app 3 has the CPU id=3 and the GPU id=1. We can see that app 1 and app 2 share one GPU together.

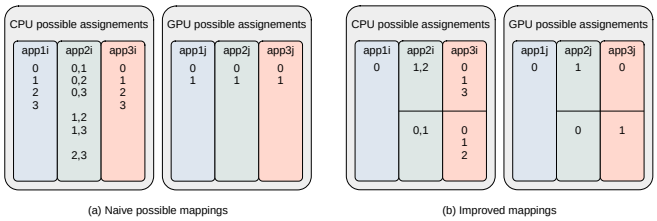


Fig. 3. Illustration of determination of possible mappings with three applications, four CPUs, and two GPUs, and given that X={1,2,1} and Y={1,1,1}.
The improved version has fewer duplicates in terms of obtained mappings 24 vs 768).

Iteration in the axis is done as follows:

- The indexes in X and Y can be seen as permutations with repetition of a possible number of assigned workers in nb_graphs positions:
  - $X_i \in$ {{0,0,0}, {0,0,1}, ... , {0,0,4}, {0,1,0}, ... ,{4,4,4}}. By looking carefully, we can see that each X_i is likened to "i" in base $(nb\_cpus + 1)$
  - $Y_j \in$ {{0,0,0}, {0,0,1}, {0,0,2}, {0,1,0}, ... ,{2,2,2}}. In like manner, Yi is likened to "j" in base $(nb\_gpus + 1)$
- Z_k (or Zijk): The k-th possible mapping, given X_i and Y_j. One naive way to get the Z_k's values is to generate the different 2-uplets, of the possible CPUs assignment given X_i and the possible GPUs assignment given Y_j. For our example, Fig. 3(b) illustrates the determination of possible mappings and how to find the Z_k. We have $Z = (app1i \times app2i \times app3i) \times (app1j \times app2j \times app3j)$. While the naive version leads to $((4 \times 6 \times 4) \times (2 \times 2 \times 2) = 96 \times 8 =)$ 768 possible mappings, the improved one gives $((1 \times 2 \times 3) \times (1 \times 2 \times 2) = 6 \times 4 =)$ 24 possible mappings. The improved version has far fewer duplicates in obtained possible mappings, and this is for only three applications.

This model presents the whole research space. However, during our research process, the search space is circumscribed around the convex zone containing our global optimum. This is done by setting for each application the minimum number of workers (CPUs and GPUs) obtained in step one, and the maximum by adding the number of workers not yet assigned. Thus, we reduce the search space and speed up the search.

Our gradient function is evaluated as a symmetric linear interpolation [47]. Our search process follows the pattern direction [48], first, we search towards the X direction, then the Y direction, and finally the Z direction. To ensure process time scaling, the learning rates for the different axes are $tau_x = 0.001 \times (\lfloor nb\_graphs/\sqrt{nb\_cpus} \rfloor + 1)$, $tau_y = 0.001 \times (\lfloor nb\_graphs/\sqrt{nb\_gpus} \rfloor + 1)$ and $tau_z = 0.0005 \times (\lfloor nb\_graphs/\sqrt{nb\_cpus + nb\_gpus} \rfloor + 1)$.

c) MinMaxWL (Min-Max Workload balancing): The MinMaxWL algorithm is a load-balancing strategy that dis-

tributes workers among applications by minimizing the maximum ideal makespan. The strategy is depicted in Algorithm 3 (in Appendix), and has four main steps.

First of all, it assigns one worker to each application having a pure workload according to the type of worker (from lines 3 to 13). While trying to assign a worker to the current application, if there are no remaining workers of the type, the application shares one with the under-loaded application related to that type. A backpropagation is employed in the case of sharing to ensure load-balancing when less-loaded applications are treated after more-loaded ones.

The second step is to ensure all the applications have at least one worker, by assigning a worker to applications without a pure workload (from lines 15 to 28). Since those applications are hybrid, the type of worker to assign is the fastest on the application. A similar sharing process is also employed, but this time the sharing is made on the worker, leading to the smallest makespan at the point.

Finally, while there are remaining workers (per type of worker), it assigns a worker to the application that will minimize the maximum ideal makespan among all the applications.

2) Distribution options: If the distribution of resources to the applications is accurate, the applications will end almost at the same time, and so will the workers. Otherwise, some resources could be idle for a long, while remaining applications may need them. To deal with that situation, a redistribution of the resources might be necessary. One crucial aspect to consider for it is the condition of resizing, the when.

The condition of resizing we employed is the following:

- An application just ended, and there remain applications to run.
- There is a significant standard deviation between the progress rate of the applications.

The estimation process time (workload) for a CPU or GPU may differ from the effective processing time during the execution. For instance, let us suppose an application with a CPU workload of 100s, and that has been assigned 10 CPU workers. Suppose 5s after executions start, there is a need for redistribution, and it remains at an overall 20s processing time for the workload. We would have expected having executed $5 \times 10 = 50$s for the application, whereas we have $100 - 20 = 80$s. The progress rate in this case is therefore equal to $80/50 = 1.6$; which means the application is running faster than expected. Now we know that possibly the application may end in $(20/1.6)/10 = 1.2$s instead of $20/10 = 2$s.

Before the redistribution, we adjust the workload of each application according to its progress rate. We have proposed two resource redistribution options.

a) One Distribution: This is the default behavior, where the distribution is done once and for all.

b) Multiple distributions: Here we do the distribution as initially, but considering the adjusted workloads of the remaining applications.

c) Inherit released workers: Here we distribute the released workers (by the just-ended application) to the remaining ones, with high privilege to those that were delaying.

B. RSCHED Implementation in StarPU

StarPU offers a platform to dynamically construct, delete, and modify Scheduling Contexts, which are used to execute several parallel kernels in an isolated way and without interference. This allows the users to assign workers to the contexts, at their creation time, or resize them during program execution. However, this is subject to the knowledge of the number of workers needed for each scheduling context. StarPU proposes online performance tools to monitor the execution of tasks, to make execution time estimations.

1) Multiple task-based applications: There are several applications implemented in StarPU. However, there is no mechanism to launch or orchestrate the execution of concurrent applications. For the sake of simplicity, instead of using several different applications, we have exploited the implementation of Cholesky factorization to have several independent applications. The Cholesky application in StarPU is implemented with a performance model for each codelet. We have added a parameter to specify the number of applications to create, and for each application, we gave the possibility to specify the size and the number of blocks via environment variables.

2) Context creation and workload determination: For each application, a separate context is created and a task scheduler is associated with it. In this work, we have chosen to use DMDA as a scheduler for all the applications. DMDA relies on a historical performance model to be able to estimate in advance the duration of a codelet on each kind of processing unit. Using StarPU historical performance model, we have been able to compute the different workloads of the concurrent applications.

IV. Performance Study

A. Experiments Setup

1) Hardware: We conducted our experiments on three configurations with different GPU models as follows:

- A100: Composed of two 32-core AMD Zen3 EPYC 7513 @ 2.60 GHz, and 2 NVIDIA A100 (40GB). We use 30 CPU cores and 16 CUDA streams per GPU.
- Quadro: Composed of 2 Icosa-core Cascade Lake Intel Xeon Gold 5218R CPU @ 2.10 GHz, and 2 NVIDIA Quadro RTX8000 (48GB). We use 30 CPU cores and 16 CUDA streams per GPU.
- K40M: Composed of 2 Dodeca-cores Haswell Intel Xeon E5-2680 v3 2.5 GHz, and 4 K40m GPUs (12GB). We use 20 CPU cores and 8 CUDA streams per GPU.

We have configured StarPU as follows. For each configuration, we set the environment variables STARPU_NCPU to the number of CPU cores, STARPU_NCUDA to the number of GPU, and STARPU_NWORKER_PER_CUDA to the number of CUDA streams. Therefore, for all the configurations, we have more GPU workers than CPU ones.

2) Task-based applications: We implemented Cholesky factorization in StarPU and tested it across twelve different configurations as follows:

- app_0: Matrix size of 3.200, with 5 blocks
- app_1: Matrix size of 3.200, with 10 blocks
- app_2: Matrix size of 6.400, with 10 blocks
- app_3: Matrix size of 6.400, with 20 blocks
- app_4: Matrix size of 9.600, with 10 blocks
- app_5: Matrix size of 9.600, with 30 blocks
- app_6: Matrix size of 19.200, with 20 blocks
- app_7: Matrix size of 19.200, with 30 blocks
- app_8: Matrix size of 25.600, with 40 blocks
- app_9: Matrix size of 25.600, with 80 blocks
- app_{10}: Matrix size of 76.800, with 80 blocks
- app_{11}: Matrix size of 76.800, with 120 blocks

3) Software configuration: For each application, we have made different affinities related to the types of compute units (CPU or GPU). By default, all the tasks of the Cholesky application have two codelets, one for CPU and one for GPU. Overall, we have used the three following affinities:

- Default (*affinity0*): each task has one CPU codelet and GPU codelet.
- Only CPU (*affinity1*): all the tasks have only a CPU codelet.
- Only GPU (*affinity2*): all the tasks have only a GPU codelet.

To analyze the influence of the number of concurrent applications, and of the workload, we have made experiments with 3, 6, and 12 concurrent applications. To be in accord with a realistic scenario, we have shuffled the list of applications in each experiment. Then we took consecutive applications to form the groups. For instance, in the case of three applications, we have executed concurrently the applications at the first, second third positions, then the fourth, fifth, and sixth positions, and so on.

B. Metrics

In our experiments, we have compared our four distribution strategies against the concurrent execution using a unique context with all the workers (DMDA_CONC), and against sequential execution (i.e. one application after the another) using a unique context with all the workers (DMDA_SEQ).

As metrics, we have considered the speedup, the data transfer, and the resource utilization efficiency (RUE). We also compared the distribution processing time of our strategies.

The RUE is a new metric hereby introduced and defined as follows:

Definition 1. *We define the RUE as the ability to maximize the utilization of the resource, that is, using the adequate number and types of resources for the execution of each application. The RUE is given by Eq. 10, which is the product of the resource utilization and the efficiency [49].*

$$RUE = \frac{\sum_{p \in USED\_WK}\{processing\_time\ of\ worker\ p\}}{\sum_{p \in USED\_WK}\{total\_active\_time\ of\ worker\ p\}} \times \frac{speedup}{|USED\_WK|} \tag{10}$$

$USED\_WK$ is the list of distinct workers (CPU or GPU) used for the execution of the applications, whether concurrently or sequentially (i.e. in DMDA_SEQ). We normalized the RUE such that the values lie between 0 and 1.

C. Experiments Results and Analysis

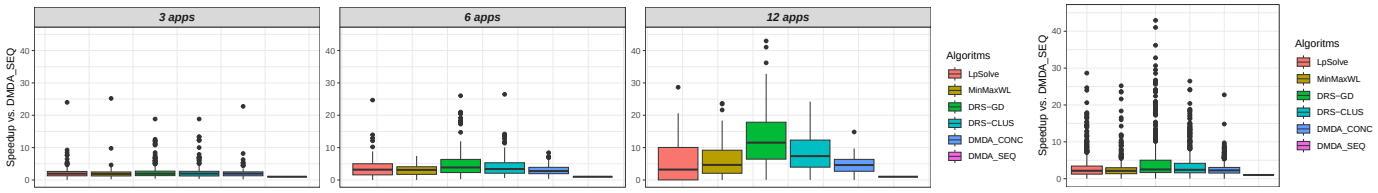
1) Default experiments: We first present the performance of our strategies (LpSolve, MinMaxWL, DSR-GD, and DSR-CLUS), and of DMDA_CONC, against DMDA_SEQ, in terms of Speedup, then in terms of data transfer, and finally in terms of RUE.

a) Speedup: The big picture of the speedup realized by the different strategies compared to DMDA_SEQ is presented in Fig. 4. For all the configurations and affinities, the LpSolve, MinMaxWL, DSR-GD, DSR-CLUS, and DMDA_CONC can significantly accelerate the execution DMDA_SEQ (Fig. 4b). Moreover, our strategies (except MinMaxWL) perform better even than DMDA_CONC with the increase in the number of applications. We observe in this study an outperformance over DMDA_CONC in more than 50% of cases for LpSolve, and more than 75% of cases for DSR-GD and DSR-CLUS (Fig. 4b).

DSR-GD and DSR-CLUS reach an acceleration of 40× compared to DMDA_SEQ. However, DSR-GD outperforms DSR-CLUS in more than 50% of situations, observed while we have an increase in applications. This means that conceptually, the Gradient Descent performs better than the clustering since the two strategies have the same building block. We observe that the speedup of the strategies increases with the number of concurrent applications (Fig. 4a). The study of the variation of the speedup according to the workload and of the GPU/CPU acceleration (see Fig. 5 and Fig. 6) reveals that DSR-GD and DSR-CLUS perform better when the percentage of the standard deviation of GPU/CPU acceleration among the application increases. This is explained by the employed bartering technique that gives GPU in preference to more accelerated applications in exchange for CPU to others.

The study of the variation of the speedup according to the number of applications over the different hardware configurations (Fig. 7) reveals that the strategies perform better on recent architectures (Quadro and A100) which have more accelerated GPU than on older ones (K40M). More specifically, DSR-GD and DSR-CLUS perform better than the other strategies, due to the same reasons as previously.

Globally, the DSR-GD realizes better speedup and in more of the situations than the others, then DSR-CLUS followed by LpSolve.



(a) Speedup obtained for the 3, 6, and 12 concurrent applications.

(b) Speedup summary

Fig. 4. Speedup of LpSolve, MinMaxWL, DSR-GD, DSR-CLUS, DMDA_CONC against DMDA_SEQ for all the affinities (affinity0, affinity1, affinity2) and all the hardware configurations (K40M, Quadro, and A100).

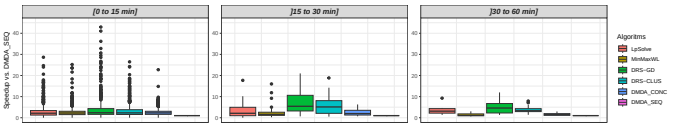


Fig. 5. Speedup of LpSolve, MinMaxWL, DSR-GD, DSR-CLUS, DMDA_CONC against DMDA_SEQ according to the average workload in minute, on K40M, Quadro or A100.

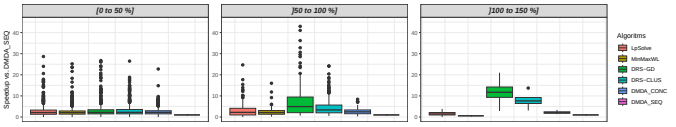


Fig. 6. Speedup of LpSolve, MinMaxWL, DSR-GD, DSR-CLUS, DMDA_CONC against DMDA_SEQ according to the standard deviation percentage of GPU/CPU speedup between the applications, on K40M, Quadro or A100.

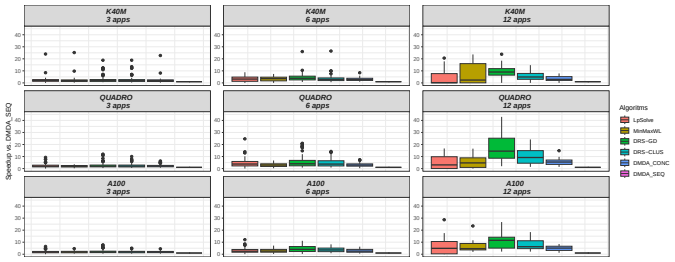


Fig. 7. Speedup of LpSolve, MinMaxWL, DSR-GD, DSR-CLUS, DMDA_CONC against DMDA_SEQ for 3, 6, and 12 concurrent applications with all the affinities, on K40M, Quadro or A100.

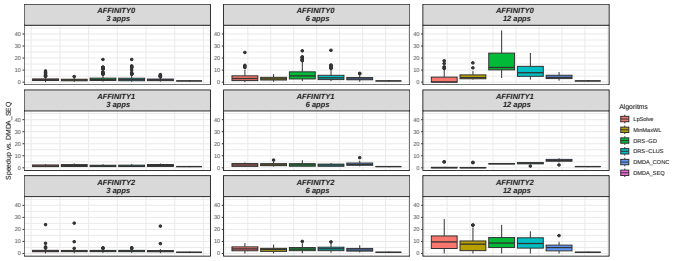


Fig. 8. Speedup of LpSolve, MinMaxWL, DSR-GD, DSR-CLUS, DMDA_CONC against DMDA_SEQ for 3, 6, and 12 concurrent applications with each affinity, on K40M, Quadro or A100.

b) Data transfer: The total amount of memory transfer obtained with the different strategies are provided in Fig. 9. All the strategies for concurrent execution used in this study (LpSolve, MinMaxWL, DSR-GD, DSR-CLUS, and DMDA_CONC) significantly reduce the total memory transfer compared to the sequential execution (Fig. 9b), DSR-CLUS been the best one.

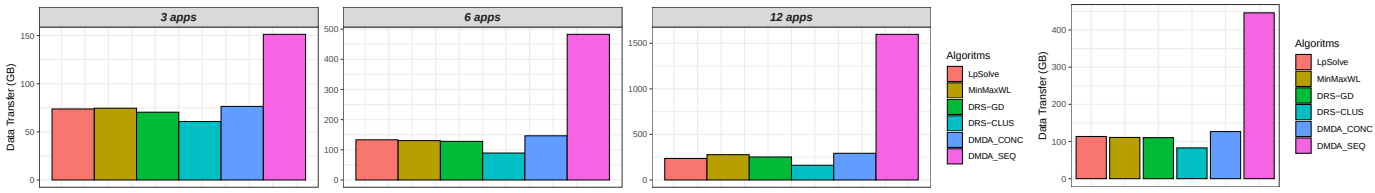
c) Resource utilization efficiency: The Normalized RUE obtained with the different strategies are provided in Fig. 10. We observe in this study that the concurrent execution of applications leads to more effective resource usage than the sequential one. The DSR-GD and DSR-CLUS strategies are more efficient in terms of resource utilization than the other (Fig. 10b), DSR-GD been the best one as the number of applications increases. Executing the applications sequentially one after the other leads to more resource wastage, which is known as a major cause of energy consumption in data centers [10], [11]. Moreover, we observe that the improvement of our strategies (DSR-GD, DSR-CLUS, and LpSolve) in terms of RUE correspond with the situations in which they are speeding up compared to the sequential execution (Fig. 10a $\equiv$ Fig. 4a). Therefore, succeeding in speeding up the sequential execution of tasks-based applications using our strategies might also help to reduce energy consumption, thanks to the effectiveness of the scheduler used.

2) Distribution processing time and options:

a) Distribution processing time: In a dynamic situation where applications arrive continuously (as we will study in the future), the decision processing has to be fast. Fig. 11 presents the evolution of processing time for each of our proposed strategies according to the acceleration of GPU compared to the CPU, and the number of executed applications.

The MinMaxWL and DSR-CLUS strategies are faster than LpSolve and DSR-GD which are meta-heuristics. However, the processing times of all the strategies are relatively small to expect good behavior even in a dynamic and computational loading environment. Moreover, even though we add the decision process time to the overall makespan, we will still have almost the same results as presented above.

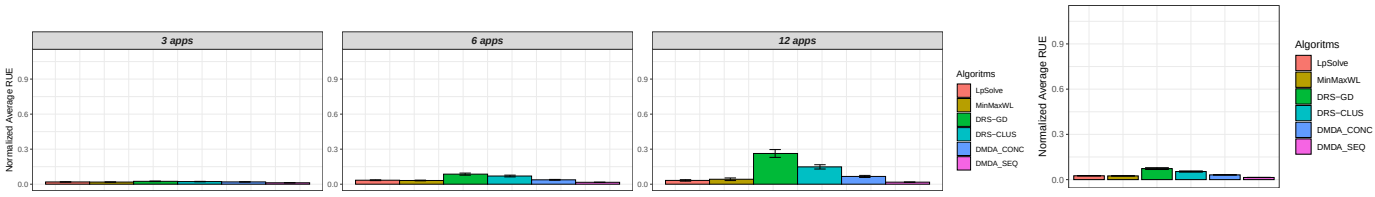
Furthermore, we realize that DSR-GD scale better than LpSolve given their processing time (Fig. 11) and speedup compared to DMDA_SEQ (Fig. 4, 5, 6, 7, 8). This performance of DSR-GD is due in part to the choice of learning rate that helped speedily converge towards the optimal solution



(a) Memory transfer obtained for the 3, 6, and 12 concurrent applications.

(b) Memory transfer summary

Fig. 9. Memory transfer of LpSolve, MinMaxWL, DSR-GD, DSR-CLUS, DMDA_CONC against DMDA_SEQ for all the affinities and all the hardware configurations (K40M, Quadro, and A100).



(a) Average RUE obtained for the 3, 6, and 12 concurrent applications.

(b) Average RUE summary

Fig. 10. Average RUE of LpSolve, MinMaxWL, DSR-GD, DSR-CLUS, DMDA_CONC against DMDA_SEQ for all the affinities and all the arch configurations (K40M, Quadro, and A100).

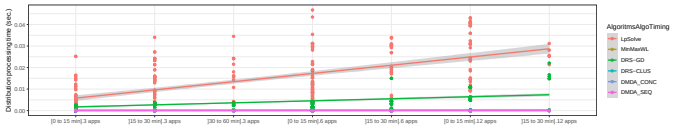


Fig. 11. Distribution processing time of LpSolve, MinMaxWL, DSR-GD, and DSR-CLUS according to the standard deviation percentage of GPU/CPU speedup between the applications, and the number of concurrent applications

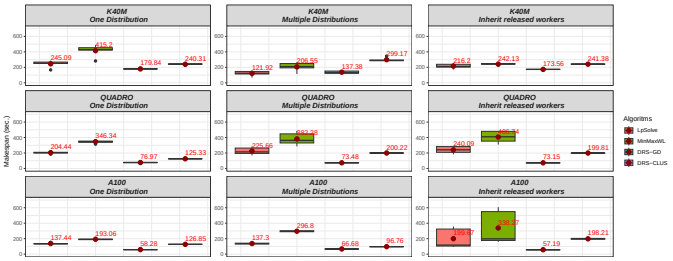


Fig. 12. **Analysis of distribution options** for LpSolve, MinMaxWL, DSR-GD, DSR-CLUS with 12 concurrent applications. *The averages makespan are displayed in red.*

no matter the number of applications and resources, and also due to the bartering technique employed (see Section III-A1). This is a promising property for a scalable system.

b) Distribution options: We carried out a study on the effectiveness of the re-distribution options ("Multiple Distributions", and "Inherit released workers") presented in Section III-A2 comparatively to the default one ("One Distribution") when combined with each of our four strategies (LpSolve, MinMaxWL, DSR-GD, DSR-CLUS). Fig. 12 presents the makespan obtained in each case, which reveals that the effectiveness of the re-distribution options ("Multiple Distributions", and "Inherit released workers") depends on the strategies and the configuration.

The combination DSR-GD/"Inherit released workers" always produces a gain for all the configurations. For the other cases, the combination strategy/option leads to a gain only for some configurations. We notify significant improvement in some cases, proving that there is hope for improving the results obtained above by using resource redistribution. However, it is imperative to do more investigations on configuration and strategy sensitivity to achieve this.

V. Conclusions

As computing resources are getting more complex and powerful, there is little doubt that we need methods to reduce the waste from the users' choices, bad application optimization, or heterogeneous workloads during executions. This is where the task-based model grants more opportunities by exposing a dynamic degree of parallelism with execution environments able to use this information in the most constructive and thus efficient way. To minimize the overall makespan and maximize resource utilization while executing multiple task-based applications, we introduce RSCHED, a two-level resource management framework that allows 1) dynamic resource distribution for concurrent execution of task-based applications, and 2) dedicated task scheduling for each application. We proposed strategies for resource distribution and implemented our proposal on the StarPU runtime system, proposing schedulers on which we rely for the second level. A new model of Gradient Descent has been proposed, among other strategies for resource distribution. We evaluated our

proposal using real applications based on the StarPU implementation of Cholesky factorization. RSCHED demonstrated the potential to speed up the overall makespan compared to consecutive execution with an average factor of 10x, and a factor of 5x when compared against the concurrent execution without resource distribution using DMDA. RSCHED also demonstrated the potential to increase the rate of resource utilization as the number of applications increases. Moreover, the decision time of our strategies and our initial study on resizing options indicate promising scalability.

In our future work, we will investigate continuous application arrivals and resizing options to have more adaptive solutions due to the variability and complexity of both the applications and the computing resources. That being said, we would like to consider different applications (instead of just Cholesky), explore multiple nodes, and improve RSCHED decisions by analyzing the structures of task graphs.

ACKNOWLEDGMENTS

This work is supported by the TExas, an Inria exploratory project. Experiments presented in this paper were carried out using the PLAFRIM experimental testbed, being developed under the Inria PlaFRIM development action with support from Bordeaux INP, LABRI and IMB and other entities: Conseil Régional d'Aquitaine, Université de Bordeaux and CNRS (and ANR in accordance to the programme d'investissements d'Avenir (see http://www.plafrim.fr/).

REFERENCES

[1] N. P. Drakenberg, "Multi-objective processor-set selection for computational cluster-systems," in *Job Scheduling Strategies for Parallel Processing: 16th International Workshop, JSSPP 2012, Shanghai, China, May 25, 2012. Revised Selected Papers 16*. Springer, 2013, pp. 56–75.

[2] C. Augonnet, S. Thibault, R. Namyst, and P.-A. Wacrenier, "Starpu: a unified platform for task scheduling on heterogeneous multicore architectures," in *Euro-Par 2009 Parallel Processing: 15th International Euro-Par Conference, Delft, The Netherlands, August 25-28, 2009. Proceedings 15*. Springer, 2009, pp. 863–874.

[3] E. Agullo, B. Bramas, O. Coulaud, E. Darve, M. Messner, and T. Takahashi, "Task-based fmm for multicore architectures," *SIAM Journal on Scientific Computing*, vol. 36, no. 1, pp. C66–C93, 2014.

[4] ——, "Task-based fmm for heterogeneous architectures," *Concurrency and Computation: Practice and Experience*, vol. 28, no. 9, pp. 2608–2629, 2016.

[5] J. M. Couteyen Carpaye, J. Roman, and P. Brenner, "Design and analysis of a task-based parallelization over a runtime system of an explicit finite-volume cfd code with adaptive time stepping," *Journal of Computational Science*, vol. 28, pp. 439–454, 2018.

[6] G. Bosilca, A. Bouteiller, A. Danalis, M. Faverge, T. Hérault, and J. J. Dongarra, "Parsec: Exploiting heterogeneity to enhance scalability," *Computing in Science & Engineering*, vol. 15, no. 6, pp. 36–45, 2013.

[7] A. Hugo, A. Guermouche, P.-A. Wacrenier, and R. Namyst, "Composing multiple starpu applications over heterogeneous machines: a supervised approach," *The International journal of high performance computing applications*, vol. 28, no. 3, pp. 285–300, 2014.

[8] C. Flint, L. Paillat, and B. Bramas, "Automated prioritizing heuristics for parallel task graph scheduling in heterogeneous computing," *PeerJ Computer Science*, vol. 8, p. e969, 2022.

[9] H. Tayeb, B. Bramas, A. Guermouche, and M. Faverge, "Multreeprio: Scheduling task-based applications for heterogeneous computing systems," in *COMPAS 2022-Conférence francophone d'informatique en Parallélisme, Architecture et Système*, 2022.

[10] J. E. Ndamlabin Mboula, V. C. Kamla, M. H. Hilman, and C. Tayou Djamegni, "Energy-efficient workflow scheduling based on workflow structures under deadline and budget constraints in the cloud," *arXiv preprint arXiv:2201.05429*, 2022.

[11] J. E. Ndamlabin Mboula, V. C. Kamla, and C. Tayou Djamegni, "Dynamic provisioning with structure inspired selection and limitation of vms based cost-time efficient workflow scheduling in the cloud," *Cluster Computing*, pp. 1–25, 2021.

[12] ——, "Cost-time trade-off efficient workflow scheduling in cloud," *Simulation Modelling Practice and Theory*, p. 102107, 2020.

[13] E. Calore and S. F. Schifano, "Porting a lattice boltzmann simulation to fpgas using ompss," in *Parallel Computing: Technology Trends*. IOS Press, 2020, pp. 701–710.

[14] J. V. Lima, G. Freytag, V. G. Pinto, C. Schepke, and P. O. Navaux, "A dynamic task-based d3q19 lattice-boltzmann method for heterogeneous architectures," in *2019 27th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*. IEEE, 2019, pp. 108–115.

[15] M. AbdulJabbar, R. Yokota, and D. Keyes, "Asynchronous execution of the fast multipole method using charm++," *arXiv preprint arXiv:1405.7487*, 2014.

[16] M. Pericàs, X. Martorell, and Y. Etsion, "Implementation of a hierarchical n-body simulator using the ompss programming model," in *Proceedings of the 1st Workshop on Irregular Applications: Architectures and Algorithms*, 2011, pp. 23–30.

[17] E. Agullo, L. Giraud, and S. Nakov, "Task-based sparse hybrid linear solver for distributed memory heterogeneous architectures," in *Euro-Par 2016: Parallel Processing Workshops: Euro-Par 2016 International Workshops, Grenoble, France, August 24-26, 2016, Revised Selected Papers 22*. Springer, 2017, pp. 83–95.

[18] X. Lacoste, M. Faverge, G. Bosilca, P. Ramet, and S. Thibault, "Taking advantage of hybrid systems for sparse direct solvers via task-based runtimes," in *2014 IEEE International Parallel & Distributed Processing Symposium Workshops*. IEEE, 2014, pp. 29–38.

[19] E. Agullo, A. Buttari, A. Guermouche, and F. Lopez, "Multifrontal qr factorization for multicore architectures over runtime systems," in *Euro-Par 2013 Parallel Processing: 19th International Conference, Aachen, Germany, August 26-30, 2013. Proceedings 19*. Springer, 2013, pp. 521–532.

[20] R. Carratalá-Sáez, M. Faverge, G. Pichon, G. Sylvand, and E. S. Quintana-Ortí, "Tiled algorithms for efficient task-parallel h-matrix solvers," in *2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE, 2020, pp. 757–766.

[21] R. Arias Mallo, "Particle-in-cell plasma simulation with ompss-2," Master's thesis, Universitat Politècnica de Catalunya, 2019.

[22] D. Sukkari, H. Ltaief, M. Faverge, and D. Keyes, "Asynchronous task-based polar decomposition on single node manycore architectures," *IEEE Transactions on parallel and distributed systems*, vol. 29, no. 2, pp. 312–323, 2017.

[23] L. Boillot, G. Bosilca, E. Agullo, and H. Calandra, "Task-based programming for seismic imaging: Preliminary results," in *2014 IEEE Intl Conf on High Performance Computing and Communications, 2014 IEEE 6th Intl Symp on Cyberspace Safety and Security, 2014 IEEE 11th Intl Conf on Embedded Software and Syst (HPCC, CSS, ICESS)*. IEEE, 2014, pp. 1259–1266.

[24] V. Martínez, D. Michéa, F. Dupros, O. Aumage, S. Thibault, H. Aochi, and P. O. Navaux, "Towards seismic wave modeling on heterogeneous many-core architectures using task-based runtime system," in *2015 27th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*. IEEE, 2015, pp. 1–8.

[25] S. Moustafa, W. Kirschenmann, F. Dupros, and H. Aochi, "Task-based programming on emerging parallel architectures for finite-differences seismic numerical kernel," in *Euro-Par 2018: Parallel Processing: 24th International Conference on Parallel and Distributed Computing, Turin, Italy, August 27-31, 2018, Proceedings 24*. Springer, 2018, pp. 764–777.

[26] B. Bramas, P. Helluy, L. Mendoza, and B. Weber, "Optimization of a discontinuous galerkin solver with opencl and starpu," *International Journal on Finite Volumes*, vol. 15, no. 1, pp. 1–19, 2020.

[27] P. Brucker and S. Knust, "Complexity results for scheduling problems," 2009.

[28] P. Baptiste, C. Le Pape, and W. Nuijten, *Constraint-based scheduling: applying constraint programming to scheduling problems*. Springer Science & Business Media, 2001, vol. 39.

[29] M. L. O. Salvana, S. Abdulah, H. Huang, H. Ltaief, Y. Sun, M. G. Genton, and D. E. Keyes, "High performance multivariate geospatial statistics on manycore systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 11, pp. 2719–2733, 2021.

[30] Q. Cao, Y. Pei, K. Akbudak, A. Mikhalev, G. Bosilca, H. Ltaief, D. Keyes, and J. Dongarra, "Extreme-scale task-based cholesky factorization toward climate and weather prediction applications," in *Proceedings of the Platform for Advanced Scientific Computing Conference*, 2020, pp. 1–11.

[31] K. Akbudak, H. Ltaief, A. Mikhalev, A. Charara, A. Esposito, and D. Keyes, "Exploiting data sparsity for large-scale matrix computations," in *European Conference on Parallel Processing*. Springer, 2018, pp. 721–734.

[32] J. M. C. Carpaye, J. Roman, and P. Brenner, "Design and analysis of a task-based parallelization over a runtime system of an explicit finite-volume cfd code with adaptive time stepping," *Journal of Computational Science*, vol. 28, pp. 439–454, 2018.

[33] H. Topcuoglu, S. Hariri, and M.-y. Wu, "Performance-effective and low-complexity task scheduling for heterogeneous computing," *IEEE transactions on parallel and distributed systems*, vol. 13, no. 3, pp. 260–274, 2002.

[34] Y. Xu, K. Li, J. Hu, and K. Li, "A genetic algorithm for task scheduling on heterogeneous computing systems using multiple priority queues," *Information Sciences*, vol. 270, pp. 255–287, 2014.

[35] K. R. Shetti, S. A. Fahmy, and T. Bretschneider, "Optimization of the heft algorithm for a cpu-gpu environment," in *2013 International conference on parallel and distributed computing, applications and technologies*. IEEE, 2013, pp. 212–218.

[36] H. J. Choi, D. O. Son, S. G. Kang, J. M. Kim, H.-H. Lee, and C. H. Kim, "An efficient scheduling scheme using estimated execution time for heterogeneous computing systems," *The Journal of Supercomputing*, vol. 65, pp. 886–902, 2013.

[37] M. A. Khan, "Scheduling for heterogeneous systems using constrained critical paths," *Parallel Computing*, vol. 38, no. 4-5, pp. 175–193, 2012.

[38] O. Beaumont, L.-C. Canon, L. Eyraud-Dubois, G. Lucarelli, L. Marchal, C. Mommessin, B. Simon, and D. Trystram, "Scheduling on two types of resources: a survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–36, 2020.

[39] A. K. Maurya and A. K. Tripathi, "On benchmarking task scheduling algorithms for heterogeneous computing systems," *The Journal of Supercomputing*, vol. 74, no. 7, pp. 3039–3070, 2018.

[40] B. Bramas, "Optimization and parallelization of the boundary element method for the wave equation in time domain," Ph.D. dissertation, Bordeaux, 2016.

[41] W. Fornaciari, G. Agosta, D. Atienza, C. Brandolese, L. Cammoun, L. Cremona, A. Cilardo, A. Farres, J. Flich, C. Hernandez *et al.*, "Reliable power and time-constraints-aware predictive management of heterogeneous exascale systems," in *Proceedings of the 18th International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation*, 2018, pp. 187–194.

[42] G. Agosta, W. Fornaciari, D. Atienza, R. Canal, A. Cilardo, J. F. Cardo, C. H. Luz, M. Kulczewski, G. Massari, R. T. Gavilá *et al.*, "The recipe approach to challenges in deeply heterogeneous high performance systems," *Microprocessors and Microsystems*, vol. 77, p. 103185, 2020.

[43] N. Grenèche, T. Menouer, C. Cérin, and O. Richard, "A methodology to scale containerized hpc infrastructures in the cloud," in *European Conference on Parallel Processing*. Springer, 2022, pp. 203–217.

[44] C. Cérin, N. Grenèche, and T. Menouer, "Executing traditional hpc application code in cloud with containerized job schedulers," in *High Performance Computing in Clouds: Moving HPC Applications to a Scalable and Cost-Effective Environment*. Springer, 2023, pp. 75–97.

[45] I. Dagal, K. Tanriöven, A. Nayir, and B. Akın, "Adaptive stochastic gradient descent (sgd) for erratic datasets," *Future Generation Computer Systems*, vol. 166, p. 107682, 2025.

[46] S. Santra, J.-W. Hsieh, and C.-F. Lin, "Gradient descent effects on differential neural architecture search: A survey," *IEEE Access*, vol. 9, pp. 89 602–89 618, 2021.

[47] A. S. Berahas, L. Cao, K. Choromanski, and K. Scheinberg, "Linear interpolation gives better gradients than gaussian smoothing in derivative-free optimization," *arXiv preprint arXiv:1905.13043*, 2019.

[48] S. C. Chapra, *Numerical methods for engineers*. Mcgraw-hill, 2010.

[49] H. Arabnejad and J. G. Barbosa, "List scheduling algorithm for heterogeneous systems by an optimistic cost table," *IEEE transactions on parallel and distributed systems*, vol. 25, no. 3, pp. 682–694, 2013.

APPENDIX

Algorithm 1: Ideal makespan Algorithm

```
function ideal_makespan(graph_info G, double coef_par_eff)
    nb_workers = G.#GPUS + G.#CPUS;
    if (G.#GPUS == 0.0) || (G.#CPUS == 0.0) then
        if (G.GPUS_PR == 0.0) then
            m = G.GPU_W / G.#GPUS;
        if (G.GPUS_PR == 0.0) then
            m = G.CPU_W / G.#CPUS;
        return m × coef_par_eff (nb_workers−1);
    Compute tgpu_minM and tcpu_minM using equations 5 and 6;
    if (G.GPU_W == G.GPU_PW) || (G.CPU_W==G.CPU_PW) then
        return MAX(tgpu_minM, tcpu_minM);
    gpu_rem_wl = compute the remaining GPU workload;
    cpu_rem_wl = compute the remaining CPU workload;
    if (gpu_rem_wl == 0.0) || (cpu_rem_wl == 0.0) then
        return MAX(tgpu_minM, tcpu_minM) + gpu_rem_wl / G.#GPUS
          + cpu_rem_wl / G.#CPUS;
    Compute ideal_makespan using equation 7;
    return ideal_makespan;
```

Algorithm 2: DSR-GD (Dedicated plus Shared Resource with Gradient Descent)

```
//-->Dedicated workers per app;
rem_gpus = nb_gpus;
rem_cpus = nb_cpus;
foreach graph G in graphs do
    gpus_dedicated = floor(G.GPU_W / SUM(GPU_W)) × nb_gpus;
    cpus_dedicated = floor(G.CPU_W / SUM(CPU_W)) × nb_cpus;
    while (gpus_dedicated) do
        Assign the (rem_gpus)^−th GPU to graph G;
        gpus_dedicated − −;
        rem_gpus − −;
    while (cpus_dedicated) do
        Assign the (rem_cpus)^−th CPU to graph G;
        cpus_dedicated − −;
        rem_cpus − −;
if (Stdd_speedup > SPEEDUP_STDD_LIMIT) || (Stdd_idealm >
  MAKESPAN_STDD_LIMIT)) then
    bartering();
//-->Shared workers between apps Using Gradient Descent;
Configure GD axis (X, Y) using rem_cpus and rem_gpus;
foreach axis in X, Y, Z do
    while (No convergence) do
        Use the best indexes from previous axes;
        Fix the value of the following axis;
        Search the best index in the axis using Gradient Descent with
          Min-Max ideal_makespan as objective function;
        keep track of the best solution;
```

Algorithm 3: MinMaxWL (Min-Max Workload balancing)

```
//-->Ensure each graph has at least one worker;
//----->Graphs with pure workload;
foreach graph G in graphs do
    if (G.GPU_PW != 0.0) then
        if (Remaining GPUs) then
            Assign one GPU to G;
        else
            Share one GPU with the under-loaded pure gpu graph, with
              backpropagation;
    if (G.CPU_PW != 0.0) then
        if (Remaining CPUs) then
            Assign one CPU to G;
        else
            Share one CPU with the under-loaded pure cpu graph, with
              backpropagation;
//----->Graphs without pure workload;
foreach graph G in graphs do
    if (G.CPU_PW == 0.0 && G.CPU_PW == 0.0) then
        if (G.CPU_W > G.GPU_W) then
            // GPU is faster;
            if (Remaining GPUs) then
                Assign one GPU to G;
            else
                Share one GPU with the under-loaded graph, with
                  backpropagation;
        else
            // CPU is faster;
            if (Remaining CPUs) then
                Assign one CPU to G;
            else
                Share one CPU with the under-loaded graph, with
                  backpropagation;
//-->Dist. Remaining Workers: Load-balancing using Min-Max;
while (Remaining GPUs Workers) do
    Assign one GPU to graph leading to Min-Max ideal_makespan;
while (Remaining CPUs Workers) do
    Assign one CPU to graph leading to Min-Max ideal_makespan;
```

Enhanced Network Bandwidth Prediction with Multi-Output Gaussian Process Regression

Shude Chen[1], Takayuki Nakachi[2]
Graduate School of Engineering and Science, University of the Ryukyus, Okinawa, Japan[1]
Information Technology Center, University of the Ryukyus, Okinawa, Japan[2]

***Abstract*—Modern network environments, especially in domains like 5G and IoT, exhibit highly dynamic and nonlinear traffic behaviors, posing significant challenges for accurate time series analysis and predictive modeling. Traditional approaches, including stochastic ARIMA and deep learning-based LSTM, frequently encounter difficulties in capturing rapid signal variations and inter-channel dependencies, often due to data sparsity or excessive computational cost. To address these issues, this paper proposes a Multi-Output Gaussian Process (MOGP) framework augmented with a novel signal processing strategy, where additional signals are generated by summing adjacent elements over multiple window sizes. Such multi-scale enrichment effectively leverages cross-channel correlations, enabling the MOGP model to discover complex temporal patterns in multi-channel data. Experimental results on real-world network traces highlight that the proposed method achieves consistently lower RMSE compared to conventional single-output or deep learning methods, thereby underscoring its value for robust bandwidth estimation. Our findings suggest that integrating MOGP with multi-scale augmentation holds promise for a wide range of predictive analytics applications, including resource allocation in 5G networks and traffic monitoring in IoT systems.**

***Keywords*—Network traffic prediction; Multi-Output Gaussian Process (MOGP); signal processing; time series analysis; predictive modeling; multi-channel data; IoT traffic monitoring; 5G networks**

I. Introduction

The proliferation of diverse network applications has led to increasingly congested environments, necessitating efficient resource allocation to maintain service quality and operational stability [1], [2]. With the rapid growth in both the volume and complexity of network traffic, the non-stationary and dynamic nature of traffic patterns has become a critical focus in telecommunications research [3], [4]. Accurate and timely network traffic prediction is essential to mitigate delays, prevent data loss, and optimize resource utilization in overloaded networks [5].

Recent advancements in traffic prediction technologies have enabled dynamic resource allocation through accurate traffic forecasting, typically categorized into short-term, medium-term, and long-term forecasts [6]. Short-term forecasting focuses on real-time predictions for the next few seconds or minutes, essential for time-sensitive applications. In contrast, long-term forecasting relies on historical trends over extended periods to guide strategic planning. Medium-term forecasting, which bridges the gap between these two, is particularly challenging due to the inherent variability in network traffic spanning minutes to hours.

Traditional prediction models, such as the Auto-Regressive Integrated Moving Average (ARIMA) [7], excel in capturing linear trends but struggle with the nonlinear and dynamic nature of modern network traffic [8], [9]. Deep learning models, such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks [6], offer superior nonlinear forecasting capabilities but often face challenges with overfitting, high computational demands, and large data requirements.

Gaussian Processes (GP) [10], [11] have emerged as a powerful tool for modeling nonlinear data and providing uncertainty quantification, making them particularly effective for network traffic forecasting with limited data [12], [13]. However, traditional GP methods limit their ability to adapt to dynamic and multi-scale traffic variations. This results in prediction errors over extended horizons, complicating medium-term forecasting tasks.

To address these challenges, Building on the strengths of existing GP-based multi-slot-ahead prediction models [12]-[15], this study introduces a Multi-Output Gaussian Process (MOGP) framework, incorporating an innovative signal augmentation strategy. The proposed approach introduces two key innovations: 1) the utilization of inter-channel correlations to improve prediction accuracy and reduce error propagation over extended periods, and 2) a signal augmentation strategy that generates additional input features by summing adjacent data points with varying window sizes. By enriching the input dataset with these newly generated signals, the model captures complex temporal dependencies and inter-channel interactions more effectively.

The main contributions of this paper are:

- Proposed a network bandwidth prediction framework based on Multi-Output Gaussian Process (MOGP), combined with a signal enhancement strategy.
- Enriched the input feature space and improved prediction accuracy by generating adjacent element sums with varying window sizes.
- Validated the effectiveness of the method through experiments on real multi-channel network traffic datasets, significantly reducing prediction errors.

The remainder of this paper is organized as follows: Section II reviews related methodologies and their limitations. Section III details the proposed forecasting approach, including the signal augmentation strategy and the MOGP framework. Section IV presents the dataset and experimental setup. Section

V discusses simulation results and evaluates the model's performance. Finally, Section VI concludes the paper and outlines future research directions.

II. Related Work

The evolution of network bandwidth prediction techniques has progressed from classical statistical approaches to sophisticated machine learning methods.

A. Methods Based on Probabilistic Processes

Early models like ARIMA have been extensively utilized for time series forecasting in network traffic. These models, though effective for linear trends, struggle with the complex and non-stationary nature of network data [15]. Enhancements like ARIMA/GARCH have been proposed to handle long-term dependencies, but these methods still face challenges in tracking rapidly changing network traffic characteristics [5]. These methods typically focus on single time series data, capturing patterns and periodicity over time, but have limitations in modeling long-term dependencies.

B. Machine Learning Methods

Recent advancements have seen the application of neural network architectures, such as LSTMs, which are adept at capturing nonlinearities and temporal dependencies in traffic data, providing significant improvements over traditional models [16]. Despite their effectiveness, deep learning models require substantial amounts of training data and computational resources, which can be impractical in dynamic network environments. Additionally, methods such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) have been employed to capture temporal and spatial dependencies in network traffic [17]. However, these machine learning methods are often limited by their high computational cost and susceptibility to overfitting in small datasets.

C. Traditional Gaussian Processes

Gaussian Processes (GP) have emerged as a powerful tool for network traffic prediction, given their ability to model uncertainty and non-linear dynamics of network traffic [8]. Particularly, Multi-Output Gaussian Processes (MOGP) have shown great potential in leveraging the correlations among multiple network channels to enhance prediction accuracy. GP models offer the advantage of providing uncertainty measures along with predictions, crucial for robust network management [18]. Incorporating convolutional structures in GPs, researchers have managed to capture spatiotemporal characteristics of traffic data more effectively, enhancing predictive performance [18]. These models have been particularly useful in scenarios involving large-scale data from multiple network sensors [19].

D. Recent Advances and Our Contribution

Recent studies, such as the work by Wang et al., explore a GP-based online learning framework for multi-slot-ahead network traffic forecasts [13]. They emphasize the evolving nature of network traffic and propose a dynamic kernel design using Spectrum Mixture (SM) functions and Process Convolution (PConv) to capture complex traffic patterns over different time scales. This approach addresses both tracking capability and prediction horizon challenges, demonstrating superior performance through simulations [13].

Emerging methodologies such as Graph Gaussian Processes (Graph-GP) [4] and Multi-channel Transformer-based models (MCformer) [20] further enhance the ability to model spatiotemporal correlations and interdependencies in network traffic. Graph-GP adapts well to scenarios with missing data and non-stationary traffic, while MCformer utilizes attention mechanisms to effectively capture multi-channel dependencies, proving highly effective in dynamic environments.However, these methods require extensive labeled datasets and suffer from high computational complexity.

E. Gap in Literature

- Existing deep learning models such as RNN and LSTM suffer from overfitting and large data requirements.
- Conventional GP-based methods do not efficiently utilize inter-channel correlations, limiting their predictive accuracy.
- Recent studies on multi-output regression models for network forecasting have not explored adjacent signal augmentation as a feature engineering technique.
- Our work extends these methodologies by introducing a signal augmentation strategy within an MOGP framework, improving multi-step forecasting accuracy.

Our proposed method extends these traditional GP approaches by using a multi-channel framework. By generating new signals from the original signals through summing adjacent signals [7], we capture the underlying characteristics of the data more effectively. Initially, the two channels with the strongest correlations are selected as the original data. Subsequently, new signals are created by using adjacent sums with varying window sizes for each channel signal. This approach enhances the prediction accuracy by exploiting inter-channel correlations more effectively, resulting in a robust framework for multi-channel network traffic forecasting.

III. Proposed Method

This section introduces the proposed Multi-Output Gaussian Process (MOGP) model for network bandwidth prediction. The method systematically integrates correlation analysis, signal transformation, and a Gaussian Process framework to improve prediction accuracy. The workflow, illustrated in Fig. 1, involves selecting highly correlated input signals, generating new signals with varying temporal scales, and training the MOGP model for predictive tasks.

A. Flowchart Overview of the Prediction Process

The overall workflow is illustrated in Fig. 1. The process begins with correlation analysis to identify a subset of highly correlated signals. These signals are then enriched by generating new signals through adjacent sums with varying window sizes. The original and generated signals are subsequently input into the MOGP model, which leverages their temporal and inter-channel dependencies for bandwidth prediction.

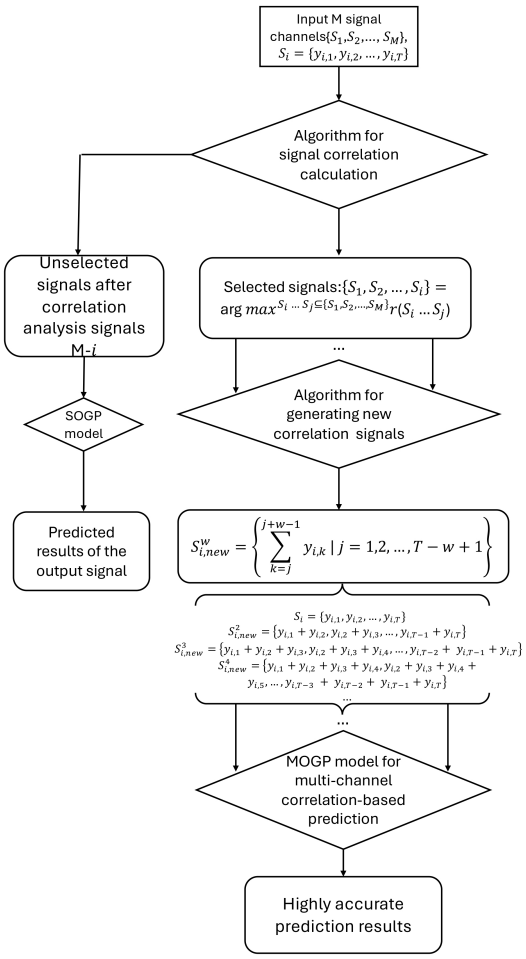


Fig. 1. Flowchart of the proposed prediction method.

B. Input Signal Definition and Correlation Analysis

Time Index and Signals. We define the discrete time axis as

$$t = 1,\, 2,\, \ldots,\, T. \tag{1}$$

For each time t, we have M signals $\{\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_M\}$, where

$$\mathbf{S}_i = \{y_{i,1},\, y_{i,2}, \ldots, y_{i,T}\}, \quad i = 1, \ldots, M, \tag{2}$$

and $y_{i,t}$ is the value of the i-th signal at time t.

Let the output signals be denoted as

$$\{\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_M\}, \tag{3}$$

where each $\boldsymbol{S}_i = \{y_{i,1}, y_{i,2}, \ldots, y_{i,T}\}$ represents a sequence of data points over time.

1) Correlation Metric: The Pearson correlation coefficient is used to quantify the linear relationship between two signals $\mathbf{S}_k$ and $\mathbf{S}_l$[21], [22]:

$$r(\mathbf{S}_k, \mathbf{S}_l) = \frac{\sum_{t=1}^{T}(y_{k,t} - \bar{y}_k)(y_{l,t} - \bar{y}_l)}{\sqrt{\sum_{t=1}^{T}(y_{k,t} - \bar{y}_k)^2}\sqrt{\sum_{t=1}^{T}(y_{l,t} - \bar{y}_l)^2}}, \tag{4}$$

where $\bar{y}_k$ and $\bar{y}_l$ are the mean values of $\mathbf{S}_k$ and $\mathbf{S}_l$, respectively.

C. Selection of Highly Correlated Signals

Identifying highly correlated signals is crucial for reducing model complexity while maintaining predictive accuracy.

To identify the most relevant signals, the following selection criterion is applied. Let L be the maximum number of signals we wish to select. We choose:

$$\{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_L\} = \{\mathbf{S}_k \mid r(\mathbf{S}_k, \mathbf{S}_l) > \text{threshold}, \quad \forall \mathbf{S}_k, \mathbf{S}_l \in \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_M\}\}, \tag{5}$$

where,

- $r(\mathbf{S}_k, \mathbf{S}_l)$: Correlation coefficient between signals $\mathbf{S}_k$ and $\mathbf{S}_l$,
- threshold: Minimum correlation value required for a signal to be included. This yields L original signals with the highest inter-correlations,
- $\{\mathbf{S}_k\}$: Subset of signals that meet the correlation criterion.

D. Generation of New Signals

To enrich the input features, new signals are generated by summing adjacent data points in the selected signals [12]. For a signal $\mathbf{S}_i$ of length T (i.e., $\mathbf{S}_i = \{y_{i,1}, y_{i,2}, \ldots, y_{i,T}\}$), and a chosen window size w, the new signal $\mathbf{S}_{i,\text{aug}}^{(w)}$ is computed as:

$$\mathbf{S}_{i,\text{aug}}^{(w)} = \Big\{ \sum_{k=j}^{j+w-1} y_{i,k} \;\Big|\; j = 1, 2, \ldots, T - w + 1 \Big\}. \tag{6}$$

where,

- $y_{i,k}$: The k-th data point of the original signal $\mathbf{S}_i$,
- k: Index of a data point within the summation window, ranging from j to $j + w - 1$,
- j: Starting index of the summation window in $\mathbf{S}_i$,
- w: Window size, i.e., the number of consecutive data points to be summed,
- T: Total length (in data points) of the original signal $\mathbf{S}_i$,
- $\mathbf{S}_{i,\text{aug}}^{(w)}$: The newly generated signal of length $T-w+1$.

For example,

- For $w = 2$ (adjacent two-term sums):

$$\mathbf{S}_{i,\text{aug}}^{(2)} = \{\, y_{i,1} + y_{i,2},\; y_{i,2} + y_{i,3},\; \ldots,\; y_{i,T-1} + y_{i,T} \}. \tag{7}$$

- For $w = 3$ (adjacent three-term sums):

$$\mathbf{S}_{i,\text{aug}}^{(3)} = \Big\{ y_{i,1} + y_{i,2} + y_{i,3}, \quad y_{i,2} + y_{i,3} + y_{i,4}, \quad \ldots, \quad y_{i,T-2} + y_{i,T-1} + y_{i,T} \Big\}. \tag{8}$$

Signal Length Note: Hence, if the original signal $\mathbf{S}_i$ has length T, the generated signal $\mathbf{S}_{i,\text{aug}}^{(w)}$ has length $T - w + 1$. By varying w, we can capture different scales of temporal dependencies, thereby enhancing the feature set for the Multi-Output Gaussian Process (MOGP) model.

E. Visualization of Signals and Generated Signals

Fig. 2 and 3 illustrate the transformation from two original signals, $\mathbf{S}_1$ and $\mathbf{S}_2$, into their augmented versions using window sizes $w \in \{2, 3, 4\}$.

a) Original Signals: Fig. 2 shows two original signals, $\mathbf{S}_1$ and $\mathbf{S}_2$, used as the starting input for the prediction process. The periodic nature and daily patterns of these signals are evident, highlighting their suitability for modeling temporal dependencies.

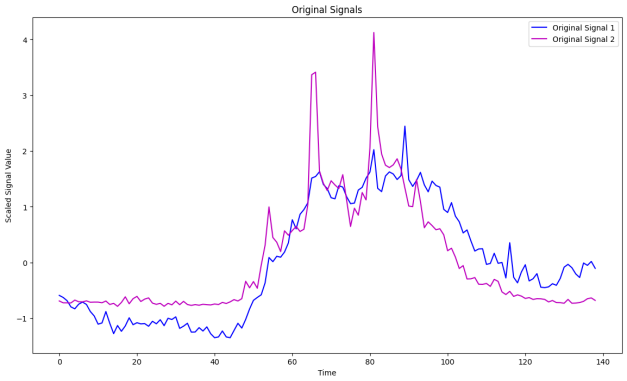


Fig. 2. Original signals ($\mathbf{S}_1$ in blue and $\mathbf{S}_2$ in magenta) with scaled values over time.

b) Augmented Signals via Adjacent Sums: Given a signal $\mathbf{S}_i = \{y_{i,1}, y_{i,2}, \ldots, y_{i,T}\}$, we define its w-term adjacent-sum (augmented signal) as

$$\mathbf{S}_{i,\text{aug}}^{(w)} = \{ \underbrace{y_{i,1} + y_{i,2} + \cdots + y_{i,w}}_{w \text{ terms}}, \underbrace{y_{i,2} + \cdots + y_{i,w+1}}_{w \text{ terms}}, \ldots \}. \tag{9}$$

As w increases, these sums capture progressively larger local trends. Fig. 3 shows newly generated signals (for $w = 2, 3, 4$) from the original $\mathbf{S}_1$ and $\mathbf{S}_2$, demonstrating how the adjacent-sum approach broadens temporal coverage.

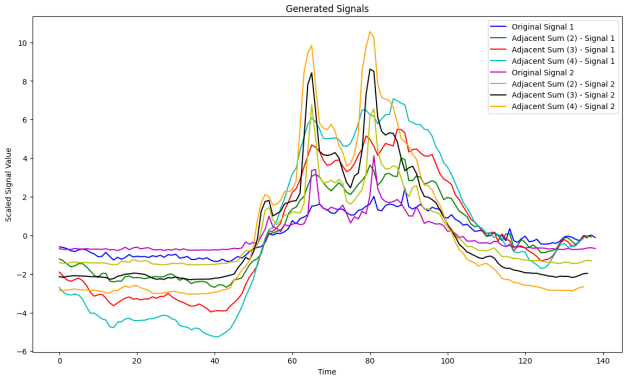


Fig. 3. Newly generated signals from $\mathbf{S}_1$ and $\mathbf{S}_2$, illustrating different window sizes w.

c) Defining MOGP$(L, \tilde{N})$: Suppose we select L original signals based on correlation analysis, and for each original signal, we generate $\tilde{N}$ adjacent-sum signals (e.g. multiple choices of w). We then denote our model by

$$\text{MOGP}(L, \tilde{N}), \tag{10}$$

meaning:

- L is the number of *original* (highly correlated) signals chosen,
- $\tilde{N}$ is the total number of *augmented* signals generated from those L originals,
- Altogether, there are $N = L + \tilde{N}$ channels in the resulting multi-output GP.

For instance,

$$\text{MOGP}(2, 6) \tag{11}$$

indicates two original signals and six augmented signals, making $N = 8$ jointly modeled channels.

d) Example: Generating Up to Eight Channels: If $\mathbf{S}_p$ and $\mathbf{S}_q$ are the selected original signals, then for $w = 2, 3, 4$ each signal spawns three augmented versions:

$$\mathbf{S}_{p,\text{aug}}^{(2)}, \quad \mathbf{S}_{p,\text{aug}}^{(3)}, \quad \mathbf{S}_{p,\text{aug}}^{(4)}, \quad \mathbf{S}_{q,\text{aug}}^{(2)}, \quad \mathbf{S}_{q,\text{aug}}^{(3)}, \quad \mathbf{S}_{q,\text{aug}}^{(4)}. \tag{12}$$

Thus, we obtain up to eight channels in total:

$$\underbrace{\mathbf{S}_p, \mathbf{S}_q}_{\text{original signals}}, \underbrace{\mathbf{S}_{p,\text{aug}}^{(2)}, \mathbf{S}_{q,\text{aug}}^{(2)}}_{\text{2-term}}, \underbrace{\mathbf{S}_{p,\text{aug}}^{(3)}, \mathbf{S}_{q,\text{aug}}^{(3)}}_{\text{3-term}}, \underbrace{\mathbf{S}_{p,\text{aug}}^{(4)}, \mathbf{S}_{q,\text{aug}}^{(4)}}_{\text{4-term}}. \tag{13}$$

Each augmented channel reflects a different local summation scale, thereby enriching the feature set that the Multi-Output Gaussian Process (MOGP) exploits for multi-step prediction. Hence, the MOGP$(L, \tilde{N})$ approach systematically combines correlated original signals with their adjacency-sum signals, allowing finer control over both short- and long-term dependencies.

F. Multi-Output Gaussian Process (MOGP) Prediction Framework

After identifying highly correlated signals and generating augmented signals (Sections III-D–III-E), we obtain N total channels (*original* + *augmented*). These N channels are collectively modeled using a *Multi-Output Gaussian Process* (MOGP), which exploits *both* temporal correlations *and* inter-channel correlations to enhance predictive accuracy.

1) Input/Output Setup and Kernel:

a) Inputs $\mathbf{X}$: We consider T time indices or feature vectors $x_1, x_2, \ldots, x_T \in \mathcal{X}$, collected in $\mathbf{X} = \{x_1, \ldots, x_T\}$. (Each x_t can be a scalar time step or a multi-dimensional feature.)

b) Outputs $\mathbf{Y}$: All original and augmented signals together form $N = L + \tilde{N}$ channels:

$$\mathbf{Y} = \{\mathbf{S}_1, \ldots, \mathbf{S}_L, \mathbf{S}_{1,\text{aug}}^{(w)}, \ldots, \mathbf{S}_{L,\text{aug}}^{(w)}\}. \tag{14}$$

We collect these N channels (each of length T) into a *multi-channel output matrix* $\mathbf{Y} \in R^{N \times T}$. In a block sense, we may write

$$\mathbf{Y} = [\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_N], \tag{15}$$

where each $\mathbf{S}_i \in R^T$. Equivalently, to index individual data values.

c) MOGP Kernel: A Gaussian Process (GP) is defined by a mean function (often taken as zero) and a covariance (kernel) function capturing similarities among inputs. For time-series forecasting, we frequently adopt an RBF+noise kernel [23], [24]:

$$k(x, x') \;=\; \theta_1 \exp\Big(-\frac{\|x - x'\|^2}{2\,\theta_2}\Big) \;+\; \theta_3\,\delta(x, x'), \qquad (16)$$

where $\theta_1, \theta_2, \theta_3$ are hyperparameters, and $\delta(x, x')$ is the Kronecker delta for noise. In a *multi-output* setting, we extend this kernel to model both *within-channel* and *cross-channel* covariances, often via a block-structured approach (e.g. $K_{\text{MOGP}} = B \otimes K_{\text{RBF}}$, where B is an $N \times N$ matrix of inter-channel correlations).

2) MOGP Posterior for Single-Step and Multi-Step Forecasts:

a) MOGP Prior: We place a joint GP prior over all N latent functions $\{f_i(\cdot)\}$ corresponding to the N channels:

$$\begin{bmatrix} f_1(\mathbf{X}) \\ f_2(\mathbf{X}) \\ \vdots \\ f_N(\mathbf{X}) \end{bmatrix} \;\sim\; \mathcal{GP}\Big(\mathbf{0},\, K_{\text{MOGP}}(\mathbf{X}, \mathbf{X})\Big). \qquad (17)$$

Observations $\mathbf{Y}$ are then linked to these $f_i(\mathbf{X})$ values (e.g. via additive Gaussian noise).

b) Single-Step Prediction: To predict the next time point $t + 1$ for each channel i, we collect all available data up to t. Let $\mathbf{X} = \{x_1, \ldots, x_t\}$ and $\mathbf{Y}_{\text{train}}$ be the corresponding outputs. Under the MOGP prior, we form the joint Gaussian $\{\mathbf{Y}_{\text{train}},\, f(x_{t+1})\}$ and condition on $\mathbf{Y}_{\text{train}}$. The resulting posterior mean for channel i is[13]:

$$\hat{y}_i(t+1) \;=\; \mu_i\big(x_{t+1}\big), \quad i = 1, \ldots, L, \qquad (18)$$

where $\mu_i(\cdot)$ is the i-th component of the MOGP posterior.

c) Multi-Step (K-Step) Prediction: To forecast K future points $\{t + 1, \ldots, t + K\}$ for each channel, we consider two main strategies.

Iterative (Recursive) Approach: We repeatedly use predicted values as though they are observed for the next step:

$$\begin{aligned} &\hat{y}_i(t+1) = \mu_i\big(x_{t+1}\big), \\ &\hat{y}_i(t+2) = \mu_i\big(x_{t+2} \mid \hat{y}_i(t+1)\big), \\ &\hat{y}_i(t+3) = \mu_i\big(x_{t+3} \mid \hat{y}_i(t+1),\, \hat{y}_i(t+2)\big), \\ &\quad \ldots, i = 1, \ldots, L, \end{aligned} \qquad (19)$$

This simple approach can accumulate errors but allows easy updates for adjacency-sum channels. Inspired by an Eq. (21)-style logic, we can define partial sums or increments. For instance, if

$$\widetilde{y}_i\big(t + k\big) \;=\; f_{i+1}\big(\mathbf{X}^*\big) \quad \text{for } k = 0, \qquad (20)$$

and

$$\widetilde{y}_i(t + k) \;=\; f_{i+2}(\mathbf{X}^*) \;-\; f_{i+1}(\mathbf{X}^*) \quad \text{for } k = 1, \ldots, K) \qquad (21)$$

then each new adjacency-sum is obtained by subtracting the previously predicted partial sum. Meanwhile, the final predicted channel values are:

$$\hat{y}_i(t + k) \;=\; \mu_i\big(x_{t+k}\big), \qquad (22)$$

ensuring consistency among all signals as they roll forward in time.

Direct Joint Approach: Instead of predicting each future time point separately, we form a single joint Gaussian over $\{x_{t+1}, \ldots, \mathbf{x}_{t+K}\}$ for all N channels[13], [25]:

$$\begin{bmatrix} \mathbf{Y}_{\text{train}} \\ \mathbf{F}_* \end{bmatrix} \;\sim\; \mathcal{N}\Big(\mathbf{0},\, \begin{bmatrix} K_{\text{MOGP}}(\mathbf{X}, \mathbf{X}) & K_{\text{MOGP}}(\mathbf{X}, \mathbf{X}_*) \\ K_{\text{MOGP}}(\mathbf{X}_*, \mathbf{X}) & K_{\text{MOGP}}(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix}\Big). \qquad (23)$$

where $\mathbf{X}_* = \{x_{t+1}, \ldots, x_{t+K}\}$ and $\mathbf{F}_* \in R^{N \times K}$ denotes the unknown function values $\{f_i(x_{t+k})\}$. Conditioning on $\mathbf{Y}_{\text{train}}$ produces a posterior with mean $\hat{\mathbf{F}}_*$. Its (i, m)-th entry gives $\hat{y}_i(t + m)$. This one-shot method often yields better long-horizon accuracy but incurs higher computational cost due to the larger covariance block.

Overall, both approaches yield $\{\hat{y}_i(t + k) \mid i = 1, \ldots, N;\, k = 1, \ldots, K\}$. The iterative method is simpler yet can accumulate errors; the direct method jointly captures cross-step correlations but is more expensive. In practice, adjacency-sum signals can still be integrated into either approach by updating or subtracting previously predicted sums as new time steps unfold.

3) Recap and Advantages: By modeling all channels (original + augmented) within a unified MOGP, we exploit *inter-channel* correlations and *different temporal scales* simultaneously. This often reduces uncertainty in multi-step forecasting, compared to treating each channel as an independent single-output GP. Hence, our MOGP framework is well-suited to network traffic scenarios, where multiple correlated signals (e.g. raw measurements and adjacent-sum signals) provide complementary information for bandwidth prediction.

G. MOGP-Based Framework for Multi-Step Traffic Prediction

Fig. 4 presents a schematic of our Multi-Output Gaussian Process (MOGP) framework for multi-step bandwidth prediction. Building on the earlier steps of selecting highly correlated signals and generating adjacency-sum augmentations, this framework highlights three major components:

- Adjacency-Sum Channels (green blocks). From each original signal S_i (or S_j), we construct multiple augmented signals $S^{(w)}_{i,\text{aug}}$ by summing adjacent data points with different window sizes w. These channels capture various temporal scales (e.g. short-term bursts for $w = 2$ or medium-range trends for $w = 3, 4, \ldots$).
- Block-Structured MOGP Inference (blue boxes). All channels (original + augmented) are modeled jointly using a multi-output Gaussian Process. The block-structured kernel encodes both temporal dependencies (within each channel) and cross-channel correlations (across original and augmented signals). From this MOGP, we obtain a predictive mean $\mu(\mathbf{X}^*)$ and covariance $\Sigma(\mathbf{X}^*)$ for the forecast horizon K.

- Partial-Sum or Difference-Based Recovery (bottom). In multi-step prediction, each adjacency-sum channel (e.g. $S^{(2)}_{i,\text{aug}}$) can be used to recover the single-step prediction $\hat{y}_i(t+k)$ by subtracting already-predicted values. For instance,

$$\begin{aligned} \hat{y}_i(t+2) &= \hat{S}^{(2)}_{i,\text{aug}}(t+1) - \hat{y}_i(t+1), \\ \hat{y}_i(t+3) &= \hat{S}^{(3)}_{i,\text{aug}}(t+2) - \hat{S}^{(2)}_{i,\text{aug}}(t+2), \\ &\dots \end{aligned} \tag{24}$$

thus ensuring consistency among predicted sums and the underlying pointwise forecasts.

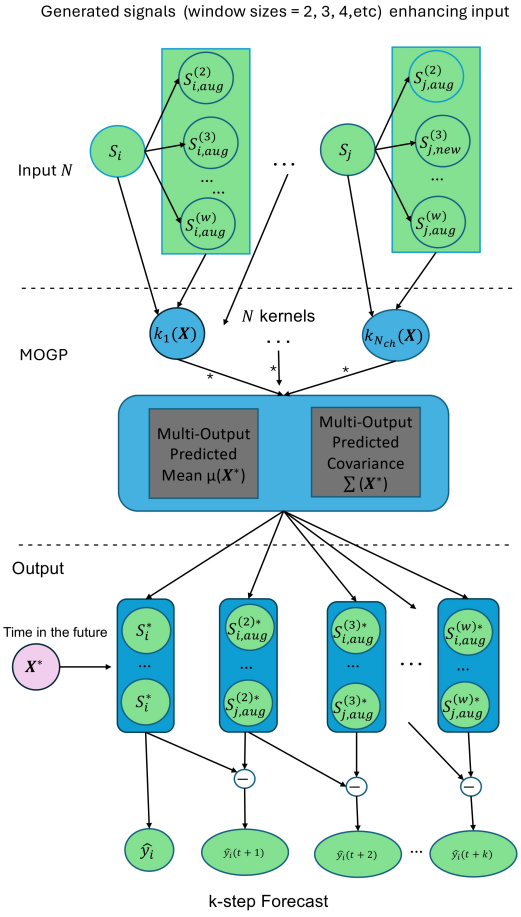


Fig. 4. The MOGP-based framework for multi-step forecasting. After generating adjacency-sum augmentations (green), a multi-output Gaussian Process (blue) provides joint predictions, which are then combined or subtracted to yield final outputs $\hat{y}_i(t+k)$.

a) Multi-Step Prediction: As discussed above, once the MOGP posterior is obtained, there are two main ways to forecast K time steps ahead:

- Iterative Approach: predict $\hat{y}_i(t+1)$, feed it back into the adjacency-sum channels, then predict $\hat{y}_i(t+2)$, and so forth. $\hat{y}_i(t+3)$ can be predicted in a similar manner until $\hat{y}_i(t+k)$.
- Direct (Block) Approach: form a single joint Gaussian over $\{\mathbf{x}_{t+1}, \dots, \mathbf{x}_{t+K}\}$ and solve for all $\hat{y}_i(t+k)$ in one shot, often achieving better accuracy (at a higher computational cost).

b) Advantages at Different Scales: By integrating adjacency-sum channels with an MOGP, the framework:

- *Captures multiple timescales* via small vs. larger summation windows,
- *Leverages cross-channel correlations*, reducing prediction uncertainty,
- *Yields robust multi-step forecasts*, as each newly predicted value can be consistently used to update the adjacency sums in subsequent steps.

In essence, this joint modeling of raw and augmented signals enables more accurate and stable bandwidth predictions across various forecasting horizons.

IV. Introduction to Experimental Model

This section provides background on the WIDE Project, describes the MAWI dataset used for our experiments, and presents an overview of traffic patterns and inter-signal correlations. The goal is to illustrate the data sources and motivations that underpin the MOGP-based forecasting framework described in Section III.

A. WIDE Project

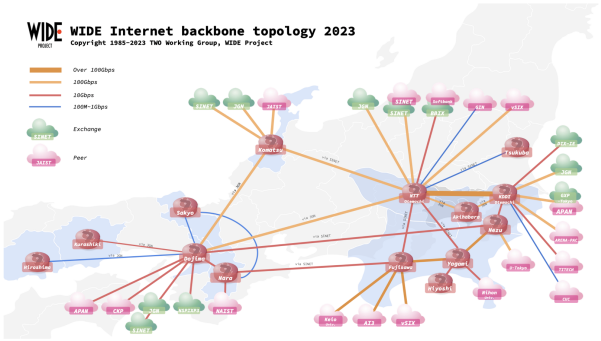


Fig. 5. 2023 WIDE backbone topology.

The Widely Integrated Distributed Environment (WIDE Project[1]) is a Japanese initiative focused on advancing Internet infrastructure. It involves a broad consortium of academic and research institutions that collaborate on designing, standardizing, and operating critical Internet technologies. Over the past decades, the WIDE Project has significantly contributed to IPv6 deployment, mobile Internet, and security innovations, and it continues to foster global partnerships among engineers and researchers.

As of 2023, the WIDE backbone (Fig. 5) consists of a Layer 2 and Layer 3 network interconnecting various Japanese and international sites, including San Francisco and Bangkok. This infrastructure provides a rich environment for collecting and analyzing real-world network traffic, which is crucial for exploring predictive modeling techniques such as Gaussian Process (GP) forecasting.

[1]WIDE Project: http://www.wide.ad.jp/.

1) MAWI Dataset Overview: The MAWI[2] (Measurement and Analysis on the WIDE Internet) group, a key part of the WIDE Project, collects and publicly shares traffic data from global ISP links. In particular, this study uses MAWI's traffic records from 2024/5/17, sampled at 10-minute intervals. As illustrated in Fig. 6, the dataset contains seven major aggregated flow signals and one time-stamp column, totaling eight columns in all. Each row represents a 10-minute observation, leading to 144 potential data points daily, though our final curated dataset has 143 valid entries after cleaning.

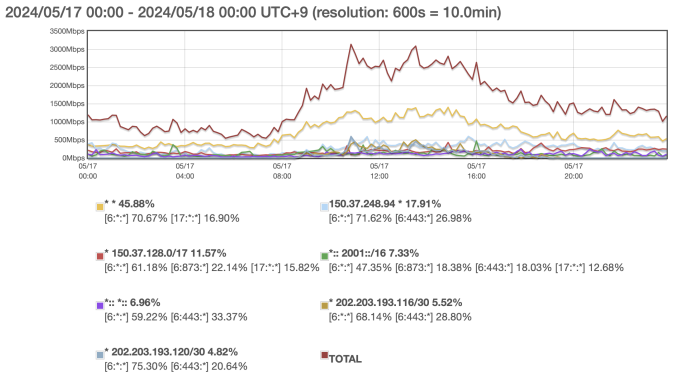


Fig. 6. Traffic data from 2024/5/17, courtesy of the MAWI dataset.

a) Data Preprocessing: We remove rows with missing values and normalize each signal to zero mean and unit variance. The time column is re-indexed to facilitate chronological analysis (e.g. minute-based or 10-minute-based indexing). This ensures that all signals are on a comparable scale before feeding them into the Multi-Output Gaussian Process (MOGP) model.

2) Traffic Pattern Analysis: A preliminary inspection of the MAWI traffic from 2024/5/17 reveals distinct diurnal patterns. Communication volumes begin to rise around 8 AM, peak during midday, and gradually decline overnight. Fig. 6 (from Section III)demonstrates these daily fluctuations. Such periodic behaviors indicate temporal correlations within each signal.

Furthermore, external factors like time zone synchronization across regions can induce correlated traffic surges among different signals. Leveraging these temporal and inter-signal dependencies is central to improving forecast accuracy in network traffic models.

3) Correlation Analysis: To quantify inter-signal relationships, we perform pairwise correlation among the seven flow signals. The resulting correlation matrix (Fig. 7) highlights various strengths of linear association. Signals that exhibit high pairwise correlation are prioritized for the MOGP input, as multi-output GP can exploit these cross-channel correlations more effectively.

The Pearson correlation coefficient, as defined in Eq. (4), is used to measure the strength of the linear relationship between two signals.

[2]MAWI Working Group Traffic Archive: https://mawi.wide.ad.jp/mawi/.

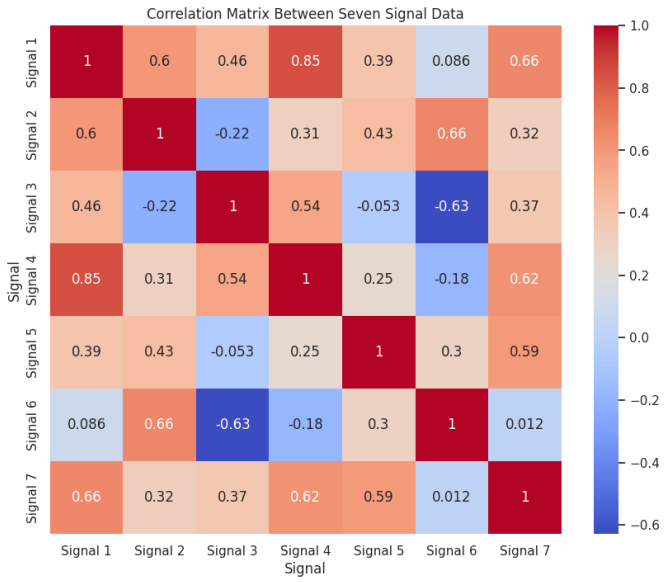


Fig. 7. Correlation matrix between seven MAWI signals, illustrating strong pairs for model input.

B. Data Setting and GP Models

Having introduced the WIDE Project and the MAWI dataset, we now discuss the GP model setup for our experiments. The experimental conditions are as follows:

- Training/Testing Split: We utilize the first T_{train} points (e.g. day 1's data) to train the GP model and reserve the subsequent points (e.g. final few observations) for testing.
- Window Size (Sliding Window): A fixed window size window_size is chosen (e.g. 50). In each step, the model sees $\mathbf{X}_{\text{train}}$ (time indices or features) and $\mathbf{y}_{\text{train}}$ (normalized signal values), then predicts the next point. The window slides forward by one time step iteratively.
- Multiple Channels (N=1,2,8): We compare three main GP setups:
  1) Single-Output GP ($N = 1$): Only one original signal (e.g. `signal1`) is modeled. This captures broad temporal trends but does not exploit cross-signal information.
  2) MOGP(2,0): We pick two highly correlated signals (e.g. $\mathbf{S}_p$ and $\mathbf{S}_q$), as identified by Pearson correlation. Then both signals are jointly modeled in a MOGP setting. This allows cross-channel correlation to inform the prediction, often improving accuracy compared to single-output.
  3) MOGP(2,6): We take two original signals $(\mathbf{S}_p, \mathbf{S}_q)$ plus their six adjacent-sum signals with window sizes $w = 2, 3, 4$. Hence, $N = 2 + 2 \times 3 = 8$ total input channels. By providing multi-scale features (short-term sums, medium-term sums, etc.), the MOGP(2,6) model can capture both short-range fluctua-

tions and daily-scale patterns across the correlated signals.

- Single-Step vs. Multi-Step Forecast:
  - Single-Step (1-step): Predict y_{t+1} given data up to time t as described in Eq. (18).
  - Multi-Step (K-step): Predict $\{y_{t+1}, \ldots, y_{t+K}\}$ either iteratively using Eq. (19) or via a direct block-covariance approach.

a) Summary of Experimental Model: Overall, the data feeding process, combined with the MOGP approach, is illustrated in Fig. 1. We use real traffic from the WIDE Project's MAWI dataset, apply standard preprocessing, select correlated signals, generate additional adjacent-sum signals if needed, and then run sliding-window GP predictions. We will perform performance metrics (RMSE) for single output versus multioutput comparisons, demonstrating how correlation exploitation and multi-scale features can enhance forecast accuracy.

1) Single-output Gaussian Process: In this scenario, only one signal is used as input for the prediction. The model captures the overall trends but struggles to predict sharp fluctuations due to the limited input information. Fig. 8 demonstrates the predicted results along with confidence intervals, showcasing the performance of the single-output GPR model for future-step prediction.

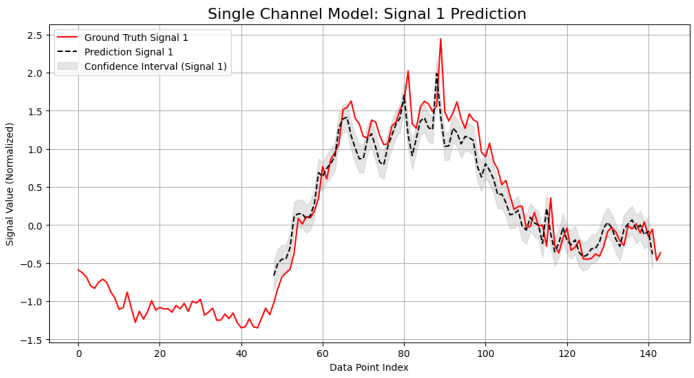


Fig. 8. Gaussian process prediction of signal trends with confidence intervals.

2) MOGP(2,0): Single-Step Prediction: In this setup, two highly correlated signals are used as inputs to the Multi-Output Gaussian Process (MOGP) model (here, MOGP(2,0) indicates two original signals and zero augmentations). This approach leverages inter-channel correlations to enhance the accuracy of predictions. Unlike the single-output model, the MOGP(2,0) model demonstrates improved robustness in capturing signal variations. Fig. 9 illustrates the predictive performance of the MOGP(2,0) model when considering two input signals, highlighting the advantage of incorporating multiple correlated channels.

3) MOGP(2,6): Single-Step Prediction: This model extends the input set by incorporating two original signals and their corresponding generated signals at varying temporal scales ($w = 2, 3, 4$), resulting in a total of eight input signals (2 original + 6 augmented). The inclusion of generated signals captures both short-term fluctuations and long-term patterns, significantly improving the accuracy of prediction. Fig. 10

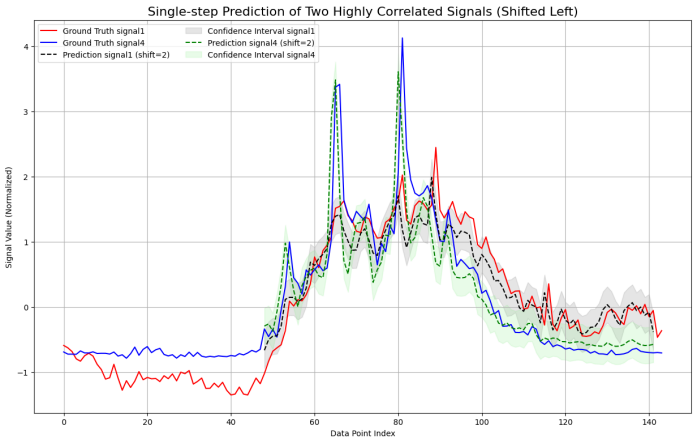


Fig. 9. Prediction results using the MOGP(2,0) model for single-step forecasting.

shows the results of the MOGP(2,6) model, demonstrating its superior predictive capacity compared to the single output and MOGP (2,0) models.

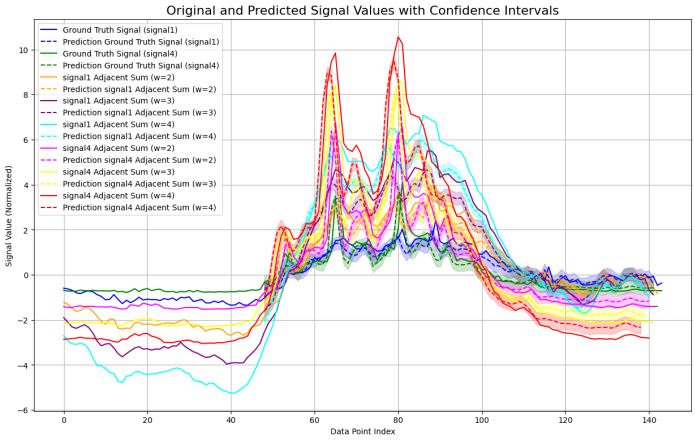


Fig. 10. Enhanced prediction accuracy using the MOGP(2,6) model for single-step forecasting.

4) MOGP(2,0): Multi-Step Prediction: In this experiment, the MOGP(2,0) model is extended to predict three future steps. This approach demonstrates the capability of the model to handle sequential forecasting tasks, utilizing inter-channel correlations effectively. Fig. 11 illustrates the predictive performance of the MOGP(2,0) model for three-step predictions.

5) MOGP(2,6): Multi-Step Prediction: The MOGP (2,6) model is used to predict three future steps, further leveraging the generated signals across varying temporal scales ($w = 2, 3, 4$). By incorporating more input signals, the model captures intricate temporal dependencies, enhancing prediction accuracy. Fig. 12 shows the prediction results for the MOGP(2,6) model.

V. Simulation Results and Discussion

A. Comparison and Summary of Experimental Results

The performance of the proposed models was evaluated using the Root Mean Squared Error (RMSE) metric, defined

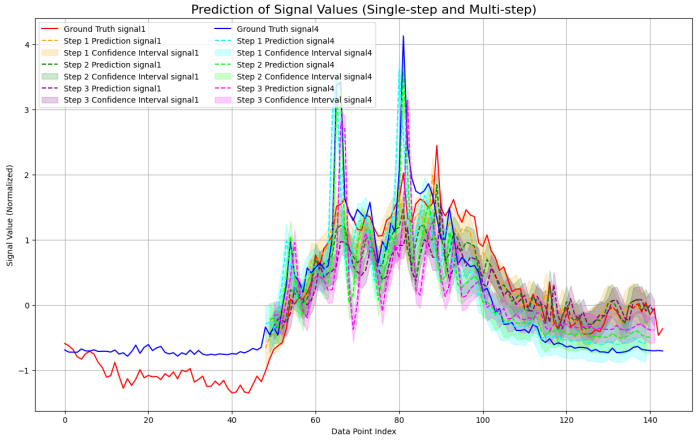


Fig. 11. Prediction results using the MOGP(2,0) model for three-step forecasting.

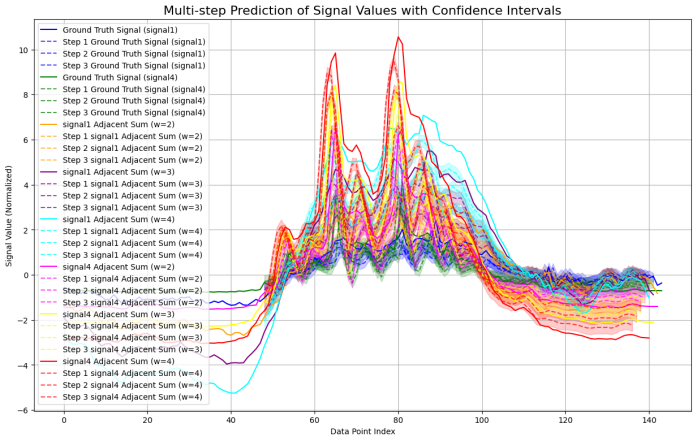


Fig. 12. Enhanced prediction accuracy using the MOGP(2,6) model for three-step forecasting.

as [26]:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum(\hat{y} - y_{\text{test}})^2}$$

where $\hat{y}$ represents the predicted values for the unknown input signal x_{test}, and y_{test} denotes the corresponding ground truth values.

Table I compares the RMSE of single-output, MOGP(2,0), and MOGP(2,6) models for predicting network traffic signals.

TABLE I. RMSE Comparison of Gaussian Process Models

Model	RMSE
Single-output Gaussian Process (1-step)	0.1967
MOGP(2,0) (1-step)	0.1552
MOGP(2,6) (1-step)	0.1273
MOGP(2,0) (3-step)	0.1943
MOGP(2,6) (3-step)	0.1607

The results in Table I demonstrate the effectiveness of the MOGP models in leveraging inter-channel correlations for prediction accuracy. The MOGP(2,6) model significantly outperforms the others, achieving the lowest RMSE value. This improvement highlights the importance of integrating generated signals with varying time scales into the prediction framework.

1) Error Analysis: Table II presents the percentage reduction in prediction error achieved by the multi-output models compared to the single-output Gaussian Process model. The error reduction underscores the effectiveness of multi-channel approaches in improving predictive performance.

TABLE II. Percentage Error Reduction by Multi-Output Models

Comparison	Error Reduction (%)
Single-output vs. MOGP(2,0) (1-step)	21.2%
Single-output vs. MOGP(2,6) (1-step)	35.3%
MOGP(2,0) (1-step) vs. MOGP(2,6) (1-step)	17.9%
MOGP(2,0) (3-step) vs. MOGP(2,6) (3-step)	17.3%

The error analysis results indicate that the multi-output models significantly reduce RMSE compared to the single-output model. MOGP(2,6), in particular, achieves substantial error reduction, confirming the advantages of utilizing generated signals to capture both short-term and long-term signal variations.

B. Comparison with Recent Methods

A closely related study by Wang *et al.* [13] has already demonstrated that Gaussian process (GP)–based forecasting can outperform statistical approaches like ARIMA as well as deep learning models such as LSTM, RC-LSTM, and RNN. In particular, their method leverages a multi-output GP framework and multi-scale adjacency-sum augmentation for multi-slot-ahead traffic prediction.

Our proposed approach builds upon and extends the work in [13] by introducing **multi-channel** input signals. Rather than relying on a single channel with adjacency-sum signals, we first select multiple original signals based on their correlations, then generate adjacency-sum augmentations for each. This strategy further exploits cross-channel relationships, allowing the model to capture subtle variations across different but related traffic flows. By doing so, we retain the core strengths of multi-output GP and adjacency-sum augmentation from [13], while enhancing predictive performance through an expanded, correlation-aware input space. Our experimental results confirm that this multi-channel extension yields higher accuracy than single-channel baselines, illustrating the practical benefits of incorporating inter-channel correlations into GP-based network traffic forecasting.

Although direct numerical comparisons are hindered by different datasets and experimental conditions, we attribute our method's strong performance to two main factors:

- *Multi-Output Modeling:* By capturing inter-channel correlations, MOGP leverages shared temporal dynamics across multiple signals—an advantage that traditional single-output methods lack.
- *Multi-Scale Adjacency-Sum Augmentation:* Generating additional signals by summing adjacent data points over varying window sizes enriches the feature space, enabling the model to capture both short-term bursts and longer-range trends more effectively.

Overall, these comparisons underscore the benefits of integrating multi-channel Gaussian Processes with adjacency-sum augmentation, offering robust predictive accuracy and principled uncertainty estimates for complex network traffic scenarios.

C. Assumptions and Limitations

This study makes primary assumptions: The selected subset of MAWI channels is sufficiently representative of typical network traffic. Although these assumptions are practical for many real-world scenarios, they may limit generalization to networks with irregular sampling patterns or strong non-stationary bursts.

Another limitation lies in the computational overhead of handling large-scale datasets. Because our approach generates multiple augmented signals for each original channel, the dimensionality grows along with the dataset size, increasing both memory usage and training time for multi-output Gaussian Processes. In future work, we plan to investigate sparsity-inducing kernels and online GP methods to better accommodate high-volume streaming data.

VI. Conclusions and Future Work

In this study, we proposed a Multi-Output Gaussian Process (MOGP) model for network traffic prediction, introducing a novel approach that integrates original input signals with additional correlated signals generated using adjacent terms. This methodology leverages inter-channel correlations and multi-scale temporal dependencies to improve prediction performance. Experimental results validated through RMSE metrics demonstrated that the MOGP model consistently outperforms single-output Gaussian Processes, showcasing robust performance across both one-step and three-step prediction tasks. The MOGP(2,6) model achieved the best overall results, reducing prediction errors by 35.3% compared to the single-output baseline. These findings confirm that incorporating correlated signals significantly enhances prediction accuracy. Furthermore, the model demonstrated consistent effectiveness in single- and multi-step forecasting scenarios, highlighting its adaptability to various temporal scales and complex datasets.

Future work will focus on exploring alternative kernel functions and hyperparameter optimization strategies to further improve prediction precision and computational efficiency. We also intend to extend our framework to 5G and IoT traffic prediction, and to real-time or large-scale environments, evaluating the model's scalability and adaptability under streaming conditions. Another promising direction is to adapt our MOGP approach for anomaly detection. Such anomalies not only consume additional network resources but also pose significant risks to overall network performance and security [27]. By exploiting the pervasive inter-channel correlations, our framework could identify deviations from typical diurnal or weekly patterns. For example, if two channels $\mathbf{S}_k$ and $\mathbf{S}_l$ with similar usage trends both show high traffic after 8 AM and taper off after 7 PM, a sudden increase in $\mathbf{S}_k$ without a corresponding change in $\mathbf{S}_l$ might signal anomalous behavior. This line of research could play a vital role in proactive network management and early threat detection.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP22K04089.

References

[1] Y. Chen, S. Jain, V. K. Adhikari, Z. L. Zhang, and K. Xu, "A first look at inter-data center traffic characteristics via yahoo! datasets," Proc. of IEEE INFOCOM 2011, pp. 1620-1628, Apr. 2011.

[2] J. Koo, V. B. Mendiratta, M. R. Rahman, and A. Walid, "Deep reinforcement learning for network slicing with heterogeneous resource requirements and time varying traffic dynamics," Proc. of IEEE CNSM 2019, pp. 1-5, Oct. 2019.

[3] A. Baiocchi, Network Traffic Engineering Stochastic Models and Applications, Hoboken, NJ, USA:Wiley, 2020.

[4] Mehrizi, S., & Chatzinotas, S. (2022). Network traffic modeling and prediction using graph Gaussian processes. IEEE Access, 10, 132644–132655.

[5] M. Beshley, et al. "End-to-End QoS 'smart queue' management algorithms and traffic prioritization mechanisms for narrow-band internet of things services in 4G/5G networks." Sensors, vol. 20, no. 8, p. 2324, 2020.

[6] C. Gijon, et al. "Long-term data traffic forecasting for network dimensioning in LTE with short time series." Electronics, vol. 10, no. 10, p. 1151, 2021.

[7] B. Zhou, D. He, Z. Sun, and W. H. Ng, "Network traffic modeling and prediction with ARIMA/GARCH," Proc. of HET-NETs 2005, pp. 1-10, Jul. 2005.

[8] A. Azari, P. Papapetrou, S. Denic and G. Peters, "Cellular traffic prediction and classification: A comparative evaluation of LSTM and ARIMA", Proc. Int. Conf. Discovery Sci., pp. 129-144, 2019.

[9] A. Azzouni and G. Pujolle, "NeuTM: A neural network-based framework for traffic matrix prediction in SDN", Proc. IEEE/IFIP Netw. Oper. Manage. Symp., pp. 1-5, Apr. 2018.

[10] A. Bayati, K.-K. Nguyen and M. Cheriet, "Gaussian process regression ensemble model for network traffic prediction", IEEE Access, vol. 8, pp. 176540-176554, 2020.

[11] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[12] Y. Wang, T. Nakachi, T. Inoue and T. Mano, "Adaptive multi-slot-ahead prediction of network traffic with Gaussian process," 2021 IEEE Global Communications Conference (GLOBECOM), Madrid, Spain, 2021, pp. 1-6, doi: 10.1109/GLOBECOM46510.2021.9685249.

[13] Y. Wang, T. Nakachi, and W. Wang, "Pattern discovery and multi-slot-ahead forecast of network traffic: A revisiting to Gaussian process," IEEE Transactions on Network and Service Management, 2022.

[14] M. van der Wilk, et al. "A framework for interdomain and multioutput Gaussian processes." arXiv preprint arXiv:2003.01115, 2020.

[15] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time Series Analysis: Forecasting and Control. John Wiley & Sons, 2015.

[16] H. Zhang, G. Wang, J. Liu, H. Hu, and S. Liu, "Network Traffic Prediction Based on Deep Learning." IEEE Access, vol. 6, pp. 23302-23310, 2018.

[17] Z. Zhao, Y. Zhang, Y. Xu, and H. Liang, "Deep learning based network traffic prediction: Methods, datasets and analysis." IEEE Access, vol. 5, pp. 5143-5153, 2017.

[18] H. Liu, J. Cai, and Y. S. Ong, "Remarks on Multi-output Gaussian Process Regression." Knowledge-Based Systems, vol. 144, pp. 102-121, 2018.

[19] E. V. Bonilla, K. M. Chai, and C. K. Williams, "Multi-task Gaussian process prediction." In Advances in Neural Information Processing Systems, pp. 153-160, 2008.

[20] Han, W., Zhu, T., Chen, L., Ning, H., Luo, Y., & Wan, Y. MCformer: Multivariate Time Series Forecasting with Mixed-Channels Transformer. IEEE Internet of Things Journal, 2024.

[21] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: appropriate use and interpretation." Anesthesia & Analgesia, vol. 126, no. 5, pp. 1763-1768, 2018.

[22] H. Xu, and Y. Deng, "Dependent evidence combination based on shearman coefficient and pearson coefficient," IEEE Access, vol. 6, 11634-11640. 2017.

[23] D. J. MacKay et al., "Introduction to Gaussian processes," NATO ASI series F computer and systems sciences, vol. 168, pp. 133–166, 1998.

[24] A. G. Wilson, "Covariance kernels for fast automatic pattern discovery and extrapolation with Gaussian processes", 2014.

[25] L. Yang, K. Wang, and L. Mihaylova, "Online sparse multi-output Gaussian process regression and learning," IEEE Transactions on Signal and Information Processing over Networks, vol.5, no. 2, pp. 258-272, 2018.

[26] T. Chai, and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)." Geoscientific Model Development Discussions, vol. 7, no. 1, pp. 1525-1534, 2014.

[27] Y. Wang, T. Nakachi, "Network traffic anomaly detection: A revisiting to Gaussian process and sparse representation," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. 107, no. 1, pp. 125-133, 2024.

Automated Subjective Perception of a Driver's Pain Level Based on Their Facial Expression

F. Hadi[1], O. Fukuda[2], W. LYeoh[3], H. Okumura[4], Y. Rodiah[5], Herlina[6], A. Prasetyo[7]
Graduate School of Science and Engineering, Saga University, Saga, Japan[1, 2, 3, 4]
Electrical Engineering Department, University of Bengkulu, Bengkulu, Indonesia[5, 6]
Faculty of Medicine, Diponegoro University, Semarang, Indonesia[7]

***Abstract*—One factor that has a positive correlation with the risk of traffic accidents is the pain experienced by drivers. This pain is sometimes expressed facially by the driver and can be subjectively perceived by others. By observing the facial expression of drivers, it can estimate the pain experienced at that point in time and intervene to prevent some accidents. A method to automatically estimate the pain level expressed by a driver using their facial expression will be proposed in this study. The model is trained by a convolution neural network based on a public dataset of facial expressions at various pain levels. This model is then used to automatically classify the pain level perceived using only the facial expressions of drivers. The result of the automated classification is then compared to ratings of subjective feelings of the driver's pain evaluated by a medical doctor. The experiment results showed that the model classified the pain level expressed facially by the drivers matched that of the classification by the medical doctor at a rate of 80%.**

***Keywords*—Pain; driver; convolution neural network; facial expression**

I. Introduction

Neck and back pain are part of chronic pain that can cause discomfort, be annoying, and interfere with a person's ability to concentrate. If a person is driving and experiencing pain, it can cause collisions [1]. In fact, pain in drivers is one of the main contributing factors to traffic accidents [2]. Furthermore, pain impairs a person's ability for quick thinking when operating a vehicle [3]. Certain medications can also impair cognitive performance and cause pain [4]. Driving over extended periods of time increases the chance of pain, and psychological elements linked to pain, such as stress and anxiety, can also influence driver behaviors and raise the risk of accidents in addition to physical discomfort [5].

The definition of pain, according to the International Association for the Study of Pain (IASP), is an unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage[6]. Pain is generally classified into two main types: acute pain dan chronic pain. Acute pain typically triggered by injury or disease, while chronic pain persists beyond the usual healing period and may not have a clear underlying cause[7]. In addition, the relationship between chronic pain and psychological disorders is complex and often bidirectional. For example, individuals with chronic neck pain are at higher risk for mood and anxiety disorders. This suggests that chronic pain may lead to psychological distress, which in turn may worsen pain perception[8].

Assessment of pain is a crucial component of pain management since it enables the determination of the intensity and consequences of pain on the patient's quality of life. Multiple instruments, including the Visual Analogue Scale (VAS) and the Numeric Rating Scale (NRS), are frequently employed to quantify pain intensity by relying on self-reported information provided by patients [9]. VAS was popularized by Aitken and Zealley in the late 1960s, who focused its function on psychological assessment [10]. They showed that VAS could effectively measure subjective feelings, which laid the foundation for its subsequent application in clinical pain measurement. In its development, VAS continued to be validated through several follow-up studies [11]. Moreover, the VAS has been used into more extensive research projects, including those involving cancer patients [12], emergency rooms [13], and measuring pain levels in fibromyalgia patients correlating pain intensity with functional disability and psychological symptoms [14].

The VAS measurement method is generally using a straight line, usually 10 cm long, with the starting point representing "no pain" and the end point indicating "maximum imaginable pain"[15]. One of the main limitations of VAS is that it is highly dependent on subjective assessment of the patient. Each individual may interpret the scale differently. This can be influenced by various psychological factors, such as anxiety and mood when the assessment is carried out. The results will be inconsistent and therefore do not reflect the actual pain[16]. Another critical limitation is that it may not provide adequate accuracy especially over a narrow scale range resulting in reduced sensitivity to detect small changes in pain levels[17].

Several studies on facial expression-based face detection have been developed. One of the leading methods is the Facial Action Coding System (FACS), which categorizes facial movements into certain Action Units (AU). This method has been used to distinguish between real and fake pain expressions in children undergoing dental treatment[18]. However, the reliability of FACS in assessing pain in various populations is still a concern, especially since this technique requires trained personnel to code facial expressions accurately [19]. The combination of facial electromyography (EMG) with electroencephalography (EEG) has also been used to explore the relationship between brain activity and facial expressions when experiencing pain. Although the results of the study showed a correlation between facial muscle movements and brain electrical activity during pain stimuli, this system is considered more complex and variability in the result[20]. In addition, the

role of observers is often distorted in interpreting facial expressions due to the influence of previous exposure to pain assessment[21]. This suggests that the context in which facial expressions are viewed can significantly affect pain recognition, indicating the need for standardized training for health care providers to reduce bias.

Since the use of Visual Analog Scale for pain assessment that has been widely used in various studies still has limitations, the use of image processing techniques and the application of machine learning in classifying features is expected to provide better results. Standardization by experts can also avoid subjectivity of assessment so that it can be applied in a wider field, including pain detection in drivers to reduce traffic accidents.

This study aims to automatically estimate the pain level expressed through a driver's facial expression. This may be a viable non-invasive approach to detecting whether a driver is experiencing pain. Additionally, a model capable of detecting a wider range of pain levels would be more beneficial than one that merely differentiates between two conditions.

II. METHODOLOGY

This research is divided into two main stages. The first stage is a classification model trained using a public dataset of facial expressions shown at different levels of pain. For the second stage, the model is used to estimate the level of pain expressed facially by the driver. Figure 1 illustrates the stages of this research process.

A. You Only Look Once (YOLO) for Pain Classification

The You Only Look Once (YOLO) is a popular detection algorithm developed by Redmon J, in 2012[22]. The YOLO algorithm process is to divide the facial image into S x S meshes. Each grid is responsible for predicting the target where the actual box will fall in the center of the grid. The total bounding box is generated from the meshes. Each bounding box contains five parameters: Target center point coordinates, target width and height dimensions (x, y, w, h), and confidence. The S x S edge predicts the category probability of the target on that edge. The prediction bounding box confidence and category probability are then multiplied to obtain the category score for each prediction box[23].

In this study, we use YOLOv5 which offers important benefits in speed and computational efficiency[24], [25] lightweight architecture[26], and capability to effectively handle dynamic and diverse facial expressions [27]. Therefore, it is very appropriate for application in real-time scenarios as it enables rapid detection of driver pain, thereby preventing traffic accidents.

YOLOv5 uses a single-stage detection methodology, where the image is processed in a single iteration across the entire network with a convolutional neural network (CNN) approach that simultaneously predicts bounding boxes and class probabilities [28]. It consists of backbone, neck, and head segment components. The main function of the backbone is to extract features, which are then aggregated by the neck to provide predictions at multiple scales. Meanwhile, the head produces the final detection findings [29].

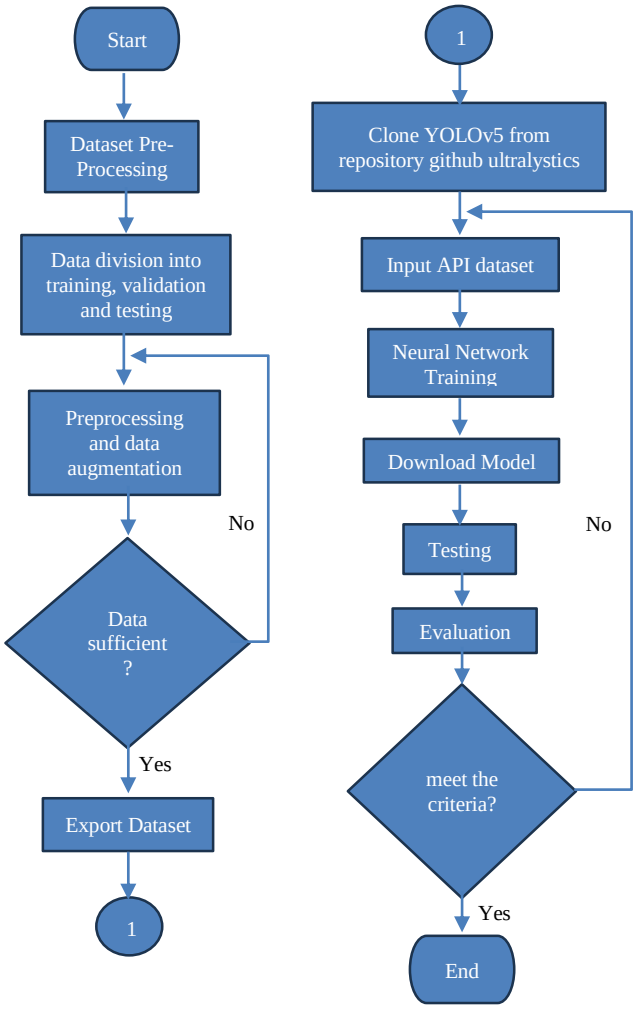


Fig. 1. Research workflow.

A significant improvement in YOLOv5 is the incorporation of anchor boxes, which are pre-defined bounding box shapes that improve the model's ability to reliably estimate object positions. This methodology uses a combination of methods, including data augmentation, to improve the robustness of the model by artificially increasing the size of the training dataset [30].

B. Datasets Pre-Processing

The dataset used in this study is from the Multimodal Intensity Pain (MIntPAIN) database[1] with 3079 images divided into 11 classes (0 - 10). Each class will be divided into three classifications based on its class range, namely mild, moderate, and severe. Fig. 2 shows an example dataset [31] with its VAS values. The image is intentionally blurred due to agreements with the dataset provider and privacy reason. The data pre-processing stage is conducted by annotating the area or pixels in the image as a region or area in the bounding box that will be used for model calculations as training or validation. In the YOLO method, annotation is carried out starting by drawing a

[1] (http://www.vap.aau.dk/mintpaon-database

bounding box on each object in the image which then stores the description of the bounding box or object class in a text file database containing the class, x coordinates, y coordinates, width, and height respectively. Data in each VAS value class is divided into 70% for training, 20% for validation, and 10% for testing. The next stage is to carry out the auto orient and resize process. Auto orient aims to adjust the position of the object in the image to ensure that the main object in the image is in the right position. While the resize, the process is to change the image size to the same, 640 x 640 pixels. The purpose of this stage is to equalize the image size because the data obtained can have variations in orientation, both in portrait and landscape formats. By resizing the images, it can ensure that all images have uniform dimensions for data processing and its use in model training.

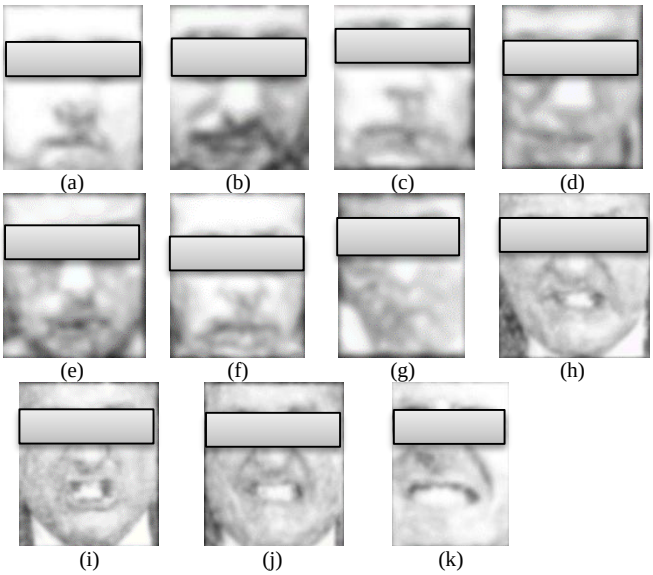


Fig. 2. VAS value dataset sample (a) 0, (b) 1, (c) 2, (d) 3, (e) 4, (f) 5, (g) 6, (h) 7, (i) 8, (j) 9 (k)10.

C. Model Training

The model-building stage is carried out on Google Colab. The first step in this process is training the YOLOv5 model available in the ultralytics github repository. The previously prepared dataset is cloned and entered using the roboflow API with the Model training configuration shown in Table I.

TABLE I. MODEL TRAINNING CONFIGURATION

Image Size	640x640 pixels
Parameters	Batch Size = 16
	Epoch = 300
Hyperparameter	default
IoU	0.5 and 0.75

The 0.5 threshold is often used as a baseline to evaluate model performance, while the 0.75 threshold is utilized for more stringent evaluation [32]. The model is not only accurate but also robust across a wide range of datasets and conditions [33]and models evaluated at the 0.5 IoU threshold performed significantly differently compared to models evaluated at lower thresholds[34].

D. Model Evaluation

The evaluations conducted in this study include precision, recall, mean Average Precision (mAP), and Intersection over Union (IoU). This process is carried out to assess the effectiveness of the object identification algorithm.

Precision measures the ratio of true positives (TP) correctly identified from all positive predictions, while recall evaluates the ratio of positive examples correctly identified from all object examples. These metrics are very important for assessing the precision of the model in object identification while avoiding excessive false positives[35]. F1-score is a metric that considers the trade-off between precision and recall, offering a single number that represents the overall performance of the model[36]. The mathematical representation for recall, precision, and F1-score are given in Eq.(1), Eq.(2), and Eq.(3).

$$Recall = \frac{TP}{TP+FN} = \frac{TP}{All\ Ground\ Truth} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} = \frac{TP}{All\ Prediction} \quad (2)$$

$$F1 - Score = \frac{2 \times (Recall+Precision}{(Recall+Precision)} \quad (3)$$

Mean Average Precision (mAP) is an important quantification in the evaluation of YOLOv5. This value is calculated by taking the average of the precision scores at various levels of recall. This value is often calculated at a certain IoU threshold, such as mAP@0.5, to assess the model's capacity to predict object locations accurately. Studies have shown that YOLOv5 can achieve high mean precision (mAP) scores, thus validating its efficacy in various detection tasks. [37], [38]. The formula for mAP is given in Eq.(4).

$$\text{mAP} = \frac{1}{n}\sum_{i=1}^{n} APi \quad (4)$$

where,

n: Number of data AP

AP: Average Precision

If the evaluated model does not meet the desired value, which is above 0.8, then the model is re-created by adding a new image dataset to improve the parameter values to match the desired so that the model is considered suitable for use.

To evaluate the overlap between predicted bounding boxes and ground truth boxes, the intersection over Union (IoU) metric is a key measure. A higher intersection over Union (IoU) coefficient indicates superior accuracy in localization. If the IoU value is greater than the threshold value of 0.5 (the value assumed to increase the accuracy of detected objects), then the results are acceptable[39] [40]. IoU can be calculated using the formula, as shown in Eq.(5).

$$IoU = \frac{Intersection\ Area}{Union\ Area} \quad (5)$$

E. Automatically Estimating Pain Level of Drivers from Their Facial Expression

We involved 15 males online taxi drivers who had been driving for 5 to 6 hours as research subject. During driving, many VAS values will be read according to the participant's facial expression because the driver has previously driven for 5-

6 hours. Therefore, for the limit, the VAS value chosen is the highest VAS value with an observation period of no more than 2 hours.

For the research environment setting, the camera is mounted in front of the driver, which is focused on the driver's head and is arranged in such a way that it does not obstruct the driver's view as shown in Fig. 3. The data taken from the camera is video which is then converted into image and then processed to detect VAS value.

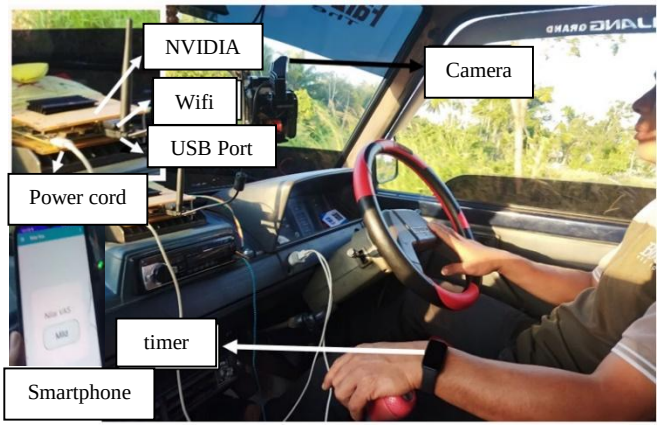


Fig. 3. A view of the experimental scenario.

III. Result and Discussion

A. YOLOv5 Model Training Result

In YOLO model training with 300 epochs and a batch size of 16, the results are shown in the confusion matrix as shown in Fig. 4. In addition, the recall, precision, F1-score, and mAP values are also calculated to measure the performance of the model. The all-normalization technique is applied in this study by dividing each element in the confusion matrix by the total number of instances in the data set. This procedure transforms each cell into a representation of the overall proportion of the data set rather than simply showing the absolute number. This is relevant to this study, where the classes correspond to different levels of pain (ranging from 0 to 10), all normalization can give an overall understanding of how well the model performs across all levels.

Based on the confusion matrix results, the prediction scores obtained can provide an understanding of the extent to which the YOLO model is able to classify each class with accuracy. The value class 8 gets the highest prediction score of 0.96. This shows that the model is very good at identifying and predicting objects included in these classes. The high prediction score in class 8 can be caused by several factors, such as clear and consistent representation in the training data, sufficient variation in the objects representing this class, and optimal hyperparameter selection. Meanwhile, the three classes of VAS values with values 2, 3, and 4 have almost the same and relatively low confusion matrix values of 0.40; 0.63; and 0.42 respectively. This is influenced by several factors, including the expression for these values has almost the same expression because pain with VAS values 0 to 4 has not changed much of the expression. Therefore, if the Interval between the classes is relatively small, difficulty in distinguishing between the classes can occur. In addition, the subjectivity of the interpretation of VAS values by individuals can cause similar perceptions of different levels and the limitations of the samples used in data collection also affect this. As for the result, the value of the train data is as shown in Fig. 5.

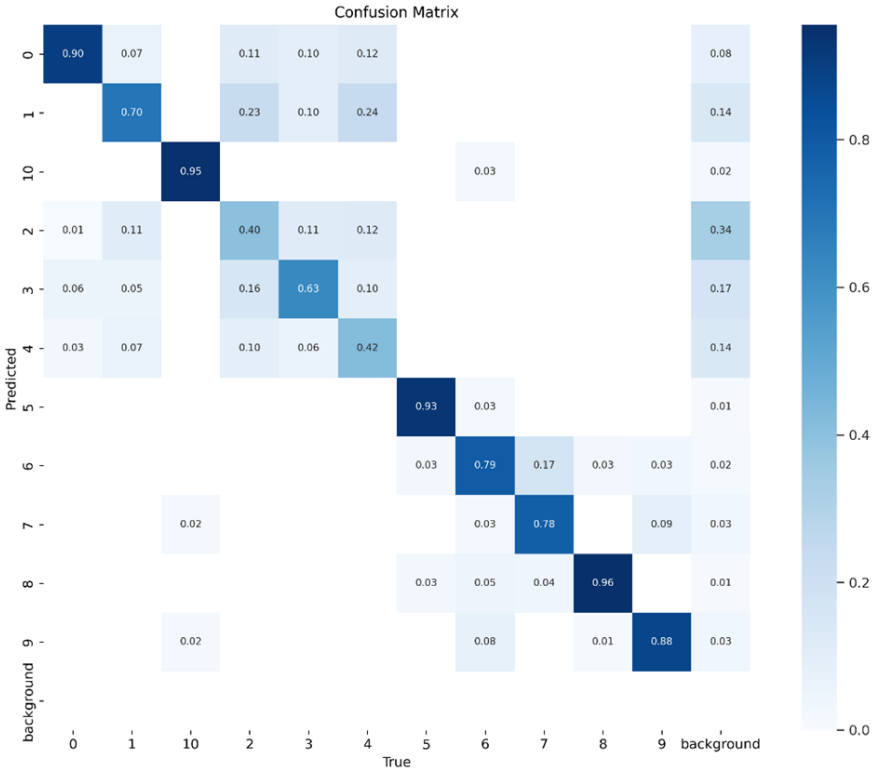


Fig. 4. Confusion matrix result.

From these results, numerous figures illustrate variations in the performance of the YOLO model throughout the training process. The graph illustrates that box loss, object loss, and class loss exhibit a general level of consistency, with minor variations over the training phase. This observation suggests that the model exhibits consistency in its learning. Nevertheless, the recall graph exhibits instability resulting from imbalances in class distribution, less than ideal model parameters and configurations, and inadequate quality of the dataset. To enhance recall stability, it is imperative to examine the distribution of samples in each class, fine-tune model parameters, and enhance dataset quality by collecting more representative data. Furthermore, to enhance the legibility of the model training outcomes, they are shown in Table II. The training results obtained an average value of mAP at 0.5 of 0.82673 and mAP at 0.95 of 0.67048 which were produced with epoch 300. The training process with default hyperparameters was stable at epoch 300 so that the model was said to have been fulfilled and could be used.

For model evaluation using the IoU value, The IoU value is varied to 0.5 and 0.75 which aims to determine the effect of the given IoU. The evaluation results on the model are shown in Fig. 6. The f1-score graph results have almost the same shape, this means that the variation of the IoU value does not have a significant effect on the f1-score value because the precision and recall values do not change much when the IoU value is varied in this model. From the graph, it can be concluded that the recommended confidence value to use is in the range of 0.2 to 0.5.

TABLE II. MODEL TRAINING RESULT

Type of Data	***Box Loss***	***Object Loss***	***Class Loss***
Training	0.013476	0.0049682	0.0095172
Validation	0.012259	0.003167	0.014258
Metrics	Precision	Recall	F1-Score
	0.73679	0.82046	0.789075908
	mAP		
	@0.5	@0.5:0.95	
	0.82673	0.67048	

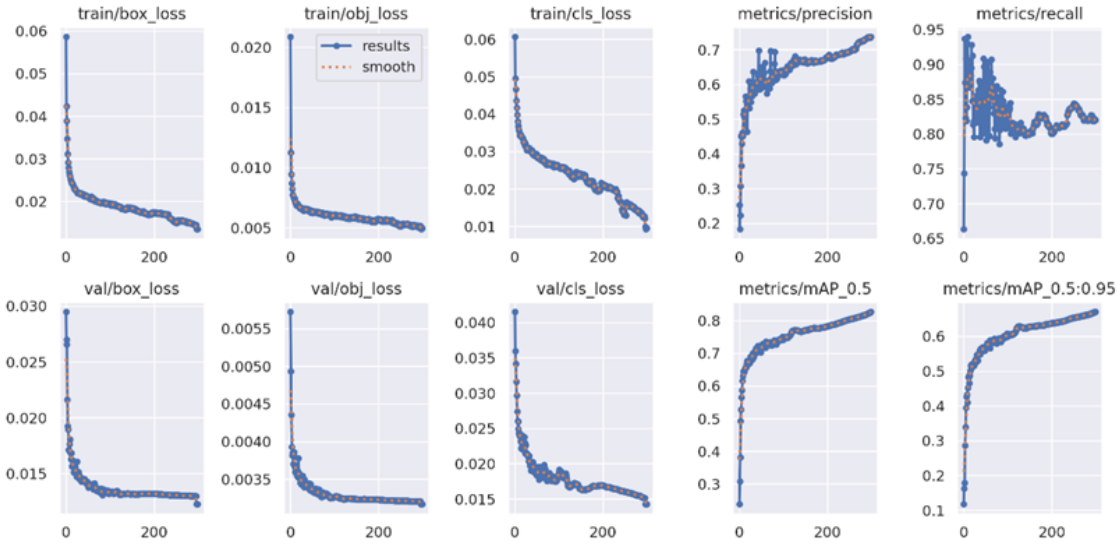
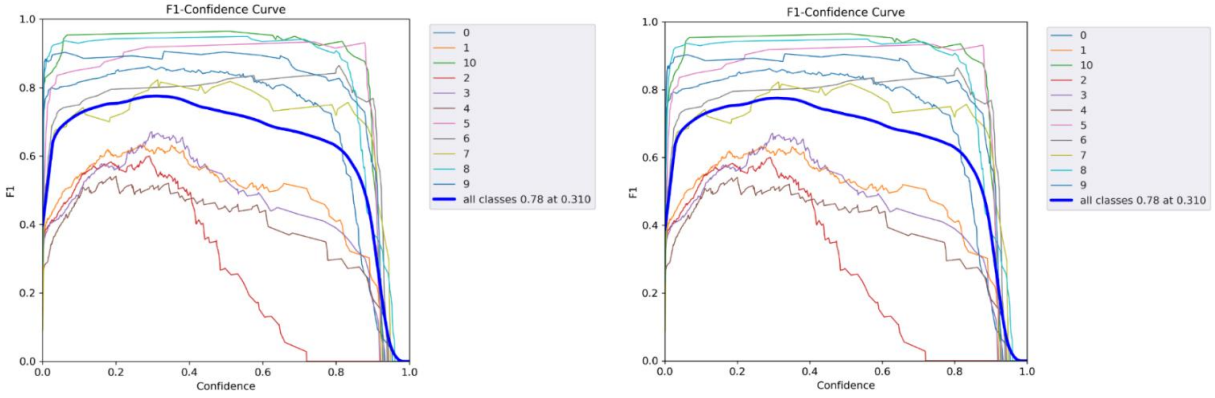


Fig. 5. Model training result.



(a) (b)

Fig. 6. F1-score (a) IoU 0.5, (b)).75.

B. Evaluation with Real World Drivers

The estimation results from the model are compared with ratings from a medical doctor. The classification results are divided into three levels, namely "mild" with a VAS value range of 0-3, "moderate" with a VAS value of 4-7, and "severe" with a VAS value of 8-10. The results from the 15 drivers recorded are shown in Table III.

TABLE III. PAIN LEVEL ESTIMATED FROM FACIAL EXPRESSION BY THE MODEL AND A MEDICAL DOCTOR

Subject	VAS Value		Classification	
	Subjective rating by medical doctor	***Proposed Model***	***Subjective rating by medical doctor***	***Proposed Model***
1	7	6	severe	moderate
2	5	6	moderate	moderate
3	6	6	moderate	moderate
4	5	5	moderate	moderate
5	8	8	severe	severe
6	5	6	moderate	moderate
7	1	1	mild	mild
8	5	5	moderate	moderate
9	1	1	mild	mild
10	9	9	severe	severe
11	9	9	severe	severe
12	3	4	mild	moderate
13	8	6	severe	moderate
14	1	1	mild	mild
15	1	1	mild	mild

Of the 15 participants, there were 3 VAS values that deviated from the subjective assessment by the medical doctor so that accuracy can be calculated by subtracting the observed value from the actual value and dividing it by the actual value, then multiplying by 100%, so that:

$$Accuracy = \frac{15-3}{15} = \frac{12}{15} \times 100\% = 80\%$$

With the test results with an error rate of 20%, the following discusses factors that can affect measurement accuracy.

This study was conducted in Indonesia involving male volunteer drivers in Bengkulu City who are predominantly Malay Austronesian ethnicity with an age range of 20-30 years. The results of the study cannot be generalized to be applied to everyone. Older people will have different detection rates. Although the existing datasets are all male, there are still limitations due to demographic, social and cultural diversity. This of course can affect the generalization of the model to different populations. In addition, the availability of datasets from the country of origin is not yet available, which will affect the reliability of the model. Therefore, more accuracy needs to be improved by adding datasets using different ethnic facial variations. Although the results of this study work well on the existing test dataset, if applied to a general context, it still needs to be validated.

Pain is a feeling that can be subjectively felt by someone. It could be that the variation shown in pain at a certain level but is different in giving expression. Variations in facial expressions are a challenge in themselves that can affect the accuracy of this study. Although it has involved external assessments involving medical doctors, it may still be influenced by bias in the dataset. This can also affect the results, especially if applied to different populations. Nevertheless, the proposed model has been proven effective in detecting pain using the Visual Analogue Scale value approach.

Model performance when applied in the real world is influenced by lighting, camera angle, driver movement and if the driver speaks or shows other expressions such as emotions and others. Although research has applied variations in datasets by adding variations in the form of blur, rotation, and brightness enhancement so that the model training process can learn from various data conditions, so that the model can learn from various data conditions, lighting conditions greatly affect the results obtained. Proper camera placement is also a major concern so that data consistency can be maintained. Another thing is if the driver speaks or gives an emotional expression that is very likely to affect the measurement results. These results also may not be generalizable to all conditions due to differences in demographic and cultural factors as discussed above. In addition, variations in pain types (acute and chronic types) should also be considered so that these results may be difficult to apply to different types of pain.

Future research can be further developed by collecting more diverse data sets from different ethnicities which is expected to improve system learning. To improve system validation, this research can be developed by involving biomedical sensors such as heart rate or tone of voice, so that the results are more accurate and can be validated independently. The selectivity and sensitivity of the research can still be improved by applying certain methods so that they can distinguish between real pain and fake pain, without the help of experts in assessing pain. The use of cameras that can work optimally and are not too affected by the environment such as vehicle movement and changes in light intensity can also be considered.

IV. CONCLUSION

In this study we propose to estimate a pain level expressed by facial expressions of drivers by applying VAS value. Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads. The proposed model can process and classify the driver's facial expression based on the Visual Analog Scale (VAS) value scale ranging from 0 to 10. The classification results are divided into three levels, namely "mild" with a VAS value range of 0-3, "moderate" with a VAS value of 4-7, and "severe" with a VAS value of 8-10.

The results of the confusion matrix show that the YOLO model can classify each class accurately. Class value 8 obtained the highest prediction score of 0.96, but the other 2 classes, namely classes 2 and 4, showed accuracy values below 0.5. This can be caused by the relatively small interval between classes or the subjectivity of the interpretation of VAS values so that the same perception is at different levels.

The experiment test results also show that the system can classify the driver's facial expression of pain level with that of classification by the medical doctor at rate of 80%. Thus, the model may be used to detect driver pain and could lead to a reduction the number of traffic accidents in the future.

ETHICS STATEMENT

Our research is based on non-invasive measurement (camera recording from the car dashboard), and we did not intervene in the everyday routine of the experimental participants. The 15 experimental participants were taxi drivers, and their facial expressions during their routine work were recorded and analyzed. All experimental participants gave informed consent before the experiment and written consent was obtained. The pain rating obtained was based solely on the driver's facial expression recorded from a camera, either using a neural-network model or by a medical doctor. We did not obtain the actual subjective pain experience by the participants. Their facial expression may not reflect the actual amount of pain they were experiencing. There is almost no or small possibility of risk or danger arising from the implementation of this research.

In accordance with the local legislation and institutional requirement[2] where this experiment was performed because the study was investigating public behavior and was purely observational (non-invasive and non-interactive), ethical approval was not required.

REFERENCES

[1] A. Aafreen, A. R. Khan, A. Khan, N. K. Maurya, and A. Ahmad, "Prevalence of Neck Pain in Car and Motorcycle Drivers: A Comprehensive Review of Primary, Secondary, and Tertiary Care," JOURNAL OF CLINICAL AND DIAGNOSTIC RESEARCH, 2023, doi: 10.7860/jcdr/2023/64993.18222.

[2] A. Vaezipour, M. S. Horswill, N. E. Andrews, V. Johnston, P. Delhomme, and O. Oviedo-Trespalacios, "How distracting is chronic pain? The impact of chronic pain on driving behaviour and hazard perception," Accid Anal Prev, vol. 178, 2022, doi: 10.1016/j.aap.2022.106856.

[3] C. Longtin, A. Lacasse, C. E. Cook, M. Tousignant, and Y. Tousignant-Laflamme, "Management of low back pain by primary care physiotherapists using the pain and disability drivers management model: An improver analysis," Musculoskeletal Care, vol. 21, no. 3, 2023, doi: 10.1002/msc.1742.

[4] E. Matsuoka, M. Saji, and K. Kanemoto, "Daytime sleepiness in epilepsy patients with special attention to traffic accidents," Seizure, vol. 69, 2019, doi: 10.1016/j.seizure.2019.04.006.

[5] C. van Vreden et al., "The physical and mental health of Australian truck drivers: a national cross-sectional study," BMC Public Health, vol. 22, no. 1, 2022, doi: 10.1186/s12889-022-12850-5.

[6] S. N. Raja et al., "The revised International Association for the Study of Pain definition of pain: concepts, challenges, and compromises," Sep. 01, 2020, Lippincott Williams and Wilkins. doi: 10.1097/j.pain.0000000000001939.

[7] M. K. Nicholas, "Time vs mechanism in chronic pain," 2022. doi: 10.1097/j.pain.0000000000002584.

[8] S. Kazeminasab et al., "Neck pain: global epidemiology, trends and risk factors," Dec. 01, 2022, BioMed Central Ltd. doi: 10.1186/s12891-021-04957-4.

[9] E. E. Krebs et al., "Development and Initial Validation of the PEG, a Three-item Scale Assessing Pain Intensity and Interference," J Gen Intern Med, vol. 24, no. 6, 2009, doi: 10.1007/s11606-009-0981-1.

[10] H. M. McCormack, D. J. de L. Horne, and S. Sheather, "Clinical applications of visual analogue scales: A critical review," Psychol Med, vol. 18, no. 4, 1988, doi: 10.1017/S0033291700009934.

[11] C. Maxwell, "Sensitivity and accuracy of the visual analogue scale: a psycho - physical classroom experiment.," Br J Clin Pharmacol, vol. 6, no. 1, 1978, doi: 10.1111/j.1365-2125.1978.tb01676.x.

[12] M. M. Alfonsin, R. Chapon, C. A. B. de Souza, V. K. Genro, M. M. C. Mattia, and J. S. Cunha-Filho, "Correlations among algometry, the visual analogue scale, and the numeric rating scale to assess chronic pelvic pain in women," Eur J Obstet Gynecol Reprod Biol X, vol. 3, 2019, doi: 10.1016/j.eurox.2019.100037.

[13] M. Jiang et al., "Elastic Silicone Occlusive Sheeting Versus Silicone Occlusive Sheeting in the Treatment of Scars: A Randomized Controlled Trial," Dermatol Ther (Heidelb), vol. 12, no. 8, 2022, doi: 10.1007/s13555-022-00763-5.

[14] L. G. R. Fernandes et al., "Correlation between levels of perceived stress and depressive symptoms in the functional disability of patients with fibromyalgia," Rev Assoc Med Bras, vol. 69, no. 11, 2023, doi: 10.1590/1806-9282.20230690.

[15] J. Aceituno-Gómez et al., "Correlation between three assessment pain tools in subacromial pain syndrome," Clin Rehabil, vol. 35, no. 1, pp. 114–118, Jan. 2021, doi: 10.1177/0269215520947596.

[16] Y. S. Park, J. Choi, and S. W. Park, "Blink reflex changes and sensory perception in infraorbital nerve-innervated areas following zygomaticomaxillary complex fractures," Arch Plast Surg, vol. 47, no. 6, pp. 559–566, 2020, doi: 10.5999/aps.2020.01130.

[17] J. Khadka, P. G. Schoneveld, and K. Pesudovs, "Comparing the measurement properties of visual analogue and verbal rating scales," Ophthalmic and Physiological Optics, vol. 42, no. 1, 2022, doi: 10.1111/opo.12917.

[18] M. Alkhouli, Z. Al-Nerabieah, and M. Dashash, "Analyzing the Facial Action Units associated with genuine and fake pain caused by inferior alveolar nerve block in Syrian children: a cross-sectional study," Jun. 21, 2023. doi: 10.21203/rs.3.rs-3044856/v1.

[19] Z. Chen, R. Ansari, and D. J. Wilkie, "Learning Pain from Action Unit Combinations: A Weakly Supervised Approach via Multiple Instance Learning," IEEE Trans Affect Comput, vol. 13, no. 1, pp. 135–146, 2022, doi: 10.1109/TAFFC.2019.2949314.

[20] C. Ma, C. Wang, D. Zhu, M. Chen, M. Zhang, and J. He, "The Investigation of the Relationship Between Individual Pain Perception, Brain Electrical Activity, and Facial Expression Based on Combined EEG and Facial EMG Analysis," J Pain Res, vol. 18, pp. 21–32, 2025, doi: 10.2147/JPR.S477658.

[21] P. J. Göller, P. Reicherts, S. Lautenbacher, and M. Kunz, "Vicarious facilitation of facial responses to pain," European Journal of Pain (United Kingdom), vol. 28, no. 1, pp. 133–143, Jan. 2024, doi: 10.1002/ejp.2169.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016. doi: 10.1109/CVPR.2016.91.

[23] R. Lakhotiya, M. Chavan, S. Divate, and S. Pande, "Image Detection and Real Time Object Detection," Int J Res Appl Sci Eng Technol, vol. 11, no. 5, 2023, doi: 10.22214/ijraset.2023.51839.

[24] G. Yang et al., "Face Mask Recognition System with YOLOV5 Based on Image Recognition," in 2020 IEEE 6th International Conference on Computer and Communications, ICCC 2020, 2020. doi: 10.1109/ICCC51575.2020.9345042.

[25] W. Yang, X. Gan, and J. He, "Defect Identification of 316L Stainless Steel in Selective Laser Melting Process Based on Deep Learning," Processes, vol. 12, no. 6, Jun. 2024, doi: 10.3390/pr12061054.

[26] O. M. Lawal, "YOLOv5-LiNet: A lightweight network for fruits instance segmentation," PLoS One, vol. 18, no. 3 March, 2023, doi: 10.1371/journal.pone.0282297.

[2] Guidelines and Ethical Standards National Health Research and Development, Ministry of Health of the Republic of Indonesia, 2021

[27] Y. Xu, "Real-Time Face Expression Recognition Monitoring Using Deep Learning," in Advances in Transdisciplinary Engineering, IOS Press BV, Mar. 2024, pp. 698–704. doi: 10.3233/ATDE240135.

[28] Y. Ma, "Target tracking and detection based on YOLOv5 algorithm," Applied and Computational Engineering, vol. 16, no. 1, 2023, doi: 10.54254/2755-2721/16/20230860.

[29] Y. Chang, D. Zhou, Y. Tang, S. Ou, and S. Wang, "An improved deep learning network for image detection and its application in Dendrobii caulis decoction piece," Sci Rep, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-63398-w.

[30] A. Alotaibi and T. Arif, "Fast and accurate automated intestinal parasites egg detection and classication from images based on YOLOv5 deep convolutional neural network," 2023, doi: 10.21203/rs.3.rs-2520494/v1.

[31] M. A. Haque et al., "Deep multimodal pain recognition: A database and comparison of spatio-temporal visual modalities," in Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018, Institute of Electrical and Electronics Engineers Inc., Jun. 2018, pp. 250–257. doi: 10.1109/FG.2018.00044.

[32] M. R. Subhi, E. Rachmawati, and G. Kosala, "Safety Helmet Detection on Field Project Worker Using Detection Transformer," Journal of Information System Research (JOSH), vol. 4, no. 4, pp. 1316–1323, Jul. 2023, doi: 10.47065/josh.v4i4.3852.

[33] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Jun. 2019, pp. 658–666. doi: 10.1109/CVPR.2019.00075.

[34] L. Lam, M. George, S. Gardoll, S. Safieddine, S. Whitburn, and C. Clerbaux, "Tropical Cyclone Detection from the Thermal Infrared Sensor IASI Data Using the Deep Learning Model YOLOv3," Atmosphere (Basel), vol. 14, no. 2, Feb. 2023, doi: 10.3390/atmos14020215.

[35] S. M. Makhdoomi, C. Khosla, and S. D. Pande, "Real Time Lung Cancer Classification with YOLOv5," EAI Endorsed Trans Pervasive Health Technol, vol. 9, no. 1, 2023, doi: 10.4108/eetpht.9.3925.

[36] W. Hao, C. Ren, M. Han, L. Zhang, F. Li, and Z. Liu, "Cattle Body Detection Based on YOLOv5-EMA for Precision Livestock Farming," Animals, vol. 13, no. 22, 2023, doi: 10.3390/ani13223535.

[37] M. M. Rana, M. S. Hossain, M. M. Hossain, and M. D. Haque, "Improved vehicle detection: unveiling the potential of modified YOLOv5," Discover Applied Sciences, vol. 6, no. 7, Jul. 2024, doi: 10.1007/s42452-024-06029-3.

[38] E. Casas, L. Ramos, E. Bendek, and F. Rivas-Echeverria, "Assessing the Effectiveness of YOLO Architectures for Smoke and Wildfire Detection," IEEE Access, vol. 11, 2023, doi: 10.1109/ACCESS.2023.3312217.

[39] M. A. Azam et al., "Deep Learning Applied to White Light and Narrow Band Imaging Videolaryngoscopy: Toward Real-Time Laryngeal Cancer Detection," Laryngoscope, vol. 132, no. 9, 2022, doi: 10.1002/lary.29960.

[40] F. M. A. Mazen, R. A. A. Seoud, and Y. O. Shaker, "Deep Learning for Automatic Defect Detection in PV Modules Using Electroluminescence Images," IEEE Access, vol. 11, 2023, doi: 10.1109/ACCESS.2023.3284043.

Mobile Application Based on Geolocation for the Recruitment of General Services in Trujillo, La Libertad

Melissa Giannina Alvarado Baudat, Camila Vertiz Asmat, Fernando Sierra-Liñan
Facultad de Ingeniería, Universidad Privada del Norte, Trujillo, Perú

***Abstract*—Currently, there is no technological solution that efficiently facilitates the offering of general services by independent workers in the city of Trujillo. This limitation reduces job opportunities, as workers secure fewer contracts due to reliance on client recommendations, a method that is often inefficient due to long response times and low accessibility. Leveraging the versatility of mobile applications. This study contributes to computer science by demonstrating how cloud-based data management, real-time communication, and location-based service matching using Google APIs optimize service efficiency and user experience. The study follows an applied research approach with a quantitative methodology, employing a pre-experimental explanatory design and a sample of 22 workers selected through non-probabilistic convenience sampling. The development was carried out using the Flutter framework and the Dart programming language, with an SQL database hosted on Microsoft Azure cloud services. The Mobile-D agile methodology guided the development process. After implementing the application, the results showed an 86.79% reduction in the average hiring process time, a 50% increase in the number of contracts completed, and a 51.27% improvement in workers' average satisfaction. These findings highlight the effectiveness of mobile and cloud computing technologies, along with ranking algorithms and geolocation services, in streamlining labor market interactions and improving user experience.**

Keywords*—Mobile application; recruitment; geolocation; general services

I. Introduction

We live in an era defined by technological advances, where virtually any activity, from education and business to finance and entertainment, can be completed through mobile devices [1]. The widespread penetration of these devices is evident in statistics showing that 45% of people in 40 countries use non-smartphones, while 43% opt for smartphones [2]. This context of mobile connectivity has led to the emergence of apps as the main communication channel in modern society, standing out for their versatility and efficiency in accessing important and reliable information at any time of the day [3].

The influence of the digital age extends even to the workplace, where the conventional job search has been replaced by digital platforms, which have proven successful in connecting professional candidates with potential employers [4]. Globally, unemployment rates are higher in wealthy countries than in poorer ones, especially among less-educated workers who face greater difficulties adapting to a labor market driven by technological progress that primarily benefits the highly skilled [5]. The deindustrialization of the manufacturing sector has led to a significant decline in "blue-collar" jobs, reflecting a polarization of employment that particularly affects intermediate-skilled occupations. While unemployment among more educated workers remains stable, less educated workers experience rising unemployment rates as GDP per capita increases. This indicates that skill-biased economic development contributes to higher unemployment rates among less-skilled workers, who often must abandon self-employment to seek salaried jobs in an increasingly competitive and precarious labor environment [6].

However, many workers offering services such as electricians, plumbers, builders, tailors, carpenters, among others, lack access to a technological tool with similar characteristics to those available in other sectors, which would facilitate finding job opportunities [7]. This situation forces them to rely heavily on word-of-mouth recommendations from previous clients to secure jobs [8]. On the other hand, for clients, finding trustworthy personnel without the help of the internet can be an arduous and tedious task, as the waiting time to contact a worker is often lengthy [9].

According to study [10], digitalization is transforming the labor market by facilitating job searches and recruitment via the internet, fostering labor mobility, increasing efficiency, and reducing structural unemployment, thereby improving the flow of workers between employers. Social media enables workers to promote themselves to employers outside their local markets, and a significant portion of European freelancers find employment through these platforms. Digital technologies are also driving new forms of employment, such as temporary work distributed through task-based service platforms. These changes have positioned private digital intermediaries as central players in labor market operations, creating new considerations for policymakers.

In the context of the COVID-19 pandemic, the International Labor Organization (ILO) assessed its impact on the labor market and the rapidly evolving situation. It analyzed employment and unemployment rates, as well as global working hours in the European Union. Results showed a 5.4% decline in the first quarter of 2020 and a 14% decline in the second quarter compared to the fourth quarter of 2019. Notably, these declines were less pronounced in Europe and Central Asia, with decreases of 3.4% and 13.9%, respectively [11].

In Trujillo, La Libertad, the absence of a digital platform to connect independent workers with clients limits job

opportunities and service quality. To address this challenge, this research aims to develop a mobile application to facilitate the hiring of general services, optimizing the process and benefiting both providers and consumers. Thus, the following question was formulated: How can the implementation of a mobile application improve the hiring of general services in Trujillo, La Libertad, in 2024? The specific questions were as follows: How does the implementation of a mobile application based on geolocation reduce the hiring process times for general services in Trujillo, La Libertad, in 2024? How does the implementation of a mobile application based on geolocation improve the number of contracts in the hiring process for general services in Trujillo, La Libertad, in 2024? How does the implementation of a mobile application based on geolocation improve satisfaction in the hiring process for general services in Trujillo, La Libertad, in 2024?

II. State of the Art

The authors in study [9] addressed the stagnation problem of manual labor startups in their early stages by proposing a geolocated mobile job portal. This portal acts as an intermediary, connecting workers with users based on geographical proximity, thereby improving job prospects and the number of hires for these workers. The methodology employed was an Incremental model, allowing for continuous evaluation and improvement of the software after each development cycle. The application consists of three main modules: login, service requests by users, and request acceptance by providers. This modular approach enhances usability and maintenance while enabling advanced features such as suggesting additional services based on the nearest available personnel. The research concluded successfully, potentially contributing to reducing unemployment in Nigeria by improving connections between employers and workers.

Similarly, aiming to improve the acquisition of local repair services in Metro Manila, [12] developed "Handy Fix," a mobile application designed to optimize the experience of contracted technicians. Using a mixed-methods approach and a sample of 414 households, the challenges faced by technicians in terms of localization, communication, and payment were assessed. The results showed a significant improvement in technician satisfaction, with 89.17% of respondents affirming that the application addressed the key challenge of verifying skills and credentials. Additionally, 85% highlighted that the lack of adequate tools and safety equipment was effectively addressed through the application's integrated ratings and review system.

In another study, [13] developed an Android application that utilizes location-based services to improve the search and management of services such as plumbing, pipefitting, and electrical work. The goal was to increase the efficiency of service management for users. Using an Agile prototyping approach and tools like Android Studio and MySQL, rapid iterations were conducted based on feedback. The methodology included planning, analysis, design, and implementation phases, resulting in a system that surpassed most functional tests. Workers found the application significantly improved profile management, accounts, services, and appointments, though areas for improvement were identified, such as text appearance and ratings functionality.

Furthermore, the authors in study [14], researched and developed an Android application for domestic services, leveraging on-demand application technology to optimize the hiring of home professionals. Through a methodology combining qualitative and quantitative techniques, such as interviews, literature reviews, and case analyses of existing applications, user needs and market trends were identified. The result was an application with three modules: administrator, workers, and clients. The administrator can modify the website after logging in, while clients can describe required services, manage payments, rate services, and initiate a refund process in cases of dissatisfaction. The application significantly improved the service hiring experience, with recommendations provided for its successful market implementation.

The study by [15] examined the evolution and impact of on-demand home service platforms. Using a qualitative and quantitative exploratory-descriptive design, market data, growth trends, and user satisfaction surveys were analyzed. The results highlight the importance of mobile applications in efficiently connecting workers and clients, improving work organization and management, and optimizing service quality and customer satisfaction. However, challenges related to security and trust in mobile technology persist.

The authors in study [16] analyzed the on-demand home service industry and proposed improvements to applications to offer a smoother and more user-friendly experience. Using a descriptive methodology based on literature reviews, market analyses, and user experience evaluations, the study emphasizes the need to enhance service quality, labor rights, and sustainability. The results suggest that applications like Urban Company and TaskRabbit could benefit workers by implementing training programs, quality control measures, and labor policies that ensure fair pay and job security. Moreover, integrating technologies such as artificial intelligence and machine learning is proposed to optimize service allocation and personalize recommendations, enhancing efficiency and job satisfaction while reducing instability and inequality in the workplace.

According to the study by [17], a technological solution was presented to optimize the work of technicians in the home appliance repair sector. The goal is to facilitate connections between technicians and users requiring home repairs through an online platform and mobile applications. Using a quantitative approach and technological development design, the methodology employed tools like Python, Django, Geopy, and the Fast2SMS API to create an accessible and efficient interface. This platform allows technicians to receive service requests in a more organized manner, improving their access to the job market and optimizing their time by reducing unnecessary travel. The results show a significant improvement in service efficiency and satisfaction for both users and providers, though no specific statistical data is provided.

Finally, the study in [18] developed an application for domestic services using the Extreme Programming (XP) methodology, integrating APIsPERU to verify user information through SUNAT and RENIEC data, enhancing security in the hiring process. The application enables real-time worker localization and optimizes logistics through its integration with

Google Maps, achieving an average response time of two seconds. Validation tests confirmed its effectiveness and integration capability with external services.

III. Objectives

A. General Objective

To determine how the implementation of a geolocation-based mobile application improves the hiring process for general services in Trujillo, La Libertad.

B. Specific Objective

- To determine how the implementation of a geolocation-based mobile application reduces the time required for the hiring process of general services in Trujillo, La Libertad.
- To determine how the implementation of a geolocation-based mobile application increases the number of hirings in the general services contracting process in Trujillo, La Libertad.
- To determine how the implementation of a geolocation-based mobile application improves satisfaction in the hiring process for general services in Trujillo, La Libertad.

IV. Materials and Methods

The methodology used for the development of the application is Mobile-D, which, according to study [19], emerges from the combination of other well-known solutions, all adhering to agile principles. It is characterized by prioritizing software functionality over extensive documentation, client participation over rigid contractual negotiation, and flexibility in the face of changes rather than strict adherence to a predefined plan. It consists of five phases: exploration, initialization, production, stabilization, and system testing.

This study employed a pre-experimental single-group design with pretest and posttest. According to study [20], this type of design allows the evaluation of the effects of an intervention or treatment on a specific sample. It is characterized by two aspects: the use of a single group of participants and a linear order requiring measurement of the dependent variable before and after the intervention or treatment is implemented.

This research was classified as applied and adopted a quantitative approach, with an explanatory level and a pre-experimental design. The study population was selected using non-probabilistic convenience sampling, comprising 22 workers, considering the time and resource constraints for the research. Therefore, the size is the same as the population.

To measure the key indicators of the study, carefully selected techniques and instruments were employed. The "Average hiring time for personnel" and the "Number of hirings" were analyzed using the ratio technique, with an observation sheet that enabled the collection of accurate and quantifiable data on these aspects. Meanwhile, the "Satisfaction level with services" was assessed using a Likert scale, through a structured questionnaire designed to capture users' perceptions and satisfaction regarding the service. These instruments were validated and approved by experts to ensure alignment with the study objectives, guaranteeing the reliability and validity of the data collected. Table I shows indicators, techniques and instruments.

TABLE I. Indicators, Techniques and Instruments

Indicator	Technique	Instrument
Average hiring time for personnel	Ratio	Observation sheet
Number of hirings	Ratio	Observation sheet
Satisfaction level with services	Likert Scale	Questionnaire

V. Methodology

The development of the mobile application was conducted using the Mobile-D methodology, which comprises the following phases: (A) Exploration, (B) Initialization, (C) Production, (D) Stabilization, and (E) System Testing.

A. Exploration

During this phase, stakeholders were identified, and the project foundations were presented, considering its objectives and requirements. Additionally, the project scope was established to ensure all aspects aligned with stakeholders' expectations. The functional requirements are detailed in Table II below:

TABLE II. Functional and No-Functional Requirements

Code	Description
RF-001	Users must be able to log in according to their type (Customer or Tasker).
RF-002	Users must be able to view their service history.
RF-003	Users must be able to view their current hired services.
RF-004	Users must be able to access their active chats.
RF-005	Users must be able to update their profiles.
RF-006	The system must allow users to rate a service after its completion.
RF-007	Users must be able to search for a service by category.
RF-008	The system must provide users with nearby services or clients based on geolocation.
RF-009	The system must allow users to cancel a service request.
RF-010	The system must enable Taskers to schedule a service for a specific date and time.

B. Initialization

Fig. 1 illustrates the development environment, utilizing Visual Studio Code for the frontend and Spring Tool Suite 4 for the backend, along with the installation of JDK 1.8 (Java 8 SE). The database will be managed using SQL Server Management Studio 20. Additionally, the development team employed the Flutter framework and the Dart programming language. Communication with clients will be facilitated through email and phone calls.

Solution Architecture

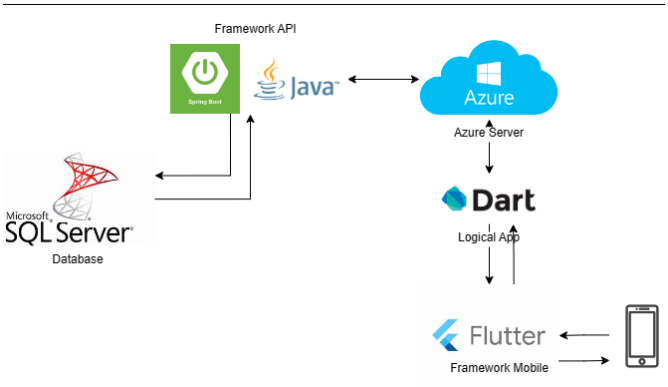


Fig. 1. Software architecture diagram.

C. Production

Planning Day (see Table III)

TABLE III. ITERATION PLANNING BY PHASE

Fase	Iterations	Description
Exploration	Iteration 0	Establishment of stakeholders, functional and non-functional requirements, scope definition, and selection of development technologies.
Initialization	Iteration 0	Creation of the mobile application architecture, requirements analysis, and interface design
Production	Iteration 1	Verification of selected tools. Development of the mobile app interfaces and database implementation and connection to Azure. Creation of the GitHub repository.
	Iteration 2	Implementation of user registration and login functionality: Tasker and Customer, using Firebase OTP service. Visualization of service categories (Service table).
	Iteration 3	User data modification through the app. Implementation of geolocation for users to search for nearby services using Google Places.
	Iteration 4	Implementation of chat between users using Firebase. Request logging implementation.
	Iteration 5	Implementation of user reviews. View of pending and completed request history.

Workday

Iteration 1. The selected tools for development were verified, and the project setup was completed, including project compilation, library downloads, credential creation for services, and proper connection to the SQL Server database. This database was uploaded to Microsoft Azure to enable simultaneous use and faster performance for the development team. For the design of interfaces, they were developed following the functional requirements outlined in the first phase, emphasizing the application's usability. Finally, the backend and frontend projects were uploaded to a GitHub repository for version control. Additionally, a web application was created on the Microsoft Azure platform, linking it to the backend project via GitHub.

Iteration 2. Using stored procedures in SQL Server and APIs in the backend project, the user registration and login functionality were developed through Firebase's OTP service. As shown in Fig. 2, users can authenticate using their phone number, receiving a unique code to be entered in the application interface. Based on stored procedure verifications of user existence in the database and the selected user type (Tasker or Customer), they are redirected to the Registration interface (as shown in Fig. 3 and Fig. 4) or the Login page. The Customer is presented with the Home screen, listing Services and Top Taskers, along with quick access to their chats. The Tasker, in turn, sees the Home screen with unopened requests and direct access to their chats. Fig. 5 show customer main menu and service search by category.

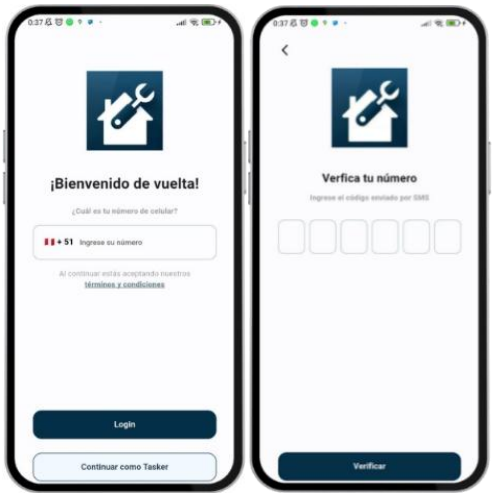


Fig. 2. OTP registration.

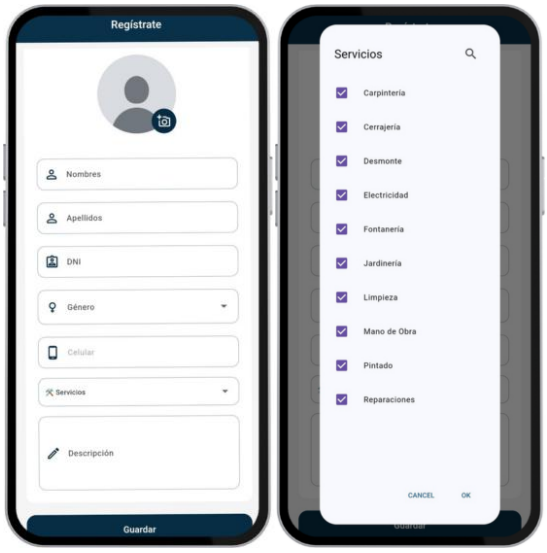


Fig. 3. Tasker registration.

Iteration 3. From the user perspective, they can now access their profile to modify their data using stored procedures. Additionally, the Google Places API will be implemented and utilized on the front end. When a user has location services enabled, their latitude and longitude data will be automatically captured and sent to the backend, allowing them to be mapped and enabling the Customer to see the closest Tasker in their area, as shown in Fig. 6 and Fig. 7.

Fig. 4. Customer registration.

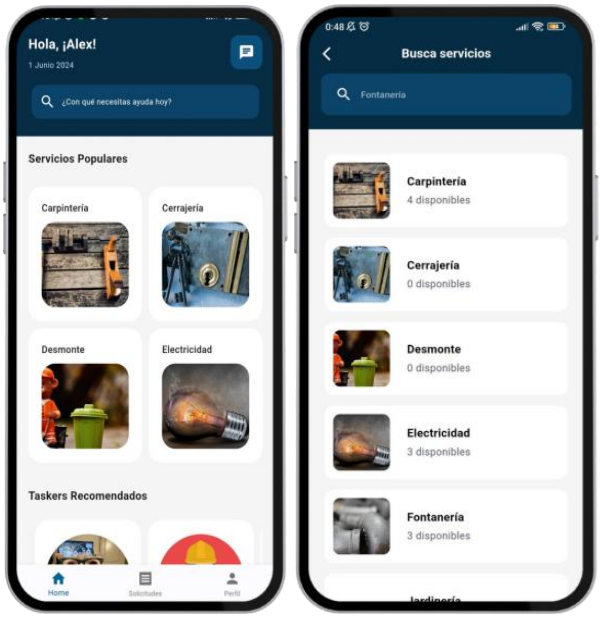


Fig. 5. Customer main menu and service search by category.

Iteration 4. During this iteration, Firebase libraries were integrated to enable chat and notification services within the application. Requests are managed with four different statuses: deleted, accepted, completed, and pending. In the Tasker view, as shown in Fig. 8 and Fig. 9, a redirection to a request registration modal was implemented, where the necessary details specified by the client are entered to confirm the request.

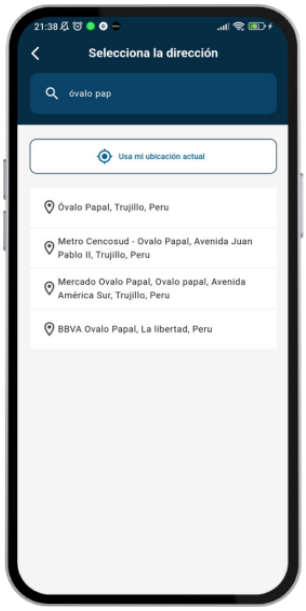


Fig. 6. Location selection via geolocation or search.

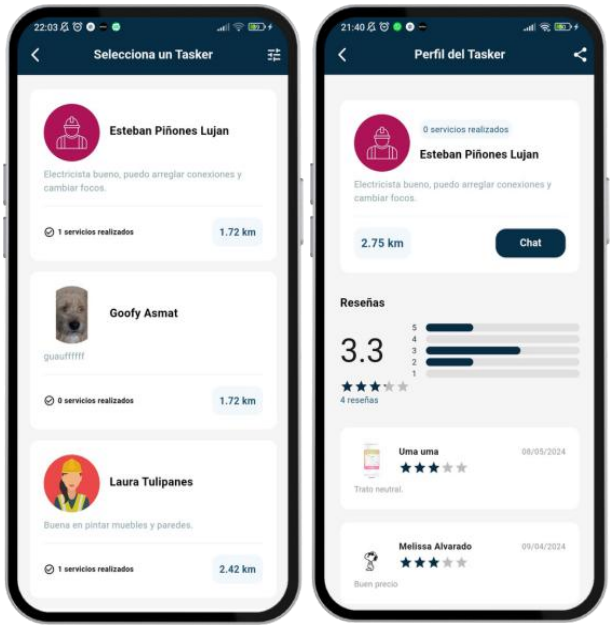


Fig. 7. Selection and profile of tasker near current location.

Iteration 5. In this iteration, as illustrated in Fig. 10, the ability to view the history of pending and completed requests was implemented, providing users with a comprehensive overview of their past interactions in the application. Additionally, a review feature was added, allowing users to leave comments and ratings on the services received once completed, as shown on Fig. 11. Consequently, Customers now see recommendations for Taskers with the highest average ratings on their home screen, while Taskers can view the ratings of the Customers requesting their services.

Fig. 8. Chat history.

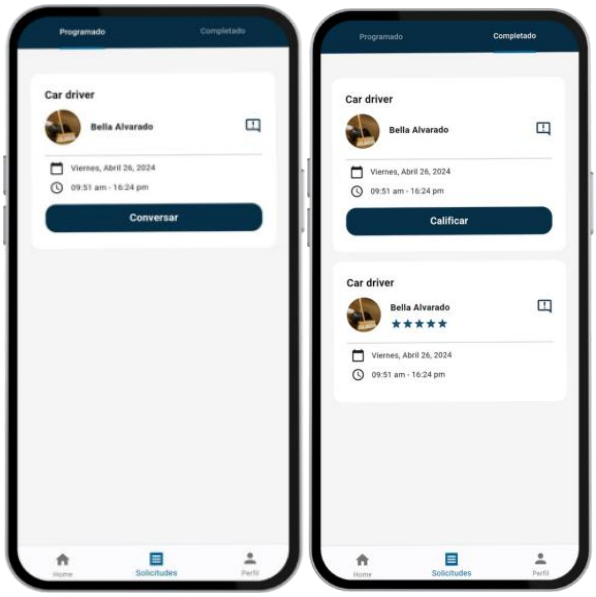


Fig. 10. History of scheduled and completed requests.

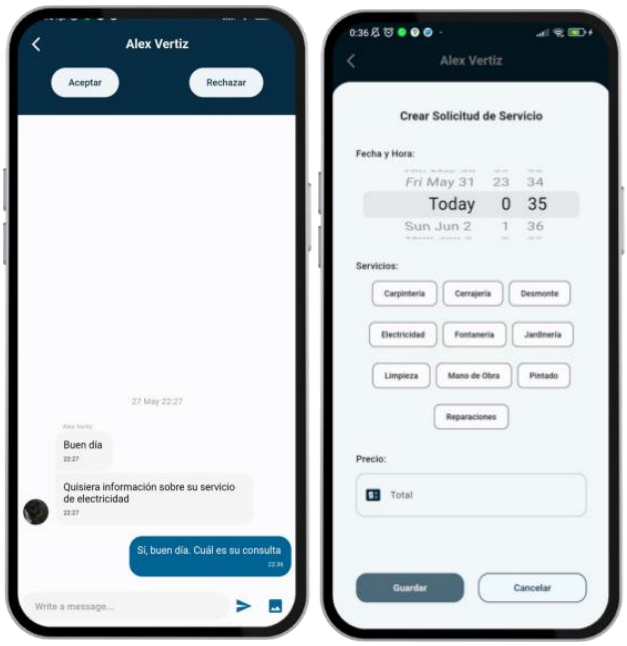


Fig. 9. Chat and service request creation.

Fig. 11. Service rating.

D. Stabilization

In this phase, a modular structure was implemented, allowing for clearer organization in both the application's architecture and interface. A bottom navigation bar was added, making it easier to access features. This improvement enhances the user experience by providing smoother interaction.

Algorithm 1: MenuTasker Navigation

```
Initialize
    Create MenuTasker as a stateful widget
    Create NavigationControllerTasker as a navigation

Compute
    Initialize controller with NavigationControllerTasker
While (the application is running) do
    Build the user interface
    For (each region of the interface) do
        Display BottomNavigationBar with three options:
            - Home
            - Requests
            - Profile
        Show the corresponding screen based on selectedIndex
    Update
        Observe changes in selectedIndex
        Update and analyze
            If (selectedIndex changes) then
                Search for the corresponding screen in the screens list
    End                 End
    End
```

E. System Testing

During this phase, unit tests were conducted to review the code and ensure that the developed and implemented product operated correctly according to the specified functionalities. Additionally, time was allocated to identify and fix any errors detected during the process.

TABLE IV. UNIT TEST

Test Scenario	Expected Outcome		Obtained result	
			RT	OK
User	Verificar Signup Tasker Service ()	searchTasker	332ms	X
		searchTaskerByUid testing	441ms	X
		saveTasker testing	285ms	X
		updateTasker testing	278ms	X
		updateProfileTasker testing	299ms	X
User	Verificar SearchCustomer Service ()	searchCustomer testing	285ms	X
		searchCustomerByUid testing	281ms	X
		updateCustomer testing	281ms	X

As shown in Table IV, the results demonstrate that the code is effective and adequately fulfills the expected functionality.

VI. RESULTS

The results presented in Table V demonstrate a high reliability of the measurement instruments used in the research. The obtained coefficients exceeded the 0.70 threshold, indicating adequate consistency in the measurements.

TABLE V. RELIABILITY

Instrument	Method	Indicator	Coefficient
Pre- and Post-Observation Sheet	Test - retest	General services contracting time	.910
Pre- and Post-Observation Sheet	Test - retest	Number of Contracts	0.843
Pre and post Questionnaire	Cronbach's Alpha	Satisfaction Level	.853

Data elaborated based on SPSS.

A. Descriptive Analysis

Indicator 1 General services contracting time

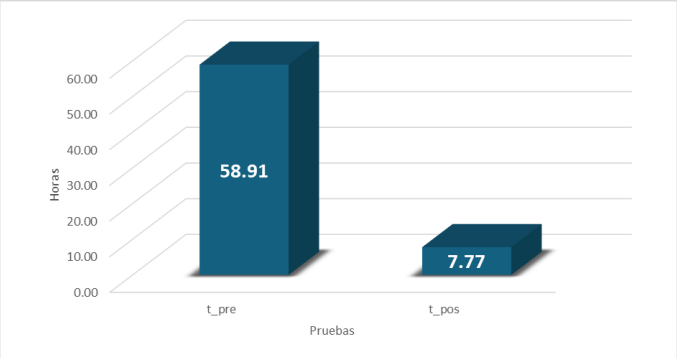


Fig. 12. Average service contracting time before and after.

According to Fig. 12, the average time to contract general services before the solution's implementation was 59 hours, and the post-implementation time was eight hours. The significant reduction in the average time after implementation suggests that the application had a positive and considerable impact on the contracting process, improving the efficiency of worker selection.

Indicator 2 Number of Contracts

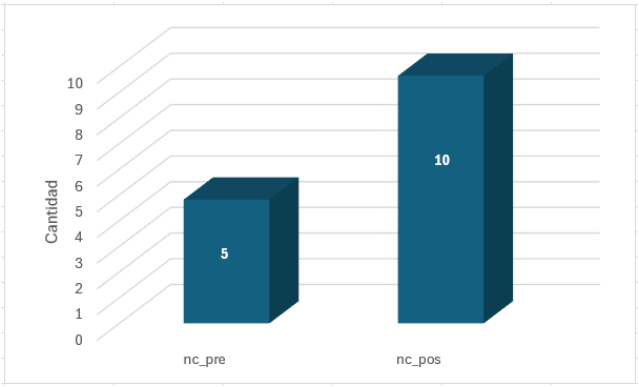


Fig. 13. Average number of contracts before and after.

According to Fig. 13, the average number of general service contracts before the implementation of the solution was 5, while after the implementation, 10 contracts were recorded. This notable increase in the number of requests demonstrates the positive impact of the solution's implementation, doubling the number of contracts and significantly improving the efficiency of the process.

Indicator 3 Satisfaction with the Contracting Process

TABLE VI. LEVEL OF SATISFACTION BEFORE IMPLEMENTATION

Category	Frequency	Percentage
Strongly disagree	7	31.82%
Disagree	13	59.09%
Neither agree nor disagree	2	9.09%
Agree	0	0%
Strongly agree	0	0%
Total	22	100%

Data elaborated from SPSS

TABLE VII. LEVEL OF SATISFACTION AFTER IMPLEMENTATION

Category	Frequency	Percentage
Strongly disagree	0	0%
Disagree	0	0%
Neither agree nor disagree	0	0%
Agree	11	50%
Strongly agree	11	50%
Total	22	100%

Data elaborated from SPSS

According to Tables VI and VII, the predominant perception of 22 workers regarding satisfaction before the implementation of the mobile application in the general service hiring process was: totally disagree 7 (31.82%), disagree 13 (59.09%), and neither agree nor disagree 2 (9.09%). However, after the implementation of the mobile application, a considerable positive impact on satisfaction was observed, with a notable increase in perception: agree 11 (50%) and totally agree 11 (50%).

B. Inferential Analysis

1) Normality test: Based on the results from Tables VIII, IX, X and XI, obtained from a sample of 22 data points, it is observed that most indicators show an asymptotic significance lower than 0.05, which suggests a non-normal distribution of the data. Therefore, the non-parametric Wilcoxon test for paired samples is used for each indicator.

TABLE VIII. SHAPIRO-WILK TEST FOR AVERAGE TIME OF THE GENERAL SERVICE HIRING PROCESS

	Statistic	*gl.*	*Sig.*
t_pre	*.887*	*22*	*.017*
t_pos	*.782*	*22*	*<.001*

Data elaborated from SPSS

TABLE IX. SHAPIRO-WILK TEST FOR AVERAGE NUMBER OF CONTRACTS IN THE GENERAL SERVICE HIRING PROCESS

	Statistic	gl.	Sig.
t_pre	.875	22	.010
t_pos	.854	22	.004

Data elaborated from SPSS

TABLE X. SHAPIRO-WILK TEST FOR AVERAGE SATISFACTION WITH THE CONTRACTING PROCESS FOR GENERAL SERVICES

	Statistic	gl.	Sig.
s_pre	.912	22	.052
s_pos	.920	22	.076

Data elaborated from SPSS

2) Hypothesis formulation

a) Alternative hypothesis: The implementation of a mobile application based on geolocation significantly improves the time, satisfaction, and number of contracts in the general service hiring process in Trujillo, La Libertad.

b) Null hypothesis: The implementation of a mobile application based on geolocation does not significantly improve the time, satisfaction, or number of contracts in the general service hiring process in Trujillo, La Libertad.

3) Hypothesis testing: Indicator 1 Contracting Process Time

TABLE XI. WILCOXON TEST FOR AVERAGE CONTRACTING PROCESS TIME FOR GENERAL SERVICES

Null Hypothesis	Test	Sig.	Decision
The median difference between t_pre and t_post is equal to 0	Wilcoxon signed-rank test for related samples	.000	Reject the null hypothesis

Data elaborated from SPSS

Since the p-value is equal to 0, which is less than 0.05, we reject the null hypothesis and accept the alternative hypothesis. This means that the implementation of a mobile application based on geolocation significantly improves the contracting process time for general services hiring process in Trujillo, La Libertad.

Indicator 2 Number of Contracts

TABLE XII. WILCOXON TEST FOR AVERAGE CONTRACTING PROCESS TIME FOR GENERAL SERVICES

Null Hypothesis	Test	Sig.	Decision
The median difference between nc_pre and nc_post is equal to 0	Wilcoxon signed-rank test for related samples	.000	Reject the null hypothesis

Data elaborated from SPSS

Since the p-value is equal to 0, which is less than 0.05, we reject the null hypothesis and accept the alternative hypothesis. This means that the implementation of a mobile application based on geolocation significantly improves the number of contracts in the general services hiring process in Trujillo, La Libertad (see Table XII and XIII).

Indicator 3 Satisfaction with the Contracting Process

TABLE XIII. WILCOXON TEST FOR AVERAGE SATISFACTION WITH THE CONTRACTING PROCESS FOR GENERAL SERVICES

Null Hypothesis	Test	Sig.	Decision
The median difference between s_pre and s_post is equal to 0	Wilcoxon signed-rank test for related samples	.000	Reject the null hypothesis

Data elaborated from SPSS

Since the p-value is equal to 0, which is less than 0.05, we reject the null hypothesis and accept the alternative hypothesis. This means that the implementation of a mobile application based on geolocation significantly improves satisfaction in the contracting process for general services in Trujillo, La Libertad.

VII. DISCUSSION

According to the results obtained, the times, satisfaction, and number of contracts in the hiring process were analyzed and compared before (pre) and after (post) the implementation of the mobile application. This app was developed using a hybrid framework to ensure cross-platform compatibility and efficient performance. The objective was to improve the efficiency of the general service hiring process through the implementation of a mobile application based on geolocation in Trujillo, La Libertad.

The results of the study reveal a significant reduction of 51.14 hours in the average time of the general service hiring process after the implementation of the mobile application. This substantial decrease demonstrates the effectiveness of mobile technology in optimizing organizational processes. The findings align with previous research [8], [13], which highlighted the potential of geolocation-based mobile applications to streamline operations and improve efficiency. Furthermore, the integration of Google Places API allowed users to easily find the nearest available taskers, further enhancing the speed and accessibility of the hiring process. These results corroborate the research proposed by study [20] regarding the importance of time as a key performance indicator in various functional areas.

After the implementation of the mobile application, a notable 50% increase in the number of contracts was observed, rising from an initial average of 5 to 10 contracts. This increase clearly reflects the ease and effectiveness the application offers in accelerating the general service hiring process. These results are consistent with those reported in study [17], where it was demonstrated that mobile applications not only optimize process organization but also significantly expand access to the job market. Similarly, the results reinforce the argument made by [9], which suggests that the development of a location-based job portal has a positive impact on increasing the number of contracts.

Regarding satisfaction levels, an increase in the satisfaction average from 55% to 83.2% was observed after the mobile application implementation, indicating that employees are more satisfied with the hiring process for their services after adopting this tool. These findings coincide with those reported in study [12], where it was determined that the implementation of a location-based mobile application significantly improved technician satisfaction by 89.17% by providing a more organized environment. Similarly, the study in [13] highlighted that the development of a location-based mobile application for facilitating service search increased satisfaction in the hiring process management. These results also reinforce the claims of [18], who argues that satisfaction is achieved through the standardization and optimization of processes, which enhances quality and, consequently, overall satisfaction.

VIII. CONCLUSION

The implementation of a mobile application based on geolocation has proven to be effective in improving the efficiency of the general service hiring process in Trujillo, La Libertad, by significantly reducing the average process time and increasing the satisfaction level and number of worker contracts.

The average time of the hiring process was reduced from 58.91 hours before implementation to 7.77 hours after, representing an 86.79% decrease. This result was confirmed by the Wilcoxon test, which showed a significant level lower than 5%, confirming the acceptance of the alternative hypothesis.

Similarly, the average number of worker contracts in the hiring process increased from 5 to 10 after the implementation of the mobile application, which equals a 50% increase. This increase was also validated by the Wilcoxon test, with a significance level lower than 5%, reaffirming the effectiveness of the technological intervention.

Additionally, the average worker satisfaction in the hiring process increased from 55% to 83.2% after the mobile application implementation, which represents a 51.27% increase. This increase was also validated by the Wilcoxon test, with a significance level lower than 5%, reaffirming the effectiveness of the technological intervention.

Finally, this study contributes to computing by demonstrating how cloud-based data management, real-time communication, and location-based service matching using Google APIs optimize service efficiency and user experience. Additionally, the platform implemented a ranking algorithm that sorts service providers based on their ratings, ensuring that users can easily find the highest-rated professionals, leading to a more reliable and efficient hiring experience

ACKNOWLEDGMENT

The authors would like to thank the Universidad Privada del Norte for providing us with the necessary knowledge, our advisor for guiding us, our parents and our pets for their unconditional support throughout these years of study.

REFERENCES

[1] K. Dharani, S. Bhatti, A. Dewani, E. Rajput, and A. Ayaz, “Renovate-It: A geo-based technical professional hiring system for repairing and maintenance services,” in 2018 International Conference on Computing, Mathematics and Engineering Technologies: Invent, Innovate and Integrate for Socioeconomic Development, iCoMET 2018 - Proceedings, 2018. doi: 10.1109/ICOMET.2018.8346318.

[2] M. N. Islam, M. Arman Ahmed, and A. K. M. Najmul Islam, “Chakuri-bazaar: A mobile application for illiterate and semi-literate people for searching employment,” International Journal of Mobile Human Computer Interaction, vol. 12, no. 2, 2020, doi: 10.4018/IJMHCI.2020040102.

[3] J. L. A. Espinoza, A. R. L. L. Yacelga, and W. G. S. Michilena, “MOBILE APPLICATIONS AND THEIR IMPACT ON SOCIETY | LAS

APLICACIONES MÓVILES Y SU IMPACTO EN LA SOCIEDAD," Universidad y Sociedad, vol. 14, no. 2, pp. 237–243, 2022.

[4] G. Karaoglu, E. Hargittai, and M. H. Nguyen, "Inequality in online job searching in the age of social media," Inf Commun Soc, vol. 25, no. 12, 2022, doi: 10.1080/1369118X.2021.1897150.

[5] E. Gil, "Employment Deindustrialization: Understanding Jobless Growth in the Manufacturing Industry," SSRN Electronic Journal, vol. 25, no. 3, pp. 30–43, 2020, doi: 10.2139/ssrn.4194319.

[6] Y. Feng, D. Lagakos, and J. E. Rauch, "Unemployment and Development," The Economic Journal, vol. 134, no. 658, pp. 614–647, Jan. 2024, doi: 10.1093/ej/uead076.

[7] H. D. Rojas Cisneros, "Aplicación móvil para la contratación de servicios profesionales," REVISTA ODIGOS, vol. 2, no. 3, pp. 27–45, Oct. 2021, doi: 10.35290/ro.v2n3.2021.475.

[8] M. Meccawy et al., "The graduate helper: Using a mobile application as a feasible resource for job hunting across Saudi Arabia," International Journal of Interactive Mobile Technologies, vol. 12, no. 4, 2018, doi: 10.3991/ijim.v12i4.7566.

[9] A. Aligbe, N. G. Ikechukwu, O. Okoyeigbo, A. Uyi, A. Olajube, and A. U. Adoghe, "Development and Implementation of Location-Based Mobile Job Portal for Blue-Collar Jobs in Nigeria," IOP Conf Ser Mater Sci Eng, vol. 1107, no. 1, p. 012098, Apr. 2021, doi: 10.1088/1757-899X/1107/1/012098.

[10] M. Gabriel and M. Thyssen, "Report of the high-level expert group on the impact of the digital transformation on EU labour markets," Publications Office of the European Union, 2019.

[11] T. Weber, J. Hurley, and D. Adăscăliței, "COVID-19: Implications for employment and working life," European Foundation for the Improvement of Living and Working Conditions, p. 86, 2021.

[12] A. A. Tandoc, S. A. Marie B. Arguiñoso, S. M. F. Carungay, and K. S. Tafalla, "HandyFix: A Service Design for a Location-Based Mobile Application for Acquiring On-Demand Local Handyman Services in Metro Manila," in Proceedings of the International Conference on Industrial Engineering and Operations Management, Michigan, USA: IEOM Society International, Sep. 2023. doi: 10.46254/AP04.20230059.

[13] K. Foo, "SERVICE FINDER: An Android-based Service Finding and Booking Application," Applied Information Technology and Computer Science, vol. 4, no. 1, pp. 466–485, 2023.

[14] H. Bhaskar, K. Rao, P. Bhandarkar, P. Prakash, and G. Laxmi, "An Android Application for Home Services," An Android Application for Home Services, vol. 7, no. 5, 2020.

[15] A. Gothankar, N. Yadav, A. Animesh, U. Rautgol, and P. Kulkarni, "A Survey of on Demand Home Services," International Journal of Advanced Research in Science, Communication and Technology, vol. 4, no. 2, pp. 561–564, Jan. 2024, doi: 10.48175/IJARSCT-15289.

[16] Y. Shubham, T. Rupesh, T. Bhimsen, and S. Vikas, "On demand home services (servizio)," i-manager's Journal on Future Engineering and Technology, vol. 18, no. 3, p. 29, 2023, doi: 10.26634/jfet.18.3.19496.

[17] I. Verma, A. Kumar, A. Tripathi, Y. Sain, and A. Sharma, "Service Providers for Home Appliances," Journal of Positive School Psychology, vol. 6, no. 3, pp. 7215–7219, 2022.

[18] D. I. N. Quispe, J. M. N. Quispe, J. L. H. Salazar, and J. P. Cruzado, "Mobile App for the Promotion of Home Services," in Proceedings of the 2020 IEEE Engineering International Research Conference, EIRCON 2020, 2020. doi: 10.1109/EIRCON51178.2020.9254079.

[19] J. R. Molina Ríos, J. A. Honores Tapia, N. Pedreira-Souto, and H. P. Pardo León, "Comparativa de metodologías de desarrollo de aplicaciones móviles," 3C Tecnología_Glosas de innovación aplicadas a la pyme, vol. 10, no. 2, pp. 73–93, Jun. 2021, doi: 10.17993/3ctecno/2021.v10n2e38.73-93.

[20] A. YILMAZ and S. DUYGULU, "Developing Psychological Empowerment and Patient Safety Culture: A Pre-experimental Study," Journal of Basic and Clinical Health Sciences, vol. 5, no. 2, pp. 94–103, May 2021, doi: 10.30621/jbachs.907526.

Development of a Software Tool for Learning the Fundamentals of CubeSat Angular Motion

Victor Romero-Alva[1], Angelo Espinoza-Valles[2]
Faculty of Engineering, Universidad Tecnológica del Perú, Peru[1]
Inter-University Department of Space Research, Samara National Research University, Russian Federation[2]

***Abstract*—The development of tools for understanding and simulating CubeSat angular motion is essential for both educational and research purposes in space technology. In this context, this paper presents the development of a MATLAB-based software tool designed to facilitate the comprehension of CubeSat angular motion. This tool allows users to simulate CubeSat dynamics by adjusting parameters, such as initial conditions and physical properties, enabling the observation of different types of motion, including rotatory, oscillatory, both stable and unstable behaviors. The mathematical models selected for simulating the CubeSat dynamics are presented. The interface of the tool, designed for intuitive parameter input and visualization of phase portraits of the system under consideration, is described. The software is demonstrated using a CubeSat 3U configuration, and simulation results, including angle of attack, angular velocity, and altitude decay, are presented. This tool aims to enhance the understanding of CubeSat angular motion, contributing to the design and operation of CubeSat missions in low Earth orbit.**

Keywords*—CubeSat; angular motion; simulation; learning tool MATLAB

I. Introduction

CubeSats, a class of nanosatellites, have become popular in educational and commercial activities in space due to their cost effectiveness and versatility [1, 2]. The proliferation of CubeSats, typically deployed in low Earth orbit (LEO), has transformed space exploration and satellite-based research. Previously, solutions to many complex space problems could be imagined possible only with the use of larger spacecraft [3-5].

One critical aspect of their operation is angular motion, which determines orientation of the spacecraft in space. Understanding how a CubeSat experiences angular motion and how to control its attitude is vital for mission success, especially for applications involving imaging, communication, and scientific measurements [6, 7]. Despite their relatively small dimensions, CubeSats adhere to the same fundamental principles of orbital mechanics and attitude control as larger spacecraft. These principles are influenced by external forces such as gravitational and aerodynamic torques, which can vary based on the altitude, mass distribution, and geometric properties of the satellite [8].

In practice, achieving stable or predictable attitude control is essential for both passive and active stabilization methods and requires a firm comprehension of how these forces are related to the physical characteristics of the CubeSat. Given the increasing complexity of CubeSat missions, there is a growing need for intuitive and interactive learning tools that simplify the study of its dynamics.

Traditional methods of teaching these concepts often rely on theoretical mathematical models, which can be difficult for learners to fully comprehend without visual aids or practical experimentation. Consequently, in recent years, the development of graphical user interfaces as learning tools has gained traction across multidisciplinary fields, demonstrating their potential to facilitate complex subject matter through interactive and intuitive designs.

For instance, Botha and Marais developed a GUI to support learning in artificial intelligence, demonstrating the effectiveness of visual and interactive tools in simplifying sophisticated AI concepts [9]. Similarly, Mohd and Hashim introduced a deep learning interface designed to assess fruit quality, which exemplifies how such interfaces can aid in applied fields like food engineering [10]. These methodologies are also employed in chemometrics, as seen in the work of Chiappini et al., who developed a MATLAB GUI for multivariate calibration [11]. Gasparic et al. developed a GUI that enhances integrated development environment usability by providing command recommendations, demonstrating through user studies that such interfaces are essential for effective developer support [12]. Mishra et al. [13] developed a MATLAB-based interface that simplifies multi-block data analysis by offering integrated visualization, classification, and pre-processing capabilities, thereby making advanced chemometric methods accessible to users in industrial applications. Victoria et al. introduced a GUI for topology design in engineering structures, which underscores the broad applicability of such interfaces in engineering design processes [14]. Furthermore, GUIs have been implemented in astronomy, for example, Errazzouki et al. created a MATLAB interface to facilitate the acquisition of single star SCIDAR data, facilitating data processing [15].

These developments collectively underscore the effectiveness of interactive graphical interfaces in making complex theoretical concepts more accessible and engaging. However, despite these successful implementations, there remains a distinct need for a simplified simulation tool that provides relatively accurate representations of CubeSat dynamics while remaining accessible to students and newcomer CubeSat developers. Thus, the objective of this paper is to present a software learning tool designed to simulate the CubeSat angular motion through an intuitive interface, while

facilitating an easy understanding of CubeSat angular motion fundamentals. This MATLAB-based tool allows users to adjust key design parameters, observe different types of motion, and visualize phase portraits of the system, providing valuable insights into CubeSat dynamics.

The remainder of the article is organized as follows: Section II reviews related work in tools for simulating spacecraft dynamics. Section III presents the theoretical concepts underlying CubeSat angular motion and describes the mathematical models implemented in the tool, with emphasis on the key assumptions made for model simplification. Section IV presents the results, demonstrating the utility of the tool through a practical example based on a CubeSat 3U configuration. Section V provides a discussion of the results, highlighting the limitations, and potential areas for future work. Finally, Section VI presents the conclusions, summarizing the key contributions and significance of the developed tool."

II. Related Work

In the field of spacecraft dynamics, several sophisticated software tools and packages have been developed to simulate both angular and orbital motion of spacecraft. For instance, Hadi and Sasongko developed a CubeSat design and visualization tool in MATLAB that not only calculates key satellite parameters but also simulates dynamic responses, thereby supporting the analysis and design of CubeSat systems [16].

Ivanov et al. introduced a software package capable of simulating passive and controlled angular and orbital motion of near-Earth satellites [17]. While highly versatile, the reliance on specific mathematical models can make it less accessible for users seeking a simpler introduction to CubeSat dynamics. Moreover, as the software is developed in C++, it may be more difficult to modify or add new modules compared to more user-friendly environments like MATLAB.

Ezzat et al. provided a tool for satellite orbit tracking based on the Keplerian system. Although this method demonstrates high precision and robust visualization for orbit tracking, it does not simulate the angular motion relative to the spacecraft's center of mass. This omission is significant because understanding the angular dynamics is critical for evaluating the attitude stability and control of the spacecraft, which are essential for mission success [18].

Similarly, Turner developed an open-source spacecraft simulation and control software based on an application programming interface intended for both research and educational purposes [19]. This framework delivers high-fidelity simulation routines and efficient, fast-running code, making it a powerful tool for detailed analysis. However, its effective use requires a robust understanding of programming, which may present a barrier for users with limited technical expertise.

Shirobokov and Trofimov developed the KIAM Astrodynamics Toolbox, a robust software library for spacecraft orbital motion that incorporates Fortran modules for implementing astrodynamical functions and Python modules for their compilation [20]. While the toolbox is effective for detailed simulations and ensures computational speed, its primary focus on orbital dynamics may overlook key aspects of CubeSat attitude control and angular motion. While this advanced tool is effective, it may not be suitable for educational purposes. The reliance on a traditional programming language, as Fortran, alongside Python interfaces, can pose accessibility challenges for users without a solid programming background.

While the reviewed tools offer detailed and accurate simulations suitable for advanced research, there is a noticeable gap when it comes to introductory educational applications. Many of these tools are designed with high-fidelity modeling in mind, which can sometimes render them less accessible for beginners or users without extensive programming experience. In summary, the literature reveals two key trends in the development of educational simulation tools: the integration of graphical user interfaces to help demystify theoretical concepts and the growing use of open-source, extensible software frameworks that allow for both accurate and customizable simulations.

III. Theory of CubeSat Angular Motion

In this section, we provide an overview of the theoretical concepts and mathematical models, which underlie CubeSat angular motion, focusing on dynamic equations generally used in CubeSat missions.

A. Simplifications and Assumptions for CubeSat Dynamics Modeling

In the development of this interactive learning tool for CubeSat angular motion, several foundational assumptions are made to simplify the complex dynamics involved and facilitate an effective educational framework.

A primary assumption is that the CubeSat exhibits planar angular motion in a circular orbit, subject to the influence of gravitational and aerodynamic torques. This assumption enables a focused analysis of the rotational behavior by limiting the scope to the two principal forces acting on the CubeSat. This assumption is particularly appropriate for CubeSats operating in LEO, allowing the model to provide a simplified and close representation of their attitude dynamics.

Given that the geometry of a CubeSat is generally a parallelepiped, it can be considered a dynamically symmetric body. Therefore, its moments of inertia about the *y*- and *z*-axes are assumed to be equal ($I = I_y = I_z$). Furthermore, for form factors larger than 1U, the longitudinal moment of inertia I_x is associated with the *x*-axis, which is the axis of symmetry and corresponds to the longest side of the CubeSat.

Due to the typical vertical stacking of components in CubeSats, it is assumed that the center of mass may be displaced from the center of pressure, primarily along the longitudinal axis, denoted as Δx. This is because the arrangement of components is more likely to create asymmetry along this axis. In contrast, displacements along the lateral *y*-axis and *z*-axis are considered negligible or zero. This assumption simplifies the analysis, particularly in assessing the impact of aerodynamic forces.

These simplifications are essential for creating a simulation that is both computationally manageable and pedagogically

valuable for understanding the principles of angular motion in CubeSats.

B. Mathematical Model for CubeSat Angular Motion Modeling

In this study, the motion of a CubeSat relative to its center of mass is examined and its implemented based on the theorem of conservation of angular momentum.

According to study [21], an approximate model for the angular motion of a dynamically symmetric CubeSat in a circular orbit can be obtained by the following the expression:

$$\ddot{\alpha} - \frac{M_{ay}}{I}\sin\alpha - \frac{M_{gy}}{I}\sin\alpha = 0, \quad (1)$$

where, α is the angle of attack, M_{ay} is the aerodynamic moment, M_{gy} is the gravitational moment and I is the CubeSat transverse moment of inertia.

At lower altitudes, aerodynamic forces play a significant role in the dynamics of the CubeSat, while at higher altitudes, the influence of aerodynamic moments diminishes, and gravitational moments become more predominant. Despite the weakening of gravitational force with increasing altitude, its effects remain relevant and must be thoroughly analyzed to ensure effective control and stabilization. In this context, the gravitational moment is calculated using the following formula:

$$M_{gy} = -\frac{3\mu}{2r^3}(I_x - I)\sin 2\alpha \quad (2)$$

where, μ is the Earth's gravitational parameter, r is the CubeSat position radius-vector, I_x is the CubeSat longitudinal moment of inertia.

For CubeSats, the angular acceleration caused by aerodynamic moments is significantly higher. These moments have a much greater influence on smaller CubeSats than on larger, more massive ones, even when both have the same relative center of mass displacement and volumetric density [21]. The moment of aerodynamic drag force is calculated as follows:

$$M_{ay} = -c_0 S\Delta\bar{x}ql\sin\alpha \quad (3)$$

where $c_0 = 2.2$ is the drag coefficient, S is the characteristic area of the CubeSat perpendicular to the flight velocity vector, $\Delta\bar{x} = \Delta x / l$ is the relative distance from the center of pressure to the center of mass along the x axis of the CubeSat, q is the velocity pressure, l is the characteristic length of the CubeSat.

In the previous model, the orbital dynamics of the CubeSat were heavily dependent on the flight altitude, as this parameter directly influences atmospheric drag and other perturbative forces. Given the significance of altitude in attitude, we aim to utilize the approximate formula to predict altitude variations over time. This allows for a close estimation of the orbital evolution, especially in LEOs where altitude plays a critical role in its long-term stability and performance. For CubeSats, the ballistic coefficient tends to be higher compared to larger satellites, which leads to a shorter orbital lifetime due to the increased drag forces acting upon them at a given altitude. Under the assumption that the descent angle of a nanosatellite in a nearly circular orbit remains small and varies slowly, the altitude change over time can be estimated using the equation [22]:

$$\dot{h} = -\frac{2\sigma qV}{g} \quad (4)$$

where σ is the ballistic coefficient, V is the flight velocity and g is the Earth's gravity at a specific altitude.

Assuming the flow is free molecular and the impact of gas molecules is completely inelastic, the ballistic coefficient of a CubeSat can be determined by the following formula [8]:

$$\sigma = -\frac{c_0 S}{m} \quad (5)$$

where m is the CubeSat mass.

IV. Results

In this section the developed software tool, the CubeSat Angular Motion Simulator, is introduced.

A. Learning Tool Interface Overview

The proposed interface was implemented and verified using MATLAB programming language. This tool provides an interactive environment where users can explore the dynamic behavior of CubeSats by modifying various physical and initial conditions of the angular motion. The primary features include real-time visualization of angular motion, user input of critical parameters, and the generation of phase portraits, allowing for an in-depth exploration of both theoretical and practical aspects of CubeSat dynamics.

The interface of the CubeSat Angular Motion Simulator is designed for ease of use, as shown in Fig. 1. It is divided into several distinct panels: one for inputting CubeSat parameters, another for initializing angular motion conditions, a third for adjusting simulation time, and finally, a visualization panel that plots the phase portraits of angular motion.

B. Initial Data for Simulation

The input data required for the simulation includes a range of essential parameters that define the dynamic characteristics of the CubeSat. Specifically, it includes the geometric properties, center-of-mass configuration, inertial characteristics, and aerodynamic features of the nanosatellite. Additionally, the simulation requires the altitude of the CubeSat during flight and the initial conditions of its angular motion. Collectively, these parameters are crucial for accurately modeling the angular dynamics of the nanosatellite and facilitating a comprehensive understanding of its behavior in space.

The leftmost panel, labeled "CubeSat Parameters", allows users to define the physical characteristics of the CubeSat. These include:

- Units: The tool supports various CubeSat form factor, such as 1U up to 12U. Users can select the required size from a dropdown menu.
- Mass: The mass of the CubeSat in kilograms.
- Moments of inertia: For determining the resistance to rotational motion users input the longitudinal moment of inertia I_x and the transversal moment of inertia I, both in kg·m².

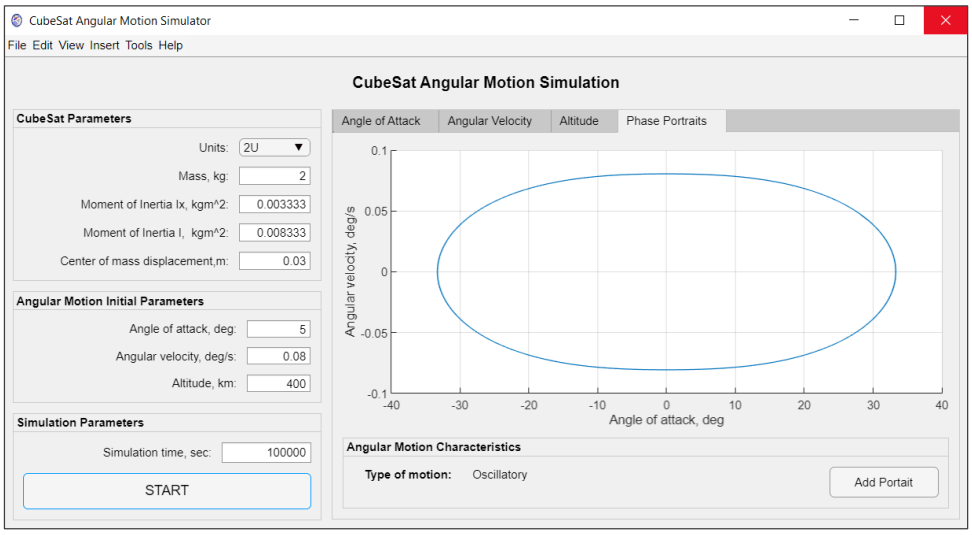


Fig. 1. Main interface of the CubeSat angular motion software.

- Center of mass displacement: A key feature of the tool is the ability to account for displacement of the center of mass along the longitudinal axis, which is input in meters.

By adjusting these inputs, users can simulate various CubeSat configurations, from simple single unit designs to more complex with multiple units.

Below the CubeSat Parameters section, users define the initial conditions of angular motion in the "Angular Motion Initial Parameters" panel.

- Angle of attack α: This parameter, expressed in degrees, defines the initial angle between the CubeSat longitudinal axis and the flight velocity vector.
- Angular velocity $\dot{\alpha}$: The initial angular velocity, given in degrees per second, specifies the rate of rotation of the angle of attack.
- Altitude: The orbital altitude of the CubeSat in kilometers.

The Simulation Parameters panel allows users to specify the total simulation time in seconds, enabling them to explore short-term or long-term behavior of the angular motion.

C. Phase Portraits Section

One of the features of the tool is the capability to generate phase portraits. A phase portrait represents the trajectory of the angular motion in phase space, where the angle of attack is plotted against angular velocity. This visual representation allows users to easily identify whether the motion is periodic, stable, or exhibits any chaotic tendencies. The phase portraits tab provides users with the option to save and compare multiple simulations, offering an understanding of how variations in initial conditions or CubeSat parameters impact the system. By selecting the "Add Portrait" button, users can overlay results from different simulation runs, facilitating comparative analysis.

For example, the phase portrait presented in Fig. 2 illustrates an analysis of angular motion in a planar representation, generated using the software tool. This example incorporates five distinct types of motion added using the "Add Portrait" button. The plot offers an insightful look into how different types of motion, rotatory and oscillatory, can be represented, along with the critical role of the separatrix in dividing the regions of possible motion.

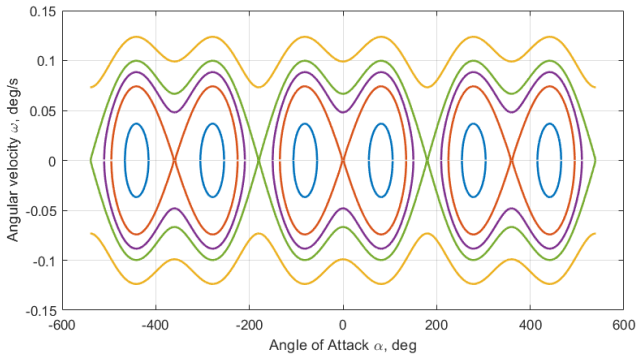


Fig. 2. Opportunity for study of phase portraits.

The phase portrait maps the relationship between the angle of attack α on the x-axis, and the angular velocity ω on the y-axis. Each trajectory corresponds to a specific type of motion, representing how the angular velocity changes as the angle of attack evolves.

1) Rotatory motion: Represented by the yellow curve, the rotatory motion is characterized by the CubeSat spinning continuously, with no return to its original orientation.

2) Oscillatory motion: In contrast, oscillatory motion is represented by closed trajectories (blue and purple curves). This type of motion implies that the angular velocity periodically reverses direction, leading to a back-and-forth movement around a stable point or equilibrium. Physically, this corresponds to a situation where the CubeSat is wobbling or oscillating without completing a full rotation.

3) Separatrix: The green and red curves are separatrices that delineate the boundary between oscillatory and rotatory

motion. This curve indicates the critical point at which a small change in angular velocity or angle of attack can result in a significant shift from oscillatory to rotatory motion.

A more detailed analysis of determination of equilibrium positions for CubeSats can be found in the work [23].

Finally, below the graph, the software provides a real-time classification of the angular motion of the CubeSat under the label "Angular Motion Characteristics". Depending on the initial conditions and the physical parameters, the tool identifies the motion as oscillatory or rotary.

D. Data Output and Visualization

Once the parameters described in section B are set, the user can start the simulation by pressing the START button. The software tool provides a range of outputs that allow users to analyze the CubeSat dynamics over time. As an illustrative example, a simulation was conducted using the parameters presented in Tables I and II, with a simulation time of 100,000 seconds.

Table I details the design characteristics for a CubeSat 3U, while Table II outlines the initial angular motion parameters used for the simulation.

TABLE I. CUBESAT SIMULATION PARAMETERS

Parameter	Value
Mass, kg	3
Longitude, m	0.3
Moment of Inertia I_x, kg·m²	0.005
Moment of Inertia I, kg·m²	0.025
Center of mass displacement Δx, m	0.03

TABLE II. INITIAL ANGULAR MOTION PARAMETERS

Parameter	Value
Angle of attack, deg	45
Angular velocity, deg/s	0.15
Altitude, km	400

Fig. 3 shows the variation of the angle of attack over time, where a clear oscillatory behavior is observed. This periodic oscillation suggests that the orientation of the CubeSat relative to its orbit changes continuously due to perturbation torques.

Analogically, Fig. 4 presents the variation of the angular velocity over time.

The phase portrait shown in Fig. 5 plots angular velocity against the angle of attack. The curve represents oscillatory motion around the equilibrium position, where $\alpha = 0$.

Finally, Fig. 6 depicts the variation in flight altitude over time for different CubeSat with form factor 2U and 3U. This plot demonstrates the gradual decay in altitude, which is typical for satellites in low Earth orbit due to atmospheric drag. Also, this graphic demonstrates that the CubeSat 2U exhibits a faster loss of altitude compared to the CubeSat 3U.

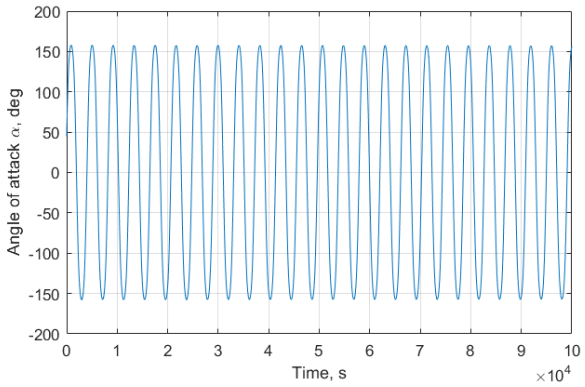


Fig. 3. Variation of the angle of attack during simulation.

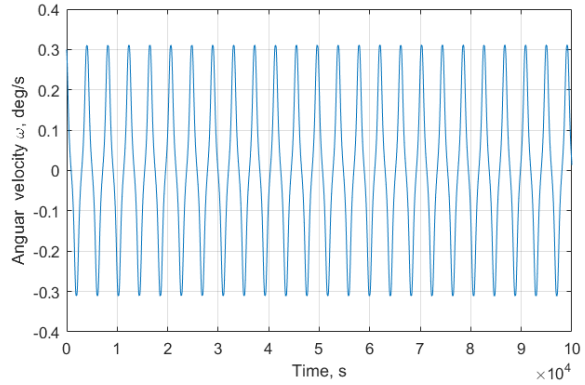


Fig. 4. Variation of the angular velocity during simulation.

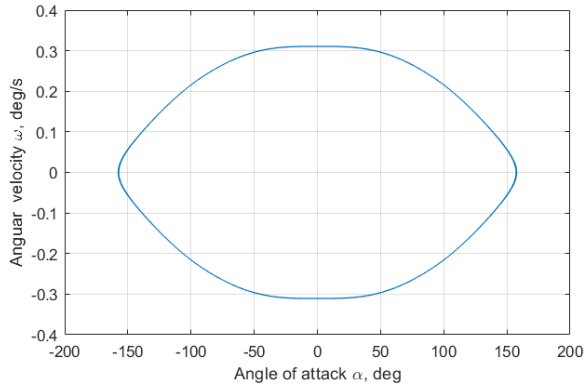


Fig. 5. Phase portrait during simulation.

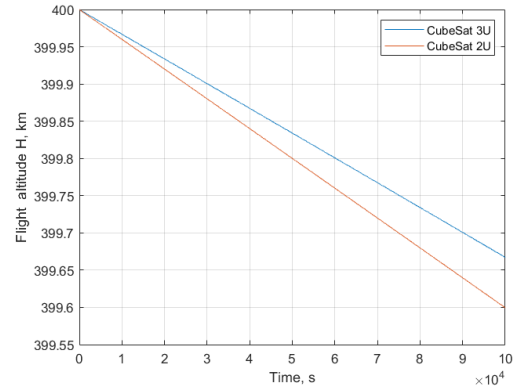


Fig. 6. Flight altitude during simulation for different CubeSats.

E. Validation and Verification

The validation and verification of the developed software tool were conducted through a combination of analytical benchmarking, sensitivity analysis, and software testing methodologies to ensure both numerical accuracy and computational reliability.

To validate the accuracy of the numerical simulations, benchmark comparisons were performed against analytical solutions derived from rigid-body dynamics. The presented mathematical models for angular motion serve as the theoretical reference for evaluating the correctness of the implemented models. For specific cases where closed-form solutions exist, numerical integration results were shown to closely replicate the expected angular velocity trajectories and phase portraits. Additionally, limiting cases were analyzed, including scenarios where aerodynamic torques were negligible, allowing direct verification against classical torque-free rotational motion solutions. The capacity of the software to capture equilibrium conditions and the expected transitions between oscillatory and rotatory regimes further reinforced the validity of the underlying mathematical framework.

Beyond validation with theoretical models, a sensitivity analysis was conducted to assess the robustness of the software in response to variations in physical parameters. Perturbations in the transverse and longitudinal moments of inertia were introduced to evaluate their effects on rotational stability, confirming that small deviations in these properties resulted in expected dynamic shifts consistent with established CubeSat dynamics theory. The center-of-mass displacement along the longitudinal axis was systematically varied to examine its impact on aerodynamic torques, demonstrating the ability to predict stability conditions influenced by structural asymmetry. Additionally, the effects on altitude were examined, confirming that at lower altitudes, increased atmospheric drag induced more pronounced oscillatory damping, while at higher altitudes, gravitational torques became the dominant perturbation force. These results closely aligned with theoretical expectations and prior research on CubeSat aerodynamics.

In addition to physical validation, software verification techniques were applied to ensure the integrity of the computational implementation. A unit testing framework was integrated within the MATLAB environment, employing modular tests for key computational functions, including numerical integration routines, moment calculations, and phase portrait generation. Each component was individually validated against known analytical solutions or benchmark datasets to detect discrepancies. Furthermore, consistency checks were performed by systematically varying input parameters and ensuring that the outputs adhered to expected trends. Edge-case testing was also conducted to evaluate system behavior under extreme parameter values, ensuring numerical stability and robustness.

V. Discussion

The results obtained from the developed CubeSat Angular Motion Simulator demonstrate its effectiveness in visualizing and analyzing the angular motion dynamics of CubeSats in low Earth orbit. The phase portraits generated by the tool confirm the expected motion behaviors, distinguishing between oscillatory and rotatory regimes based on initial conditions. The clear identification of separatrices highlights the critical transition points between these regimes, providing valuable insights for both educational and preliminary design applications.

From an educational standpoint, the simulator bridges the gap between theoretical learning and practical application, raising critical questions about how simulation-based learning affects conceptual understanding of real-world physical phenomena. The reliance on digital models may shape how students perceive and interact with physical reality, potentially impacting their approach to problem-solving and experimental validation in aerospace engineering. By providing real-time visualization and interactive parameter adjustments, the tool enhances comprehension in aerospace engineering curricula, where abstract mathematical models often present significant learning challenges.

Despite its strengths, the current version of the tool has certain limitations. The assumption of planar angular motion simplifies the analysis but excludes three-dimensional effects, which may become significant in more complex mission scenarios. Additionally, the exclusion of external perturbations such as geomagnetic torques and solar radiation pressure limits the applicability of the model for higher-fidelity mission simulations. Future work should focus on extending the model to include these effects, as well as incorporating real CubeSat telemetry data for validation. Enhancing the tool with additional control system simulations could further increase its relevance for practical mission planning and research on space dynamics.

VI. Conclusion

This work introduced an interactive software tool designed to simulate and visualize CubeSat angular motion, providing an accessible platform for exploring the principles of rotational dynamics. The features incorporated into the tool ensure computational efficiency while maintaining pedagogical value, making the tool ideal for both educational settings and preliminary CubeSat design evaluations.

Thus, this study contributes to the growing body of educational tools designed to make knowledge about space technology more accessible and comprehensible for the next generation of space professionals. The presented approach not only serves as an educational resource but also as a means for CubeSat developers to better assess attitude control strategies in early design and mission planning stages.

References

[1] A. Toorian, K. Diaz, and S. Lee, “The CubeSat Approach to Space Access,” 2008 IEEE Aerospace Conference. IEEE, Mar. 2008.

[2] A. Poghosyan and A. Golkar, “CubeSat evolution: Analyzing CubeSat capabilities for conducting science missions,” Progress in Aerospace Sciences, vol. 88. Elsevier BV, pp. 59–83, Jan. 2017.

[3] D. Selva and D. Krejci, “A survey and assessment of the capabilities of CubeSats for Earth observation,” Acta Astronautica, vol. 74. Elsevier BV, pp. 50–68, May 2012.

[4] N. Saeed, A. Elzanaty, H. Almorad, H. Dahrouj, T. Y. Al-Naffouri, and M.-S. Alouini, “CubeSat Communications: Recent Advances and Future Challenges,” IEEE Communications Surveys & Tutorials, vol. 22, no. 3. Institute of Electrical and Electronics Engineers (IEEE), pp. 1839–1862, 2020.

[5] G. Benedetti et al., "Interplanetary CubeSats for asteroid exploration: Mission analysis and design," Acta Astronautica, vol. 154. Elsevier BV, pp. 238–255, Jan. 2019.

[6] G. P. Candini, F. Piergentili, and F. Santoni, "Miniaturized attitude control system for nanosatellites," Acta Astronautica, vol. 81, no. 1. Elsevier BV, pp. 325–334, Dec. 2012.

[7] M. Ovchinnikov, V. Pen'ko, O. Norberg, and S. Barabash, "Attitude control system for the first swedish nanosatellite 'MUNIN,'" Acta Astronautica, vol. 46, no. 2–6. Elsevier BV, pp. 319–326, Jan. 2000.

[8] I. V. Belokonov, I. A. Timbai, and P. N. Nikolaev, "Analysis and Synthesis of Motion of Aerodynamically Stabilized Nanosatellites of the CubeSat Design," Gyroscopy and Navigation, vol. 9, no. 4. Pleiades Publishing Ltd, pp. 287–300, Oct. 2018.

[9] N. Botha and S. Marais, "Development of a Graphical User Interface as a Learning Tool for Artificial Intelligence," 2021 Rapid Product Development Association of South Africa - Robotics and Mechatronics - Pattern Recognition Association of South Africa (RAPDASA-RobMech-PRASA). IEEE, pp. 1–6, Nov. 03, 2021.

[10] M. Mohd Ali and N. Hashim, "Development of deep learning based user-friendly interface for fruit quality detection," Journal of Food Engineering, vol. 380. Elsevier BV, p. 112165, Nov. 2024.

[11] F. A. Chiappini, H. C. Goicoechea, and A. C. Olivieri, "MVC1_GUI: A MATLAB graphical user interface for first-order multivariate calibration. An upgrade including artificial neural networks modelling," Chemometrics and Intelligent Laboratory Systems, vol. 206. Elsevier BV, p. 104162, Nov. 2020.

[12] M. Gasparic, A. Janes, F. Ricci, G. C. Murphy, and T. Gurbanov, "A graphical user interface for presenting integrated development environment command recommendations: Design, evaluation, and implementation," Information and Software Technology, vol. 92. Elsevier BV, pp. 236–255, Dec. 2017.

[13] P. Mishra et al., "MBA-GUI: A chemometric graphical user interface for multi-block data visualisation, regression, classification, variable selection and automated pre-processing," Chemometrics and Intelligent Laboratory Systems, vol. 205. Elsevier BV, p. 104139, Oct. 2020.

[14] M. Victoria, O. M. Querin, C. Díaz, and P. Martí, "liteITD a MATLAB Graphical User Interface (GUI) program for topology design of continuum structures," Advances in Engineering Software, vol. 100. Elsevier BV, pp. 126–147, Oct. 2016.

[15] Y. Errazzouki, A. Habib, A. Jabiri, M. Sabil, and Z. Benkhaldoun, "Developing MATLAB graphical user interface for acquiring single star SCIDAR data," Astronomy and Computing, vol. 49. Elsevier BV, p. 100878, Oct. 2024.

[16] R. F. Hadi and Dr. Rianto Adhy Sasongko, "CubeSat Dynamic Design and Visualization Tools in MatLab ® Environment," Unpublished, 2014.

[17] D. S. Ivanov et al., "Software Package for Simulating the Angular and Orbital Motion of a Satellite," Mathematical Models and Computer Simulations, vol. 12, no. 4. Pleiades Publishing Ltd, pp. 561–568, Jul. 2020.

[18] K. A. Ezzat, L. N. Mahdy, A. E. Hassanien, and A. Darwish, "Robust Simulation and Visualization of Satellite Orbit Tracking System," Advances in Intelligent Systems and Computing. Springer International Publishing, pp. 505–514, Aug. 29, 2018.

[19] Andrew J. Turner, "The Development and Use of Open-source Spacecraft Simulation and Control Software for Education and Research," 2nd IEEE International Conference on Space Mission Challenges for Information Technology (SMC-IT'06). IEEE, pp. 330–336, July 2006.

[20] M. G. Shirobokov and S. P. Trofimov, "KIAM Astrodynamics Toolbox for Spacecraft Orbital Motion Design," Programming and Computer Software, vol. 50, no. 1. Pleiades Publishing Ltd, pp. 42–52, Feb. 2024.

[21] I. V. Belokonov, I. A. Timbai, and E. V. Barinova, "Design Parameters Selection for CubeSat Nanosatellite with a Passive Stabilization System," Gyroscopy and Navigation, vol. 11, no. 2. Pleiades Publishing Ltd, pp. 149–161, Apr. 2020.

[22] Hicks, K. D. Introduction to AstrodynamicReentry. Books Express Publishing, 2009.

[23] E. V. Barinova and I. A. Timbai, "Determining of Equilibrium Positions of CubeSat Nanosatellite under the Influence of Aerodynamic and Gravitational Moments," 2020 27th Saint Petersburg International Conference on Integrated Navigation Systems (ICINS). IEEE, pp. 1–4, May 2020.

Performance Evaluation and Selection of Appropriate Congestion Control Algorithms for MPT Networks

Naseer Al-Imareen, Gábor Lencse
Department of Telecommunications, Széchenyi István University, Győr, Hungary

***Abstract*—Recent academic research highlights a growing interest in multipath technologies, which offer promising solutions to networking challenges in complex environments. This interest is reflected in the emergence of protocols such as Multipath TCP (MPTCP) and Multipath UDP-in-GRE (MPT-GRE). The development of network protocols, particularly various iterations of the Transmission Control Protocol (TCP), has been distinguished by congestion detection and control algorithms, such as HighSpeed, CUBIC, Reno, LP, BBR, and Illinois. This paper evaluates the performance and suitability of these algorithms for multipath MPT-GRE networks under varying conditions, including delay, jitter, and data loss at different transmission speeds (both symmetric and asymmetric). Using StarBED resources, we applied delay, jitter, or packet loss to one of two physical paths to simulate congestion. The results demonstrate that some algorithms, HighSpeed and BBR among them, significantly enhance Quality of Service (QoS) metrics and network throughput in multipath MPT-GRE networks. These findings provide valuable insights into their performance and practical applications.**

***Keywords*—Packet loss; congestion control; MPT-GRE; delay; throughput; jitter**

I. Introduction

The rapid advancement of network applications has increased demands on network infrastructure, posing challenges to its capacity and efficiency. Many emerging applications require high bandwidth and low latency to function effectively. These requirements strain current network capabilities, exacerbating bottlenecks and highlighting the need for robust, efficient data transmission methods [1] [2].

Modern communication technology supports a variety of devices equipped with multiple interfaces, enabling networks and applications to handle complex communication demands. However, the effectiveness of these communication sessions is constrained by the TCP/IP protocol architecture, which, by default, supports only single-session handling. Leveraging multiple network interfaces during communication sessions enhances flexibility and reliability, particularly in addressing network disruptions. By dynamically switching traffic between available paths, communication systems can ensure uninterrupted data transmission, even in the face of link failures, congestion, or performance degradation [3].

This approach significantly improves fault tolerance, load balancing, and network stability, making it an essential feature for modern networking environments that demand high reliability and performance.

To address the growing demand for efficient multipath solutions, numerous methods have been developed, including MPT-GRE [4] and MPTCP [5]. MPT-GRE enables the creation of a virtual tunnel across multiple physical paths, distinguishing it from alternatives like MPTCP and Huawei's Generic Routing Encapsulation (GRE) Tunnel Bonding Protocol. While multipath approaches offer significant advantages, throughput performance in these networks often suffers due to delays and congestion [6].

Various TCP congestion control algorithms, such as HighSpeed, LP, Reno, Vegas, and CUBIC, have been designed to mitigate these challenges. These algorithms detect, control, and preempt congestion, reducing packet loss and delays. Their effectiveness stems from their ability to monitor and manage packet transmission from source to destination. Some algorithms dynamically adjust the congestion window size based on round-trip time (RTT), while others, particularly those optimized for high-bandwidth networks, expand the window to enhance throughput.

Efficient congestion management in multipath networks improves stability, ensures proper packet reordering, and minimizes data loss and delays. Furthermore, the fair allocation of network resources among competing packets prevents bandwidth monopolization, safeguarding throughput and overall network performance [7] [8].

Multipath congestion control is an active area of research focused on maximizing resource utilization by leveraging the available bandwidth across multiple paths while maintaining fairness toward competitive single-path transfers—a constraint referred to as TCP-friendliness. Congestion control techniques are crucial in optimizing network resource use, enabling throughput aggregation, and reducing bandwidth waste.

However, the rise of multipath communication has introduced new challenges. Research has identified side effects, particularly the lack of TCP-friendliness in some implementations. For example, the uncoupled congestion control approach in Multipath TCP (MPTCP) treats each sub flow as an independent TCP connection. This can result in imbalanced resource allocation, where individual subflows dominate bandwidth, leading to unfairness and performance degradation in multipath network environments [8][9].

The MPT-GRE software is designed to enhance data transmission by distributing the load across multiple paths, often achieving throughput capacities close to the combined total of the physical paths [10]. The integration of congestion control algorithms in multipath networks holds significant promise for building robust and efficient network infrastructures. These

algorithms ensure fair resource allocation and improve overall network performance by enhancing fault tolerance, reducing congestion, and increasing throughput through effective traffic management across multiple paths.

Our contributions to this paper are as follows:

1) Evaluation of congestion control algorithms: We analyze multiple congestion control algorithms within multipath MPT-GRE network environments.

2) Performance assessment under diverse network conditions: We assess the performance of these algorithms under various network conditions, including delay, jitter, combined delay and jitter, and packet loss. The evaluation also considers both symmetric and asymmetric transmission speed environments.

3) Identification of throughput-optimizing algorithms: We identify algorithms that significantly enhance network throughput under the evaluated network conditions.

II. Related Work

Many studies have explored congestion control approaches in multipath networks, each focusing on optimizing network performance in various ways. These studies have examined how congestion control algorithms can effectively handle multiple paths, reduce packet loss, minimize latency, and ensure fair resource allocation among packet flows. Additionally, researchers have worked on utilizing algorithms that can adapt dynamically to changing network conditions, such as fluctuations in bandwidth, jitter, and delay, to maximize throughput and minimize congestion.

Szabolcs Szilágyi and Imre Bordán [11] examined the impact of various TCP congestion control algorithms on multipath communication technologies, specifically MPTCP and MPT-GRE. The researchers compared the performance of seven congestion control algorithms (CUBIC, Reno, Illinois, Scalable, Veno, High-Speed, and Vegas) in quad-path IPv4/IPv6 Fast Ethernet environments. Their findings show that CUBIC provided the best performance for MPTCP and MPT-GRE, while Vegas had the lowest performance. The study used these comparisons to emphasize CUBIC's effectiveness as the default algorithm in modern operating systems and aimed to extend the evaluation to more advanced network environments and recent TCP algorithms.

To address the energy consumption challenge in Multipath TCP (MPTCP), the authors [3] analyzed existing congestion control algorithms and identified the key factors influencing energy efficiency. They conducted real-world experiments using the MPTCP Linux kernel and found that energy consumption is closely related to throughput, path delay, and varying network scenarios. To improve energy efficiency, they proposed a congestion control model with a window-increasing factor to direct traffic toward low-delay paths and an energy-aware compensatory parameter for hierarchical Internet topologies. Their experiments confirmed that the enhanced model can increase energy efficiency without compromising transmission performance.

Yu Cao et al. [12] addressed the limitations of coarse-grained load balancing in multipath congestion control, which relies heavily on packet loss as a congestion indicator. They formulated the "Congestion Equality Principle," showing that fair and efficient traffic shifting occurs when flows equalize perceived congestion across all paths. To achieve this, they proposed the delay-based algorithm Weighted Vegas (wVegas), which uses queuing delays for fine-grained load balancing. Simulations showed that wVegas responds faster to congestion changes than loss-based algorithms, improving intra-protocol fairness and reducing packet loss. The study highlights wVegas as a complement to algorithms like TCP-Vegas and TCP-Reno.

Balancing fairness, responsiveness, and window oscillation in Multipath TCP (MPTCP) congestion control is crucial for efficient multipath communication. MPTCP distributes traffic across multiple paths to enhance resource utilization and connection robustness. However, this distribution poses the challenge of adjusting transmission rates across these paths without disrupting other network traffic. To address this, researchers [13] proposed a novel fairness-based congestion control algorithm (FCCA) designed to enhance fairness among subflows while maintaining key performance metrics such as responsiveness and stability of the congestion window. FCCA dynamically adjusts the congestion window for each path based on real-time congestion feedback, optimizing bandwidth usage, improving network performance, and ensuring smooth traffic flow even under varying congestion conditions. The introduction of FCCA marks a significant step toward achieving more equitable and efficient traffic management in MPTCP, contributing to enhanced network performance and fairness in multipath communication scenarios.

III. Background

This section discusses the MPT-GRE multipath network and congestion control algorithms in detail.

A. MPT-GRE Network Technology

The MPT-GRE network is a multipath technology based on the GRE-in-UDP tunnel specification (IETF RFC 8086 [2]). MPT-GRE extends the traditional GRE-in-UDP architecture by supporting multiple physical paths and enhancing network performance through load balancing. Using the UDP source port for hashing distributes traffic more efficiently across numerous equal-cost multipath (ECMP) routes. This architecture shares some similarities with Multipath TCP (MPTCP) in its utilization of multiple paths. Still, it differs significantly in its underlying technology. MPT-GRE relies on UDP at the transport layer, building on GRE-in-UDP for a tunnel IP layer that supports TCP and UDP protocols. Huawei's GRE Tunnel Bonding Protocol has a similar objective but lacks UDP encapsulation, limiting its scalability to just two physical interfaces. In contrast, MPT-GRE's use of GRE-in-UDP offers greater flexibility, enabling a more scalable and robust multipath solution.

The MPT-GRE software architecture, shown in Fig. 1, introduces a logical tunnel layer that works independently of the physical network paths. This unique approach sets MPT apart from traditional TCP/IP protocols by directing application layer data to a tunnel path instead of directly to a physical path. The MPT-GRE software distributes incoming packets from the

logical tunnel interface across available physical paths in this environment. It allows for seamless multipath communication for applications, which only need to interact with the tunnel interface. This mechanism enables efficient traffic redistribution without changing the application's communication model. Additionally, the MPT-GRE software supports the transition to IPv6, as both IP tunnel and path versions can operate independently within the MPT-GRE library. Fig. 2 shows how data packets are transmitted and received in MPT-GRE.

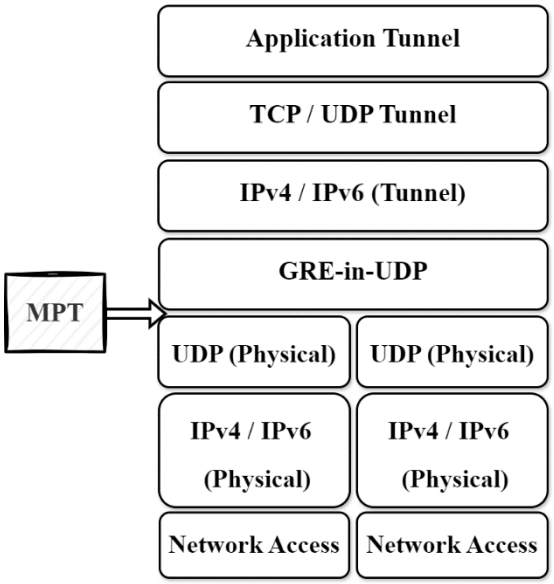


Fig. 1. Conceptual architecture of MPT-GRE [4].

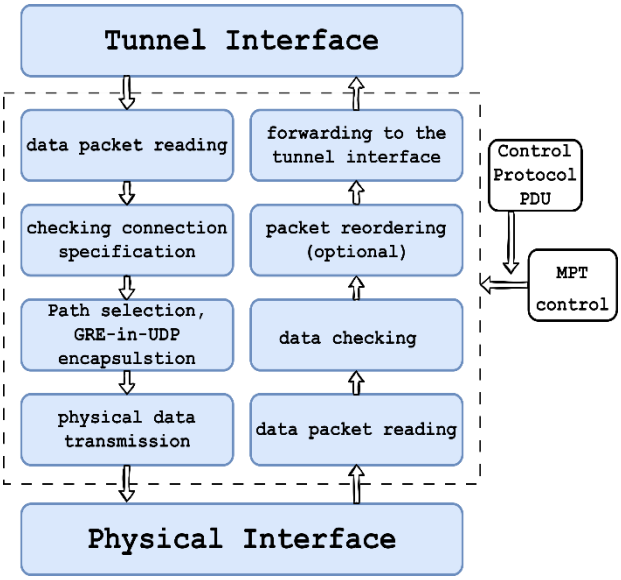


Fig. 2. Theoretical process of the MPT-GRE mechanism [4].

Upon receiving a packet from the tunnel interface, the MPT-GRE software identifies the connection specification and determines traffic distribution across multiple paths. The user data packet is then encapsulated into a GRE-in-UDP data unit. This unit may include optional GRE Sequence Numbers for reordering. The GRE header contains 16 bits of zeros and identifies the protocol type (e.g., 0x0800 for IPv4 or 0x86DD for IPv6). The GRE-in-UDP data unit is encapsulated within a UDP/IP data unit with the destination port set to 4754, which is the GRE-in-UDP port. The packet is then transmitted via the designated physical interface. Upon arrival, the MPT-GRE software verifies the packet by checking the destination port, validating the connection based on the tunnel IP header, and performing checks on the GRE sequence number or GRE Key value, if present. If the packet passes all checks without reordering, it is forwarded to the transport and application layers via the tunnel interface. If reordering is necessary, the packet is temporarily stored in a buffer for reordering before being transmitted [14].

B. Congestion Control Algorithms

Congestion control in computer networks is crucial for ensuring efficient data transfer and preventing issues such as increased latency, packet loss, and reduced throughput [15]. When network demand exceeds available bandwidth or specific data flows dominate resources, congestion control algorithms are essential for maintaining network stability. These algorithms handle transmission data by dynamically adjusting transmission rates and implementing mechanisms like slow start and fast retransmission, which help to alleviate congestion and reduce packet loss. They are designed to respond to dynamic and unpredictable network traffic, employing different strategies to manage congestion by monitoring packet flows and adjusting sending rates accordingly [11][16]. Widely used congestion control algorithms, such as LP, Veno, H-TCP, CUBIC, Reno, Hybla, Vegas, HSTCP, BBR, Westwood, BIC, and Scalable, adjust the rate of transmission to adapt to varying network structures and conditions. For example, CUBIC and Vegas monitor estimated round-trip times (RTT) to detect congestion and dynamically adjust transmission rates, while other algorithms like BBR estimate bandwidth to optimize throughput [17]. These algorithms are generally classified based on their method of detecting and responding to congestion as the following: loss-based, delay-based, hybrid (loss + delay), and bandwidth estimation-based algorithms [18][19][20][21].

1) Loss-based algorithms: Loss-based congestion control algorithms detect network congestion by identifying packet loss. Typically, this occurs when network buffers overflow due to congestion. When packet loss is detected, the sender reduces the congestion window, which is the amount of data allowed in transit without acknowledgment. The sender then gradually increases the congestion window to probe the available bandwidth. These algorithms use packet loss as a congestion signal, detected through duplicate acknowledgments or timeouts. Upon congestion detection, the sender reduces the congestion window, often by half, and slowly increases it using strategies like additive increase multiplicative decrease (AIMD). Examples of loss-based algorithms include Reno, shown in Eq. (1) and (2), a traditional algorithm that reduces the congestion window after detecting packet loss, and CUBIC, shown in Eq. (3), which employs a cubic function to manage window growth but still relies on packet loss for congestion detection.

Reno equations:

$$cwnd = cwnd + \frac{1}{cwnd}\ (\ Additive\ Increase\ ) \quad (1)$$

$$cwnd = \frac{cwnd}{2}\ (\ Multiplicative\ Decrease\ ) \quad (2)$$

CUBIC equation:

$$cwnd(t) = C \cdot (t - K)^3 + W_{max} \quad (3)$$

where C is a scaling element that holds the window growth rate, and t is the time since the last congestion event (loss), whereas $K = \sqrt[3]{\frac{W_{max} \cdot \beta}{C}}$ is the time when the window moves W_{max} again. And W_{max} is the congestion window size at the last congestion event (maximum window size before packet loss). β represents a multiplicative decrease factor, usually set at 0.7.

HighSpeed, Binary Increase Congestion Control (BIC), Scalable, and Hamilton TCP (H-TCP) are advanced loss-based congestion control algorithms designed specifically for high-speed, high-latency networks. HighSpeed aggressively increases the standard congestion window, enhancing throughput in environments with a large bandwidth-delay product (BDP) shown in Eq. (4). BIC uses a binary search method for adjusting the window size, effectively balancing rapid bandwidth probing with a cautious approach to packet loss, as shown in Eq. (5). Scalable adopts a fixed-increment growth strategy based on the current window size, allowing it to maintain high throughput in high-capacity networks, as shown in Eq. (6). On the other hand, H-TCP dynamically adjusts its window growth by monitoring changes in network congestion and round-trip time (RTT), enabling it to respond effectively to fluctuations in network conditions and improve performance in high-BDP scenarios shown in Eq. (7).

HighSpeed equation:

$$cwnd = cwnd + \frac{a(cwnd)}{cwnd} \quad (4)$$

where a ($cwnd$) is an adaptive increase factor that scales with the size of $cwnd$. This factor becomes more significant as $cwnd$ grows.

BIC equation:

$$cwnd = \frac{(cwnd_{max} + cwnd_{min})}{2} \quad (5)$$

where $cwnd_{\max}$ and $cwnd_{\min}$ represent the maximum and minimum congestion window sizes that have been reached so far.

Scalable equation:

$$cwnd = cwnd + a.cwnd \quad (6)$$

where α is a constant scaling factor typically set to a small value, such as 0.01 to 0.1.

H-TCP Equation:

$$cwnd = cwnd + (2 \times (t - T)) \quad (7)$$

where t is the elapsed time since the last congestion event, and T is a constant.

2) Delay-based algorithms: Delay-based algorithms monitor round-trip time (RTT) changes or delays to detect network congestion, allowing them to adjust the sending rate before packet loss occurs. For example, Vegas continuously monitors RTT to adjust the congestion window size based on observed delays, as shown in Equation 8. Hybla compensates for longer RTTs by scaling window growth proportionally to RTT values, improving performance in high-latency networks, as shown in Eq. (9). On the other hand, LP (Low Priority) reduces its window size when delays increase, prioritizing higher-priority traffic and ensuring smooth operation in mixed-traffic environments, as shown in Eq. (10) and Eq. (11).

Vegas equations:

$$\Delta = \left(Expected\ Throughput - Actual\ Throughput\right) \quad (8)$$

$$cwnd = cwnd + \alpha, \quad if\ \Delta < \gamma$$
$$cwnd = cwnd - \beta, \quad if\ \Delta > \delta$$

where α is the increase factor. β is the decrease factor. Δ is the difference between expected and actual throughput. γ and δ are thresholds for adjusting the window.

Hybla equation:

$$\rho = \frac{RTT_{reference}}{RTT_{current}} \quad (9)$$

where $RTT_{reference}$ is a reference for round-trip time (RTT), usually for a low-latency network

LP equation:

$$\Delta = \frac{(RTT_{currnt} - RTT_{min})}{RTT_{min}} \quad (10)$$

If Δ > γ, LP reduces the congestion window as follows:

$$cwnd_{new} = \frac{cwnd_{current}}{2} \quad (11)$$

where the threshold for the increase in RTT is denoted by γ.

3) Hybrid (Loss + Delay) algorithms: These algorithms leverage packet loss and delay to adjust the congestion window. For example, Illinois uses delay and loss metrics to adjust the window size dynamically, enabling more adaptable congestion control. Veno combines Reno and Vegas's strategies, using delay and packet loss signals to optimize the balance between performance and congestion avoidance. Similarly, YEAH uses delay as an early indicator of congestion and relies on packet loss as a secondary, more conservative signal to adjust the transmission rate.

Illinois equation:

$$cwnd = cwnd + \frac{\alpha(delay)}{cwnd} \quad (12)$$

Veno equation:

$$cwnd_{new} = cwnd_{current} + \frac{1}{cwnd_{current}}\ (Add.Increase) \quad (13)$$

YEAH Equation:

$$\mathrm{cwnd}_{\mathrm{increase}} = \frac{\mathrm{MSS}}{\mathrm{RTT}_{\mathrm{min}}} \quad (14)$$

where, MSS stands for Maximum Segment Size, referring to the most significant amount (in bytes) of data that a TCP segment can carry in the payload part of a packet, excluding the TCP and IP headers.

4) Bandwidth estimation-based algorithms: These algorithms estimate the available bandwidth in the network and adjust the sending rate accordingly, optimizing performance based on network capacity. For example, TCP Westwood estimates bandwidth using packet loss and throughput measurements. It adjusts the congestion window dynamically, making it particularly effective in wireless networks where packet loss is common. Similarly, Bottleneck Bandwidth and Round-trip propagation time (BBR) models available bandwidth and round-trip delay to efficiently manage the congestion window and transmission rate.

Westwood equation:

$$\text{BWE} = \frac{\sum ACKed\_data}{\Delta t} \quad (15)$$

where *BWE* is the bandwidth estimate, *ACKed_data* is the amount of acknowledged data, and Δt is the time interval between receiving ACKs.

BBR equation:

$$\text{Throughput} = \frac{\text{Bottleneck Bandwidth}}{RTT_{min}} \quad (16)$$

where Bottleneck Bandwidth is the maximum rate at which data can be transmitted along the path.

IV. Test Environment

A. Hardware Environment Setting

As illustrated in Fig. 3, the test setup consisted of three Dell PowerEdge R650 servers, each equipped with the following:

- CPU: Intel Xeon Gold 6330N 2.2G, 28C/56T, 11.2GT/s, 42M cache, turbo, HT (165W) DDR4-2666 x 2.
- Memory: 32GB RDIMM, 3200MT/s, dual rank 16Gb base x 16 = 512GB.
- Storage: SSD 480GB SSD SATA Read Intensive 6Gbps 512, 2.5inch Hot Plug x 2 (RAID1).
- Network Interface: 25GigE×4.

This experimental design was carefully structured to evaluate the performance of the MPT-GRE multipath network, mainly focusing on the impact of congestion control algorithms on throughput and network efficiency. The experiments were conducted using Ubuntu 22.04.4 LTS (Jammy Jellyfish) / Linux operating system servers. These experiments were designed to evaluate the performance of the MPT-GRE tunnel throughput library and to monitor and prove the effectiveness of various congestion control algorithms.

B. MPT-GRE Configuration Setting

The version of MPT used in this experiment, *mpt-gre-lib64-2019.tar.gz*, is available publicly from GitHub [22]. Two primary configuration files within the MPT installation directory were modified to enable the effective operation of the MPT-GRE multipath system across the network infrastructure.

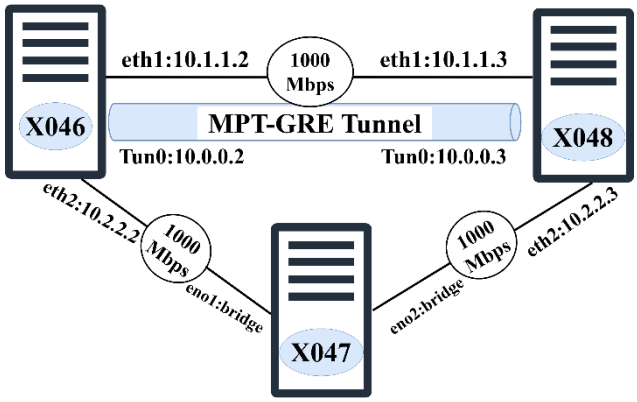


Fig. 3. Experimental topology of the MPT-GRE multipath network.

The first file, conf/interface.conf, contains essential parameters for network interfaces and tunnels. Adjustments were made to specify the Local Command UDP Port Number for managing communication between MPT-GRE and network interfaces; define the Interface Number, maximum transfer unit (MTU), and permissions for remote requests; and establish the tunnel interface name, IPv4 address, and subnet prefix. This file also manages tunnel traffic protocols, with the same configuration mirrored on the second server to maintain consistency in the multipath environment.

The second file, conf/connections/IPv4.conf, manages logical connections and defines transmission paths. Each MPT-GRE tunnel has its own connection file, specifying IP addresses for both tunnel endpoints and using IPv4 encapsulation within an IPv4 GRE tunnel. Previous publications [10][23] have provided detailed information on MPT-GRE configuration settings, and the updated configuration files for this study have also been made available on GitHub [24].

V. Experiments, Results, and Performance Analysis

The primary objective of this study is to confirm the influence of congestion control algorithms on MPT-GRE multipath networks and determine the optimal throughput aggregation value for the tunnel. A detailed evaluation of tunnel throughput is critical for enhancing our understanding of MPT-GRE operations.

Various scenarios were designed for the experiments to ensure precision and effectiveness. A Python script, available on GitHub [25], was employed to automate and facilitate the experimental process over a period of 30 seconds. This script not only enabled frequent execution but also automated the saving of tunnel throughput results, thereby enhancing consistency and repeatability. Additionally, a set of congestion control algorithms was integrated into the Python code and executed using iperf3 [26].

To evaluate the impact of various Quality of Service (QoS) metrics, we used a middle server to adjust traffic parameters with the Linux *tc* command, experimenting with delay, jitter, packet loss, and transmission speed limits. Each set of

congestion control algorithms was tested under various scenarios using the same QoS metrics.

In the first scenario, the transmission speeds were configured with eth1 set at 1000 Mbps, while eth2 varied incrementally from 100 Mbps to 1000 Mbps in steps of 100 Mbps. In the second scenario, similar network conditions were applied; the transmission speeds were reduced tenfold, with eth1 set to 100 Mbps and eth2 ranging from 10 Mbps to 100 Mbps in increments of 10 Mbps.

Due to the large number of results without loss of generality, the remaining transmission cases follow the same approach. The following cases are discussed in detail:

- The symmetrical case: *eth1 = 1000 Mbps and eth2 = 1000 Mbps* under the effect of all QoS metrics.
- The case of asymmetric paths: *eth1 = 100 Mbps* and *eth2 = 10, 20, 30... 100 Mbps* under the effect of delay and packet loss.

A. First Scenario: Analyzing QoS of Symmetrical Paths

1) Assessment impact of delay metric: We applied delays across both server network interfaces, ensuring bidirectional effects on transmitted packets. The delay parameter *x* was set at intervals of 10 ms, 20 ms, 30 ms, 40 ms, and 50 ms, using the *tc* command:

tc qdisc add dev eth1 root netem delay xms

For example, when symmetric paths (eth1=1000 Mbps and eth2=1000 Mbps) are used, the impact of the delay metric on loss-based algorithms, delay-based algorithms, hybrid algorithms, and bandwidth estimation-based algorithms is as is shown in Fig. 4. Analyzing throughput values across different congestion control algorithms under various delay conditions reveals key performance characteristics for each category.

The HighSpeed algorithm demonstrates the throughput achieving 1160 Mbps at 10 ms in the loss-based algorithms category. It sustains a relatively high performance of 202 Mbps even under moderate delays, making it particularly effective in environments with low to medium delays. The Scalable and CUBIC algorithms also perform well in this category, especially at lower delays; the Scalable algorithm reaches 531 Mbps at 20 ms. While CUBIC maintains competitive throughput as delay increases, it is less effective than the HighSpeed algorithm. The BIC algorithm offers balanced performance, achieving high throughput at low delays (1160 Mbps at 10 ms) and proving suitable in mixed delay environments. Therefore, HighSpeed and Scalable are the optimal loss-based algorithms for their peak performance and resilience under medium delays.

In the delay-based algorithms, at lower delay values (10 ms), Hybla achieves a higher throughput of 1180 Mbps compared to LP's 1080 Mbps, demonstrating better performance under minimal delay conditions. However, as the delay increases from 20 ms to 50 ms, LP outperforms Hybla, maintaining higher throughput at greater delays. For example, LP achieves 239 Mbps compared to Hybla's 178 Mbps at 40 ms. In contrast, Vegas experiences a steep drop in throughput as delay increases, making it less suitable for high-delay conditions.

Hence, while Hybla excels in low-delay situations, LP is the better option for higher-delay scenarios.

Within the hybrid (loss + delay) algorithms, Illinois stands out for its high throughput, reaching up to 1150 Mbps at 10 ms and demonstrating adaptability even at higher delays, achieving 215 Mbps at 40 ms. This makes it highly suitable for environments characterized by both loss and delay. Veno and YEAH exhibit medium performance but show sharper reductions in throughput as delay increases.

Therefore, Illinois is the most effective hybrid algorithm, excelling in mixed loss and delay conditions while consistently maintaining high throughput. In contrast, Veno and YEAH perform well in low-delay scenarios but underperform in higher-delay environments.

2) Assessment impact of jitter metric: We measured the throughput of the MPT-GRE tunnel under jitter values of 2 ms, 4 ms, 6 ms, 8 ms, and 10 ms while keeping the delay set to zero. This was done both with and without congestion control algorithms. To enable a more comprehensive assessment, we also combined delay with jitter using the following command:

tc qdisc add dev eth1 root netem delay dms jms

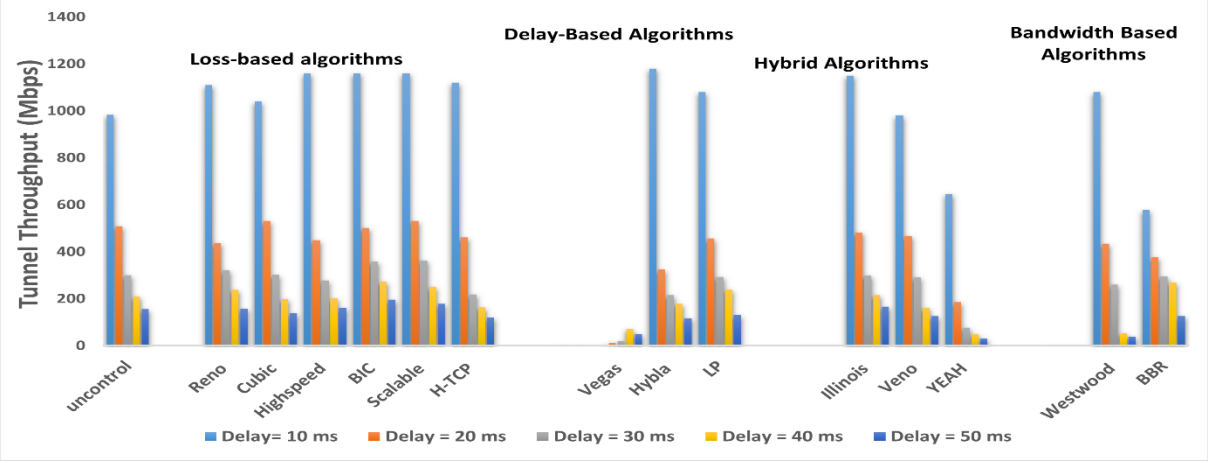


Fig. 4. Throughput performance of the MPT-GRE tunnel under varying delay metrics (eth1 = 1000 Mbps, eth2 = 1000 Mbps)/.

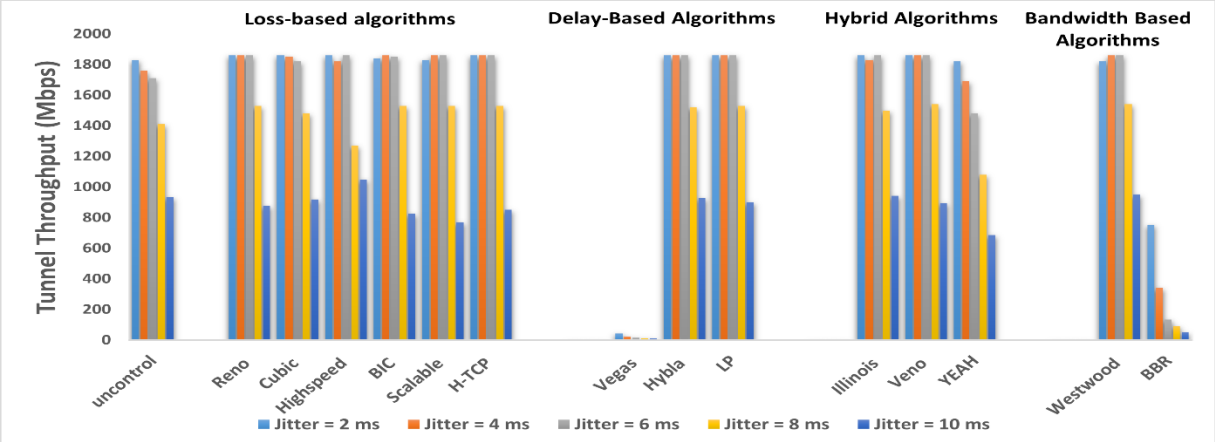


Fig. 5. Throughput performance of the MPT-GRE tunnel under varying jitter metrics (eth1 = 1000 Mbps, eth2 = 1000 Mbps).

The analysis presented in Fig. 5 illustrates the performance trends of various congestion control algorithms based on throughput at different jitter levels. Reno exhibits a consistent throughput of 1860 Mbps in the loss-based algorithms under low to moderate jitter conditions (ranging from 2 to 6 ms). However, its performance significantly declines to 1530 Mbps at 8 ms and drops further to 876 Mbps at 10 ms, indicating that it is sensitive to higher jitter levels. Other algorithms like CUBIC and H-TCP maintain relatively good throughput, showing reductions as jitter increases. Scalable and BIC perform well up to 8 ms, but they experience a significant decline in performance at 10 ms. In contrast, HighSpeed displays stable performance under low jitter levels and maintains a better throughput of 1270 Mbps at 8 ms and 1050 Mbps at 10 ms than other algorithms under higher jitter conditions.

HighSpeed's balanced performance makes it more suitable for environments with increased jitters, while Reno performs better in certain low-jitter situations.

In the delay-based algorithms, throughput remains high for Hybla and LP, maintaining a consistent 1860 Mbps under low to moderate jitter conditions (2–6 ms). However, as jitters increase to 8 ms and 10 ms, their throughput decreases, with Hybla reaching 1520 Mbps and 929 Mbps and LP achieving 1530 Mbps and 900 Mbps, respectively. In contrast, Vegas performs poorly across all jitter levels. Hybla and LP effectively manage moderate jitter, whereas Vegas is unsuitable for jitter-sensitive environments.

Of the hybrid (loss + delay) algorithms, Illinois and Veno show robust throughput despite jitter. Illinois performs slightly better at higher jitter levels, achieving 943 Mbps at 10 ms, while Veno reaches 895 Mbps. However, YEAH is less adaptable in high-jitter scenarios.

Finally, in the bandwidth estimation-based algorithms, there is a considerable difference in the performance of Westwood and BBR under jitter conditions. For example, at 10 ms of jitter, Westwood significantly outperforms BBR, achieving 949 Mbps compared to BBR's 51 Mbps.

Overall, some congestion control algorithms performed better than the Uncontrolled case (the performance baseline without applying any congestion control algorithm). For instance, the HighSpeed congestion control algorithm outperformed the Uncontrolled case in all scenarios.

To better understand the effects and identify the optimal congestion control algorithm, delay and jitter were applied simultaneously, with the jitter value set to 20% of the delay. Fig. 6 analyzes the performance of various congestion control algorithms compared to the Uncontrolled case. Several algorithms performed better than the Uncontrolled throughput, particularly under different delay and jitter conditions.

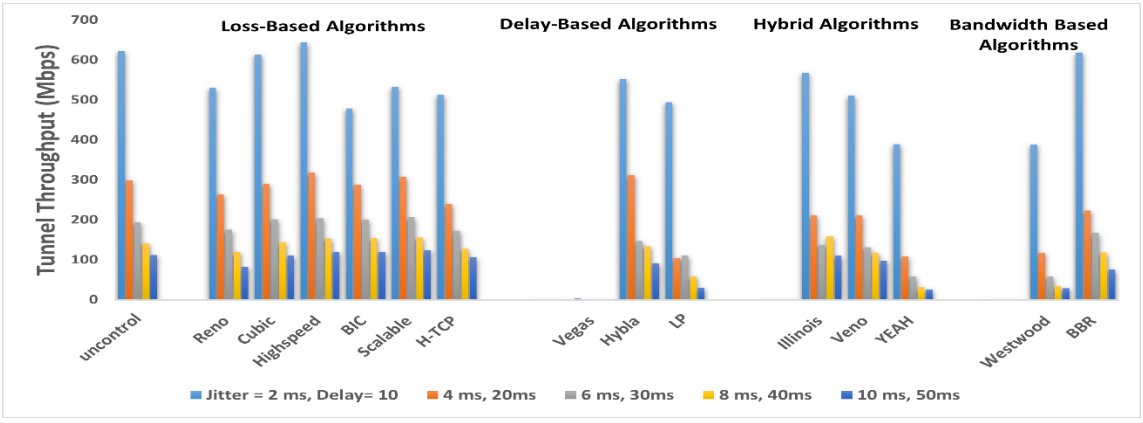


Fig. 6. Throughput performance of the MPT-GRE tunnel under combined delay and jitter metrics (eth1 = 1000 Mbps, eth2 = 1000 Mbps).

In comparison, Reno shows a more rapid and consistent decline in throughput, starting at 530 Mbps and dropping to 82.6 Mbps under higher jitter and delay, indicating a less adaptive response to congestion than the Uncontrolled case. The Uncontrolled case performs better than CUBIC and BIC under low delay and jitter conditions, specifically at 2 ms/10 ms and 4 ms/20 ms. However, as network conditions worsen, these algorithms outperform the Uncontrolled case.

Notably, the BIC algorithm excels under higher delay and jitter conditions, achieving throughput values of 200 Mbps, 154 Mbps, and 120 Mbps at 6 ms/30 ms, 8 ms/40 ms, and 10 ms/50 ms, respectively. Similarly, the Scalable algorithm outperforms the Uncontrolled case as delay and jitter increase, with throughput values of 308 Mbps, 206 Mbps, 156 Mbps, and 124 Mbps under jitter from 4 ms to 10 ms and delay from 20 ms to 50 ms. In contrast, the Uncontrolled case achieves 193 Mbps, 142 Mbps, and 112 Mbps under the same conditions, clearly showing lower throughput than Scalable and BIC as delay and jitter increase.

The HighSpeed algorithm significantly outperforms the Uncontrolled case and other algorithms, achieving 644 Mbps at jitter = 2 ms, delay = 10 ms, and maintaining a higher throughput of 120 Mbps even under the highest jitter and delay conditions. This reflects its superior congestion control capabilities.

Some of these algorithms, such as H-TCP, Hybla, and Illinois, achieved throughput levels close to those in the Uncontrolled case. In contrast, algorithms like Vegas, Westwood, YEAH, Veno, and LP consistently performed worse than the Uncontrolled values. Additionally, their performance declined as delay and jitter increased.

3) Assessment impact of packet loss metric: We tested packet loss conditions with loss rates set at 1%, 2%, 3%, 4%, and 5%. These rates were applied to network interfaces as follows:

tc qdisc add dev eth1 root netem loss x

Fig. 7 illustrates the impact of packet loss on the performance of various congestion control algorithms. Analyzing these results, it is evident that packet loss leads to a decline in throughput for all algorithms as the percentage of loss increases. This decrease is expected, as packet loss commonly results in retransmissions and delays, negatively affecting network throughput.

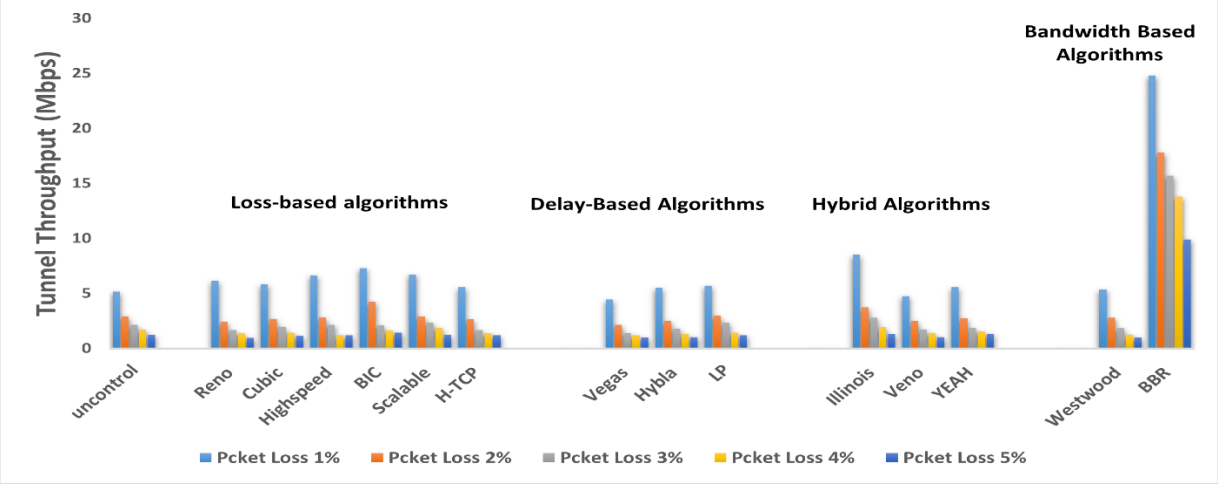


Fig. 7. Throughput performance of the MPT-GRE tunnel under varying packet loss metrics (eth1 = 1000 Mbps, eth2 = 1000 Mbps).

Among the congestion control algorithms examined, BBR is the most resilient to packet loss, consistently achieving significantly higher throughput across all levels of packet loss. Even at a packet loss rate of 5%, BBR maintains a throughput of 9.9 Mbps, far surpassing the performance of other algorithms. BBR's effectiveness minimizes the adverse effects of packet loss on throughput.

In contrast, the Uncontrolled case (i.e., without any congestion control algorithm applied) experiences substantial reductions in throughput as packet loss increases, with values dropping below 1.5 Mbps at 5% packet loss. These results highlight that congestion control algorithms significantly influence MPT-GRE tunnel throughput under varying packet loss conditions. While some algorithms perform better than others in the presence of packet loss, BBR stands out as the most promising option for maintaining higher throughput in challenging conditions.

B. *Second scenario: Analyzing QoS of Appropriate Algorithms*

In the second scenario, we analyzed:

- The HighSpeed congestion control algorithm compared to the Uncontrolled case on symmetric and asymmetric paths, where eth1 is set to 100 Mbps, and eth2 varies from 10 to 100 Mbps in increments of 10 Mbps under the delay metric.
- The BBR algorithm under the packet loss metric.

The results showed a significant increase in tunnel throughput under these specified delay and packet loss conditions. An analysis of the results, as demonstrated in Fig. 8, indicates that the HighSpeed algorithm achieved improved tunnel throughput. The MPT-GRE tunnel throughput performance between the HighSpeed algorithm and the

Uncontrolled case was assessed across symmetric and asymmetric transmission speeds and delay conditions. This analysis highlights how the HighSpeed algorithm enhances tunnel throughput in the MPT-GRE multipath network compared to the Uncontrolled case, particularly under increased network delays.

In the HighSpeed algorithm case, the MPT-GRE tunnel throughput shows a relatively modest decrease as delay increases. For example, at asymmetric transmission speeds (eth1 = 100 Mbps and eth2 = 10 Mbps), the throughput drops only slightly, from 103 Mbps to 102 Mbps, as the delay increases from 10 ms to 50 ms. At symmetric transmission speeds (eth1 = 100 Mbps and eth2 = 100 Mbps), the throughput decreases from 186 Mbps to 184 Mbps. This throughput stability under higher delays suggests that the HighSpeed congestion control algorithm efficiently manages congestion and minimizes throughput loss even as network delay increases.

In contrast, the Uncontrolled case exhibits a more significant degradation in MPT-GRE tunnel throughput with increasing delays. At asymmetric transmission speeds (eth1 = 100 Mbps and eth2 = 10 Mbps), throughput decreases from 102 Mbps at a 10 ms delay to 93.3 Mbps at a 50 ms delay, indicating greater sensitivity to delay. At symmetric transmission speeds (eth1 = 100 Mbps and eth2 = 100 Mbps), throughput drops from 183 Mbps at a 10 ms delay to 137 Mbps at a 50 ms delay. This reduced responsiveness to delay leads to more pronounced degradation in MPT-GRE tunnel throughput.

On the other hand, Fig. 9 illustrates how varying levels of packet loss affect the throughput aggregation of the MPT-GRE tunnel. Under packet loss conditions, the BBR algorithm maintains higher throughput across all transmission speeds compared to the Uncontrolled case. An exception occurs when eth1 = 100 Mbps and eth2 = 10 Mbps at 1% packet loss, where throughput for the Uncontrolled case and BBR is 24.4 Mbps and 23.9 Mbps, respectively—a slight difference.

For example, at transmission speeds of eth1 = 100 Mbps and eth2 = 20 Mbps, BBR achieves a throughput of 33.5 Mbps, while the Uncontrolled case drops to 14.8 Mbps, representing nearly a 56% reduction in MPT-GRE tunnel throughput. This trend persists at all packet loss rates, with BBR consistently mitigating the negative effects of packet loss more effectively than the Uncontrolled case.

As packet loss increases to 5%, both scenarios experience a decline in MPT-GRE tunnel throughput. However, the degradation is significantly more pronounced in the Uncontrolled case, particularly at higher transmission speeds. For instance, when eth1 and eth2 are set to 100 Mbps, throughput in the Uncontrolled case plummets to 1.05 Mbps at 5% packet loss, while BBR maintains a throughput of 7.83 Mbps.

This comparison highlights the resilience of the BBR algorithm, which consistently maintains significantly higher throughput levels than the Uncontrolled case under packet loss conditions.

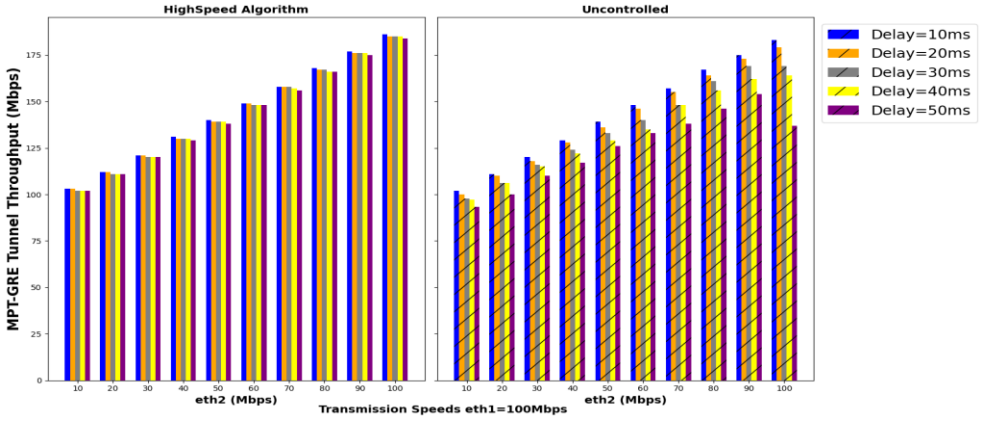


Fig. 8. Impact of the HighSpeed algorithm and the Uncontrolled case on MPT-GRE tunnel throughput under varying delay conditions.

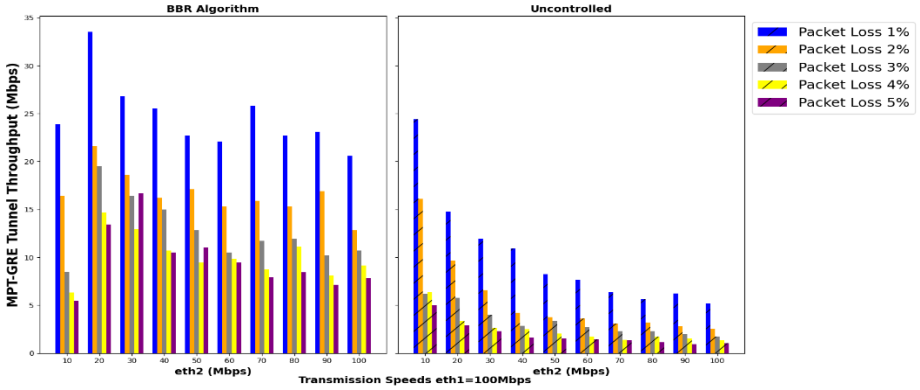


Fig. 9. Impact of the BBR algorithm and the Uncontrolled case on MPT-GRE tunnel throughput under varying packet loss conditions.

C. Determine the Appropriate Algorithm

Upon analyzing the results from the two scenarios, the HighSpeed congestion control algorithm outperforms others when network conditions are affected by delay, jitter, or a combination of both. This algorithm consistently maintains high throughput despite fluctuations in delay and jitter, making it particularly suitable for MPT-GRE multipath environments where these factors are common.

In contrast, the BBR algorithm excels under packet loss conditions, effectively adapting to varying levels of packet loss. Therefore, HighSpeed is the most appropriate choice in scenarios where delay and jitter are the primary challenges, while BBR is the preferred solution for networks where packet loss is the dominant issue.

When considering delay, jitter, and packet loss across various network conditions, the HighSpeed and BBR algorithms consistently outperform other options, making them optimal choices for maximizing throughput in diverse scenarios.

VI. Conclusion

In this study, we evaluated the impact of various congestion control algorithms on tunnel throughput in a multipath MPT-GRE network under different network conditions. The analysis focused on loss-based, delay-based, hybrid, and bandwidth estimation-based algorithms, including HighSpeed, CUBIC, H-TCP, LP, BBR, and Illinois. These algorithms were tested under varying delay, jitter, and packet loss conditions at symmetric and asymmetric transmission speeds.

The results indicated that some algorithms significantly improved MPT-GRE network performance, particularly when specific congestion control mechanisms were applied. HighSpeed significantly improved tunnel throughput as delay and jitter increased, while BBR enhanced throughput under packet loss conditions. BBR consistently outperformed the Uncontrolled case, exhibiting stable performance across various QoS conditions.

Our findings emphasize the importance of selecting appropriate congestion control algorithms for MPT-GRE networks. For example. HighSpeed performs well under delay and jitter, while BBR excels in packet loss scenarios, maximizing throughput and leveraging MPT-GRE's aggregation capabilities. This enables service providers to deliver more reliable and efficient communication solutions.

One of our future works is to create and design a specific congestion control algorithm for the MPT library.

Acknowledgment

The measurements were conducted remotely using the facilities provided by the National Institute of Information and Communications Technology (NICT) StarBED, located at 2–12 Asahidai, Nomi-City, Ishikawa 923-1211, Japan.

The authors thank Bertalan Kovács for reading and commenting on the manuscript.

The authors thank Brant von Goble from Széchenyi István University for proofreading the English-language version of the manuscript.

References

[1] B. Nawaz, K. Mahmood, J. Khan, M. Ul, A. Munir, and M. Kashif, "Congestion Control Techniques in WSNs: A Review," International Journal of Advanced Computer Science and Applications, vol. 10, no. 4, 2019, doi: 10.14569/IJACSA.2019.0100423.

[2] L. Yong, E. Crabbe, X. Xu, and T. Herbert, "GRE-in-UDP encapsulation," 2017.

[3] J. Zhao, J. Liu, H. Wang, C. Xu, and H. Zhang, "Multipath Congestion Control: Measurement, Analysis, and Optimization From the Energy Perspective," IEEE Transactions on Network Science and Engineering, vol. 10, no. 6, pp. 1–12, 2023, doi: 10.1109/TNSE.2023.3257034.

[4] B. Almási, G. Lencse, and S. Szilágyi, "Investigating the multipath extension of the GRE in UDP technology," Computer Communications, vol. 103, pp. 29–38, May 2017, doi: 10.1016/j.comcom.2017.02.002.

[5] A. Ford, C. Raiciu, M. Handley, O. Bonaventure, and C. Paasch, "TCP Extensions for Multipath Operation with Multiple Addresses," Mar. 2020. doi: 10.17487/RFC8684.

[6] Y. Thomas, G. Xylomenos, and G. C. Polyzos, "Multipath congestion control with network assistance," Computer Communications, vol. 153, pp. 264–278, 2020, doi: https://doi.org/10.1016/j.comcom.2020.01.071.

[7] Ł. Łuczak, P. Ignaciuk, and M. Morawski, "Experimental Assessment of MPTCP Congestion Control Algorithms for Streaming Services in Open Internet," in FedCSIS (Communication Papers), Oct. 2023, pp. 359–364, doi: 10.15439/2023F9991.

[8] H. Moradiya and K. Popat, "Evaluating TCP Performance with RED for Efficient Congestion Control," in International Conference on Advancements in Smart Computing and Information Security, Springer, 2024, pp. 403–414.

[9] J. Zhang, Z. Yao, Y. Tu, and Y. Chen, "A Survey of TCP Congestion Control Algorithm," in 2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP), 2020, pp. 828–832, doi: 10.1109/ICSIP49896.2020.9339423.

[10] N. Al-Imareen and G. Lencse, "Effect of Path QoS on Throughput Aggregation Capability of the MPT Network Layer Multipath Communication Library," Infocommunications journal, vol. 15, no. 2, pp. 14–20, 2023, doi: 10.36244/ICJ.2023.2.3.

[11] S. Szilágyi and I. Bordán, "The effects of different congestion control algorithms over multipath fast ethernet IPv4/IPv6 environments," CEUR Workshop Proceedings, vol. 2650, pp. 341–349, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:221662730.

[12] Y. Cao, M. Xu, and X. Fu, "Delay-based congestion control for multipath TCP," Proceedings - International Conference on Network Protocols, ICNP, pp. 1–10, 2012, doi: 10.1109/ICNP.2012.6459978.

[13] R. Melki, M. M. Mansour, and A. Chehab, "A fairness-based congestion control algorithm for multipath TCP," IEEE Wireless Communications and Networking Conference, WCNC, vol. 2018-April, pp. 1–6, 2018, doi: 10.1109/WCNC.2018.8377078.

[14] G. Lencse, S. Szilagyi, F. Fejes, and M. Georgescu, "MPT Network Layer Multipath Library," 2021. [Online]. Available: https://datatracker.ietf.org/doc/html/draft-lencse-tsvwg-mpt-10.

[15] M. B. M. Kamel, I. Ahmed Najm, and A. Khalaf Hamoud, "Congestion Control Prediction Model for 5G Environment Based on Supervised and Unsupervised Machine Learning Approach," IEEE Access, vol. 12, pp. 91127–91139, 2024, doi: 10.1109/ACCESS.2024.3416863.

[16] R. Al-Saadi, G. Armitage, J. But, and P. Branch, "A Survey of Delay-Based and Hybrid TCP Congestion Control Algorithms," IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3609–3638, 2019, doi: 10.1109/COMST.2019.2904994.

[17] H. Jamal and K. Sultan, "Performance analysis of TCP congestion control algorithms," International Journal of Computers and Communications, vol. 2, no. 1, pp. 30–38, 2008, [Online]. Available: http://w.naun.org/multimedia/UPress/cc/cc-27.pdf.

[18] S. Patel, Y. Shukla, N. Kumar, T. Sharma, and K. Singh, "A Comparative Performance Analysis of TCP Congestion Control Algorithms: Newreno, Westwood, Veno, BIC, and Cubic," in 2020 6th International Conference on Signal Processing and Communication (ICSC), Mar. 2020, pp. 23–28, doi: 10.1109/ICSC48311.2020.9182733.

[19] A. Roy, J. L. Pachuau, and A. K. Saha, "An overview of queuing delay and various delay based algorithms in networks," Computing, vol. 103, no. 10, pp. 2361–2399, Oct. 2021, doi: 10.1007/s00607-021-00973-3.

[20] M. A. Yousuf, M. M. Islam, M. S. Hosen, and M. L. Ali, "Round-Trip Time and Available Bandwidth Estimation Based Congestion Window Reduction Algorithm of MPTCP in Lossy Satellite Networks," Journal of Physics: Conference Series, vol. 1624, no. 4, p. 042024, Oct. 2020, doi: 10.1088/1742-6596/1624/4/042024.

[21] R. Gonzalez, J. Pradilla, M. Esteve, and C. E. Palau, "Hybrid delay-based congestion control for multipath TCP," in 2016 18th Mediterranean Electrotechnical Conference (MELECON), Apr. 2016, pp. 1–6, doi: 10.1109/MELCON.2016.7495389.

[22] F. Fejes, "MPT – multi-path tunnel," precompiled version can be downloaded from:, 2019. http://github.com/spyff/mpt.

[23] N. Al-Imareen and G. Lencse, "On the Impact of Packet Reordering in MPT-GRE Multipath Networks," in 2023 46th International Conference on Telecommunications and Signal Processing (TSP), Jul. 2023, pp. 82–86, doi: 10.1109/TSP59544.2023.10197737.

[24] N. Al-Imareen and G. Lencse, "MPT connections files," 2023. [Online]. Available: https://github.com/NaseerAJabbar/MPT_Connections_files.

[25] N. Al-Imareen and G. Lencse, "Implement congestion control algorithms using iperf3 for MPT-GRE multipath network," 2024. https://github.com/NaseerAJabbar/Congestion-Control-Algorithms.

[26] K. P. Jon Dugan, Seth Elliott, Bruce A. Mah, Jeff Poskanzer, "Iperf3 documentation," 2018. https://iperf.fr/iperf-doc.php.

A Chatbot for the Legal Sector of Mauritius Using the Retrieval-Augmented Generation AI Framework

Taariq Noor Mohamed[1], Sameerchand Pudaruth[2], Ivan Coste-Manière[3]
ICT Department-FoICDT, University of Mauritius, Reduit, Moka, Mauritius[1, 2]
SKEMA Business School, Sophia Antipolis, France[3]

***Abstract*—Mauritius is known to have a hybrid legal system as the logical consequence of being both a former French and English colony. From its independence in 1968 to date, the legal environment has changed to reflect the constant need to provide a framework to address the country's diverse needs. With over 1200 pieces of legislation available for consultation, including those which are no longer in force, it is very difficult to know all of them. Yet, there is a legal maxim that says, "nemo censetur ignorare legem". In other words, ignorance of the law is no excuse. This study aims to provide a solution for professionals and non-professionals to have better access to the law through the development of a chatbot. A Retrieval Augmented Generation (RAG) chatbot system has been developed to achieve this objective. A RAG system is one that leverages the use of Large Language Models (LLM) to process a query and generate a response, while ensuring accuracy by performing similarity searches against documents stored in a vector database. A sample of 46 legal documents (acts and regulations) were retrieved from the website of the Supreme Court of Mauritius. They were broken down into chunks and stored as vectors in Chroma, a vector database. The chatbot combines and processes the queries with a text prompt, searches the relevant legal texts, and generates an appropriate response using OpenAI GPT-4o-mini or MistralAI Open-Mixtral-8x22B. Since most legal texts are in English, a translation layer is included for queries in French. Sources for the answers are also displayed for easy cross-validation. This chatbot will undoubtedly be a useful tool for the Mauritian people.**

Keywords*—*Law; chatbot; retrieval augmented generation; large language model; OpenAI; Mistral AI

I. Introduction

Artificial Intelligence (AI) is the technology that enables machines to mimic humans in performing complex tasks such as learning, solving problems, making decisions, and creativity. From its humble beginnings in the 1950s, when Alan Turing first introduced the Turing Test, to the recent explosion in popularity of generative AI [1], AI has had such an impact on our everyday lives that the European Union has even introduced a regulation recently on the use of AI [2]. The use of AI can be seen in a multitude of sectors, for example, in healthcare to detect skin cancer [3] or in education [4]. The legal field is no exception, and given the famous saying that none shall be ignorant of the law, there is a need to always improve access to justice for everyone.

Over the past few decades, our world has experienced rapid growth in the technological sector. Thus, there are a multitude of AI-related solutions that have been developed to tackle issues all around the world. The importance of having a Legal Information Retrieval system (LIR) was emphasized by [5], where they focused on three kinds of LIR systems using NLP techniques, ontology-based approaches, and deep learning-based methodologies. Amato et al. [6] presented the conversational agent CREA2 designed for tasks in the legal domain, giving users advice on legal procedures or legal document drafting. It can also help in resolving disputes between individuals with regard to the European Union's legislation. In Canada, Quedot et al. [7] investigated methods to create chatbots for immigration and corporate issues. Morgan et al. [8] delved into the creation of a chatbot framework to help children have a better understanding of their legal rights. Another interesting approach is from Alam et al. [9], where a chatbot model was created to predict the outcome of legal cases in New Zealand. Firdaus et al. [10] developed a Question-and-Answer bot on Telegram with regard to the Law on Information and Electronic Transactions in Indonesia. Devaraj et al. [11], Ioannidis et al. [12], Cao [13], and Nguyen [14] developed several solutions for the legal sector where LLMs were used.

All countries have different laws specific to them, including Mauritius, known for its hybrid legal system as a result of its long history of colonialism. The laws of Mauritius are drawn from a wide range of sources, such as the French Civil Code and the English Common Law [15]. The legal field in Mauritius has evolved significantly since the enactment of the Constitution in 1968. Over the years, Mauritius has introduced various laws and regulations to promote various sectors of the economy and to bring stability and justice to the country. However, the dynamic world in which we live experiences new events all the time, and this creates a need to regularly amend existing laws or implement new ones. This can be seen throughout the multitude of parliamentary processes [16]. Certainly, this dynamic legal environment poses a challenge to the population in general, especially to those who might not be well-versed in the legal field. It is difficult for most people to get a clear answer on legal matters without having to dig into lengthy government sources and legal databases or seek help from legal professionals.

There have not been any developments for an AI solution to address this problem in Mauritius. A Legal Information Retrieval System has been developed by study [17], which can be accessed at lawanswers.me. The platform enables users to input queries in natural language. The system processes these queries by extracting pertinent keywords and provides relevant sections of Mauritian law and Supreme Court cases from 1968

to 2017. Users can customise the number of results displayed and filter searches to show only case judgments or act titles. The platform is also accessible via mobile devices, ensuring user-friendly access to legal information.

Pudaruth et al. [18] also used deep learning techniques to classify 490 legislations in the Republic of Mauritius. The development of a chatbot for the legal field will ensure that any Mauritian citizen will have better access to justice, in particular, to the statutory provisions contained in the Acts of Parliament. By leveraging the use of LLMs, an RAG system will be developed to create a chatbot that can answer legal queries using natural language. The system will be fed with a list of relevant acts and regulations so that an appropriate response can be obtained by performing a similarity search on the query and the available documents. A successful implementation of this project will save hours or even days for the Mauritian population in getting their queries answered.

The paper proceeds as follows. An overview of the legal system of Mauritius is provided in Section II. Section III covers the related works. The methodology is described in Section IV, while the results are provided in Section V. Section VI concludes the paper.

II. Background Study

"At his best, man is the noblest of all animals; separated from law and justice, he is the worst." (Aristotle). Since time immemorial, human survival and progress have always been attributed to the fact that Man is a social animal. From ancient tribes to the modern civilisations, mankind has always been associated with communal living, having one common component, rules and regulations. The Law has always been a key aspect of any society. It helps in maintaining the structure and safety in society. Without this pillar, society crumbles and gives rise to anarchy.

Mauritius is one of the countries with a hybrid legal system that encompasses both the French Civil Law and English Common Law. The country has been subjected to a long history of colonisation attempts, starting with the Dutch from 1598 to 1710, during which there was no administration of justice as we know it today. Later, from 1715 to 1810, the French colonised the island where they introduced their laws and jurisdiction to the country. A temporary "Conseil Provisoire " was set up in 1721 to deal with civil and criminal matters. In 1723, a "Conseil Provincial," a court of First instance, was put in place to treat those issues, later replaced by "Conseil Superieur" in 1734, with appeals made to the "Conseil Supérieur de Bourbon." A two-tier jurisdiction was then introduced in 1771 with a "Jurisdiction Royale" hearing cases at first instance, with the "Conseil Supérieur" being a court of higher instance, which were afterward renamed as "Tribunal de Première Instance" and "Tribunal d'Appel" respectively. When the British colonised the island in 1810, all prior judicial institutions were kept, but justice had to be delivered in the name of the King. In 1851, "La Cour de Première Instance" was abolished, and the Supreme Court was established in Mauritius, replacing the "Cour d'Appel," but the right to appeal to the "Judicial Committee of the Privy Council" was maintained. Gradually, the system returned to a two-tier judicial system with the establishment of the subordinate courts [19].

With this long history of colonialism, the Mauritian legal system draws from a wide array of sources to establish and enforce laws in the country. The French Civil Law, mainly for civil rights, criminal law, and commercial transactions; the English Common Law for evidence, tort, negligence, and health and safety; domestic legislations, that is, Acts of Parliament, customs, and international treaties and conventions [15]. Sitting on top of all laws in the country is the Constitution, also known as the supreme law, in which all the basic set of principles and laws of a country are written, such as determining the powers and duties of the government and the rights of the population. Any law deemed inconsistent with the Constitution is deemed to be void [20].

Without a proper system, all the laws might be meaningless and not be practiced at all; hence, the judicial system in Mauritius, the Judiciary, is one of the three main pillars of our country. This system is structured into two main tiers: the Supreme Court and the subordinate courts. The Supreme Court is composed of the Chief Justice, the Senior Puisne Judge, and twenty-three Puisne Judges and has unlimited jurisdiction to hear and determine any civil and criminal proceedings. It functions in a similar fashion to the High Court of England and has complete authority over all subordinate courts to ensure that proper justice is delivered. The Supreme Court functions both as a court of first instance and as an appellate court. As a first instant court, it is subdivided into multiple divisions, such as the Master's Court, Family Division, Commercial Division, Criminal Division, Mediation Division, Financial Crimes Division, and Land Division. Each division handles specific types of cases, such as financial crimes, land disputes, family matters, and commercial issues. Subordinate courts, on the other hand, typically handle less complex cases or involve claims of a lower monetary value and consist of the Intermediate Court, the Industrial Court, the District Courts, the Bail and Remand Court, and the Court of Rodrigues. Cases can be escalated to the Supreme Court from subordinate courts through appeals. Any decisions made by subordinate courts can be heard again and reviewed by the Supreme Court. Additionally, the Supreme Court has divisions specifically for civil and criminal appeals: the Court of Civil Appeal and the Court of Criminal Appeal. These divisions handle appeals from decisions of the Supreme Court in its original jurisdiction as well as from subordinate courts. In some cases, the decision from the Supreme Court can even be reviewed if an appeal is made to the Judicial Committee of the Privy Council. The decision that will be taken by the Privy Council is then deemed the final one [21].

There are also multiple parties involved in any typical court proceedings. These include the judge, the lawyers, the plaintiff, the defendants, the jury, and witnesses. The judge is the one who will ensure the smooth running of the hearing and will give the final verdict. The plaintiff is the individual or entity who brought the case to a court to seek some form of legal remedy, while the defendants are the ones on whom the legal action is taken. Lawyers are also typically assigned to each of the two opposing parties and represent each of the parties. Witnesses are the ones who can provide testimonies on specific

events that occurred for the case being heard. The jury also forms an important part of proceedings in court. It consists of 9 random civilians between the ages of 21 and 65 that are tasked to identify the true evidence in specific cases [22]. The basic idea for having a system of jury is to ensure that the verdict of the case is what society would expect it to be.

It is a known fact that engaging in a lawsuit means, in most cases, the need to invest a lot of time, effort, and money. In this optic, it is necessary that not all legal matters need to be tried in a typical court of justice; hence, the Mediation Division of the Supreme Court has been established for this purpose [21]. Mediation is one way in which disputes can be kept out of courts to save time, money and prevent extra stress from all involved parties, and even giving a more flexible judgement. The key distinction is that mediation involves a neutral, independent third party who challenges the parties on their positions and assists them in finding a compromise. The decision is not taken solely by one person, but instead the parties are guided into reaching a proper decision to resolve their dispute [23].

III. Literature Review

Legal chatbots are becoming very popular because they are now use artificial intelligence techniques to help people find relevant legal information. This section looks at some of the studies that have been done on legal chatbots. We also explain how they have been developed, how they work and what are their limitations?

Morgan et al. [8] delved into a chatbot framework that integrates machine learning, a dialogue graph and information to ease access for children about their legal rights. The user first initiates a conversation with the chatbot, and afterward, based on the dialogue graph and legal type, the chatbot identifies the legal circumstances being addressed by the user. The different parties in the conversation are identified, and based on this information, a case is created so that a legal advisor can take over. Since obtaining legal conversation in a child's way of speaking is difficult, the authors created fictional data from "artificial" statements that approximate the way a child talks and from "real" statements obtained from a study whereby adults imitated children's language. The dataset was divided into "speech act," such as greetings or affirmations, and "legal type," relating to the legal issues being tackled, such as abuse or cyber-crime. Throughout the conversation, the chatbot's neural network classified the messages as speech act or legal type to give the appropriate response. This is achieved through the tokenization of the messages and converting them into word vectors of 200 dimensions where semantic similarities can be computed. To encode the impact of word order in the different sentences, two Long Short-Term Memory (LSTM) layers and a type of Recurrent Network Layer (RNN) were used, and the outputs were processed by a dense layer with a ReLU activation function. Finally, the sentences were categorised into one of the two classes by two parallel dense layers with a softmax activation. Named entities were also recognised and extracted from the conversation as these are used later in the generated report.

In the pre-chatGPT era, Queudot et al. [7] focused on enhancing access to legal information using chatbots for two different matters: for immigration issues and for legal issues related to the corporate environment in Canada. In both cases, the chatbots make use of the MLFlow and RASA frameworks. They made use of the MLflow framework for two purposes: to keep track of script execution for data collection, data preparation, model training, and evaluation and to simplify script executions with different values. For the first chatbot, the dataset used was obtained from 1088 different web pages from the Canadian Immigration and Citizenship help desk. Then, two different intent classifiers were used: the standard model in RASA using the StarSpace algorithm and a modified Information Retrieval (IR) inspired model. The StarSpace is an algorithm in RASA used in learning the embeddings for different entities in a supervised manner. There was not much difference in the performance of the two algorithms. In the second chatbot, the dataset used was two formulations for a series of 275 questions developed by the National Bank of Canada (NBC). Three approaches were used to compare the performance of the chatbot: the classic StarSpace Model, the IR variant of StarSpace, and the BERT transformer model. For the BERT model, fine-tuning a pre-trained model was more feasible rather than training a model from scratch. The results, in this case, were in favour of using the BERT model.

Firdaus et al. [10] focused on providing a legal chatbot to search for legal documents for the Law on Information and Electronic Transactions in Indonesia. NLP techniques were used for the chatbot development and Telegram, the messaging application, as it provides much support for chatbot development. For the system to understand the query of the user and understand the appropriate context, NLP components such as parser, lexicon, understander, knowledge base, and response generator are used. In case the query contains "non-standard" words, the Levenstein distance method is used to obtain the closest "standard" word that can be processed by the system. TF-IDF cosine similarity is then applied to compare user queries with the stored documents in the database to find the appropriate answers. Questions found directly in the knowledge base had a 100% match. Random questions showed an average of 77% match, and any questions that were still related to the knowledge base showed an accuracy of 83%.

Sansone and Sperlí [5] emphasised on the importance of having a Legal Information Retrieval (LIR) in the legal field given the increasing volume of Electronically Stored Information (ESI). Given the complexity of legal documents, efficient searching algorithms are required. The authors focused on three types of LIR systems: NLP techniques, ontology-based approaches, and deep learning-based methodologies. For each type of system, different models were tested, evaluated and categorised. NLP-based LIR approaches were further classified into three categories, rule-based, NLP-based and a combination of NLP and ontologies. The main limitation of this approach is the manual definition of rules required, which is hindered by the large size of the vocabulary and can introduce noise during analysis, hence affecting the accuracy of the results. Ontology-based approaches were classified into top-down and bottom-up strategies. Ontology refers to the way of organising and representing knowledge in

a particular domain. Key limitations of this approach include the few available domain-specific ontologies and the complexity involved in their maintenance and contextual dependencies. Deep-learning methodologies were classified into pre-trained embedding models, domain specific embedding models and hybrid embedding models. Limitations include the limited training samples available.

Amato et al. [6] presented an AI-powered conversational agent named CREA2, designed specifically for the legal domain. The purpose of the agent is to help in guiding users on legal advice and procedures and give recommendations in drafting legal documents. CREA2 also helps in resolving disputes between individuals pertaining to the European Union legislations. They also discuss the different approaches that they could consider in building the project, namely, intent classification, question-and-answer-retrieval and similar question retrieval. Given their relatively small amount of dataset, the last approach has been used. The system makes use of the Natural Language Processing (NLP) algorithm to interpret the user's questions and provide the required response. Sentence-BERT (SBERT) is used to compare the user's questions to the set of available questions in the dataset. SBERT works as follows. It first creates embeddings (vectors) for each pair of questions and then computes a score using the cosine similarity function. The scores are then sorted in a descending order, and the highest scoring one is selected as it provides the most appropriate response from the database. Data augmentation techniques are also used, given the relatively small training data size. The paraphrase-based utterance augmentation framework called Parrot is used to create variations for different questions available in the training dataset. Adequacy and fluency can also be controlled in the framework. Finally, the data are stored in JSON format in a question-answer format as separate strings. Two pre-trained models and a fine-tuned version of SBERT were used. The results demonstrated that the fine-tuned version of quora-distilbert-multilingual-v2 had a higher accuracy than all the other models.

Alam et al. [9] developed a chatbot that can predict the outcome of legal cases based on past cases made available to the system. Data from the New Zealand Employment Relations Authority (NZERA) was used. Data preprocessing was used on each PDF prior to the experimentation. This included using regular expressions to identify and extract paragraphs. A manual scan had to be done on each paragraph to identify the document preamble (P) feature and the case determinations (D) feature. Eventually, two kinds of data were found: Full Documents (FD) having both P and D features, and Full Documents with the result removed (FD - D) so that independent ruling can be done based only on the circumstances. For the semantic analysis, an unsupervised topic detection method called Latent Dirichlet Allocation (LDA) was used. This enabled the creation of 10 topic-clusters each consisting of distinct features (d) and a number of top-words (n). Cosine similarity was further performed on the different clusters by keeping only words related to each cluster, performing tokenisation on the data and converting them to 128-dimensions word embeddings. For the predictive analysis, supervised learning was further performed on the data using different deep neural network models and tested on both FD and FD-D data. Each model used the Gated Recurrent Unit (GRU) variant of the RNN on TensorFlow, trained for 26 epochs. Sigmoid and SoftSign activation functions were used to test each model. RNN was used as a based model to evaluate the performance of the different LDA parameters, namely, K, n, and d. In the semantic analysis with LDA and RNN, it was observed that having 5 LDA topics, 5000 features and 300 top words per topics provided the best results.

Devaraj et al. [11] created a chatbot capable of answering questions based on documents made available to it. The chatbot was divided into three parts. Langchain, an open-source NLP framework that enables LLMs to be combined with external data, was used in the creation of the chatbot's custom knowledge. It breaks down the different documents into small chunks and stores them as vectors in a vector database, using Open AI's text embeddings. When provided with a prompt, it is compared with the data in the vector store using the cosine similarity function to determine the most appropriate response. A Flask Application was created as the backend for the entire application. It can be hosted on a server, thus enabling the query processing feature to be used remotely. Flask was used because it is lightweight, it provides for RESTful API support, Python integration and it is scalable. To make the system user friendly, the authors also designed a mobile application. Kotlin was used for the backend and XML for the frontend.

Ioannidis et al. [12] took advantage of the potential that LLM offers to create generative AI solutions regarding regulatory compliance for businesses. These were a horizon scanning tool, an obligations generator tool, and an LLM-based expert system. For the horizon scanning tool, web scraping was performed periodically on different websites that relates to the different regulatory news to fetch and store up to date information in a database. In the study, Australian regulatory bodies' websites were scraped. GPT prompts were then applied to the scraped data so that the following can be performed: generate a 150 words summary of the data, categorise the level of impact of the regulatory update as either low, medium or high and generating hashtags for easy identification of the scraped data. For the Obligations Generator, a prompt is always used along with the legislations, made available to the system, to generate a summarised list of obligations that a company should comply with. It was noticed that GPT-4 did a better job than GPT-3.5 in understanding the prompt and writing the text. The lists generated were then stored in a local database before being processed further to remove duplicate obligations using the cosine similarity measure. GPT-4 was also used to create an expert consultation tool where users can converse with the chatbot, and the latter responds in accordance with legislation or obligations available to it. This part of the system made use of the LangChain framework and embedding tool from Open AI. The authors also mentioned that LLMs are prone to hallucination. To combat this issue, a "human in the loop" system is used whereby a human verify and validate any output from the LLM before pushing them to the database.

Cao [13] highlighted how LLMs can easily answer general questions but sometimes struggle to give an appropriate answer

in specific topics such as medical or law consultations. For such kind of scenarios, the author investigated a Task-Oriented Dialogue approach so that the AI agent can ask appropriate questions to the user so that it can properly diagnose the issue encountered by the user. A multi-agent, with GPT4 as base LLM, was used. It is made up of different parts: a chat agent, a Topic Manager, a Topic Enricher, and a Context Manager. Each LLM used is provided with a different and well detailed prompt so that they can perform their actions. The Topic Manager has the user query, the action list, the status of the topic stack, and the chat history. All of these enable the topic manager to plan and make the appropriate decision throughout the conversation. The Topic Enricher uses the current topic and enriches it based on the current dialogue context. With the context provided by the context manager, the chat agent can generate an answer.

Nguyen [14] from the National Institute of Informatics in Japan gave a brief overview of LawGPT, a chatbot that provides legal assistance, since the model developed is protected by a non-disclosure agreement. LawGPT 1.0, uses a similar architecture to GPT-3, that is, a transformer architecture which enables a better understanding of the context of a sentence provided, by weighing the importance of different words in the sentence. Being a fine-tuned model of GPT3 on a large corpus of legal texts using standard deep learning techniques, such as stochastic gradient descent and backpropagation, LawGPT is designed for the legal context. Evaluation of LawGPT involves answering legal questions, generation of legal documents and the legal advice it can propose. The author mentions that the results were highly positive. However, no specific implementation details about LawGPT were to be found in the paper as the model is protected by a non-disclosure agreement.

Medeiros et al. [24] exploited the use of LLMs to create AI tools for obtaining answers from a car manual. Three approaches were selected by the authors, and in each case, the LLM models from the GPT-3 family were used, and a chunk size of 1000 and a chunk overlap of 20 were used. The performance of each was evaluated using different kinds of prompts, namely zero-shot, one-shot, and few-shot. For Doc Chatbot, the user directly interacted with the chatbot by typing the questions in a Python notebook. LlamaIndex and LangChain libraries were used to build the chatbot to interact with the PDF documents. Open AI API and the text-davinci-003 model are used to create embeddings. It was found that the bot showed low accuracy by providing accurate answers for only 1 out of 4 evaluated questions across different prompts. The cost for each question and pair was between 0.03 and 0.04 USD. In Ask Your PDF, a user-friendly interface was created to interact with the chatbot using the Streamlit tool. The LangChain library is used in the chatbot development and to interact with the PDFs. Open AI API and the text-davinci-003 model are used to create the embeddings. Here, the bot demonstrated average accuracy by providing accurate answers for 2 out of 4 evaluated questions when using the zero-shot prompt. The cost for each question and answer was between 0.02 and 0.03 USD. The Question-and-Answer System had a user-friendly interface, with the frontend built using React and the backend with FastAPI and a vector database called Qdrant. Libraries used were LangChain and the Sentence Transformers. Open AI API and all-mpnet-base-v2 model was used to generate the embeddings. However, it was noticed that the bot had a low accuracy, but it had a much lower cost of operation averaging below 0.005 USD for each question and answer.

This review on legal chatbots shows that these tools are getting more advanced due to massive progress in the field of artificial intelligence. Furthermore, these studies show how LLMs are being used for different purposes. For example, some legal professionals are using it to answer legal queries while others are using it to sift through information quicker and more accurately. Nevertheless, since the laws of each country are different, we still need to develop a customised system for the judiciary of Mauritius.

IV. Methodology

To make the laws of the Republic of Mauritius more accessible, a chatbot with a Retrieval Augmented Generation (RAG) system is proposed. Contrary to traditional rule based chatbots, our solution leverages the power of generative AI such that a human-like conversation can be achieved. The general structure of the system is as follows: legal texts, acts, and laws are fed to the system in the form of PDF documents. The PDFs are broken into chunks, and each chunk is converted into vectors using an embedding algorithm. A chunk size of 1000 characters with an overlapping of 80 was initially used. Chunk overlapping is important as a sentence can be split into two chunks, which can have an adverse effect on the generator in understanding the proper context of that sentence. To create the embeddings, the *mistral-embed* model has been used for MistralAI Open-Mixtral-8x22B while the text-embedding-ada-002 model has been used for OpenAI GPT-4o-mini. These embeddings are then stored in the Chroma vector database. The user inputs his query, which is embedded and compared with documents in the vector database. An answer is generated based on the contents of the selected documents. The decision to use a RAG system was mainly due to its ability to minimise hallucinations, provide contextually accurate answers while maintaining a level of transparency by showing what documents are used to generate a response.

By consulting a legal professional, it was found that there is a multitude of acts and regulations that are not properly updated and that the acts only are not sufficient as sources of information. Given the huge amounts of acts and regulations available, two acts, namely the Employment Relations Act and Worker's Rights Act, were selected. All the regulations supporting these two acts were also selected. A few other general and relevant acts, such as the Constitution, the Civil Code, and the Data Protection Act, were also made available to the system. The full list of legal documents used are listed in Appendix 1. The architectural diagram of the RAG system with the tools and technologies used is illustrated in Fig. 1.

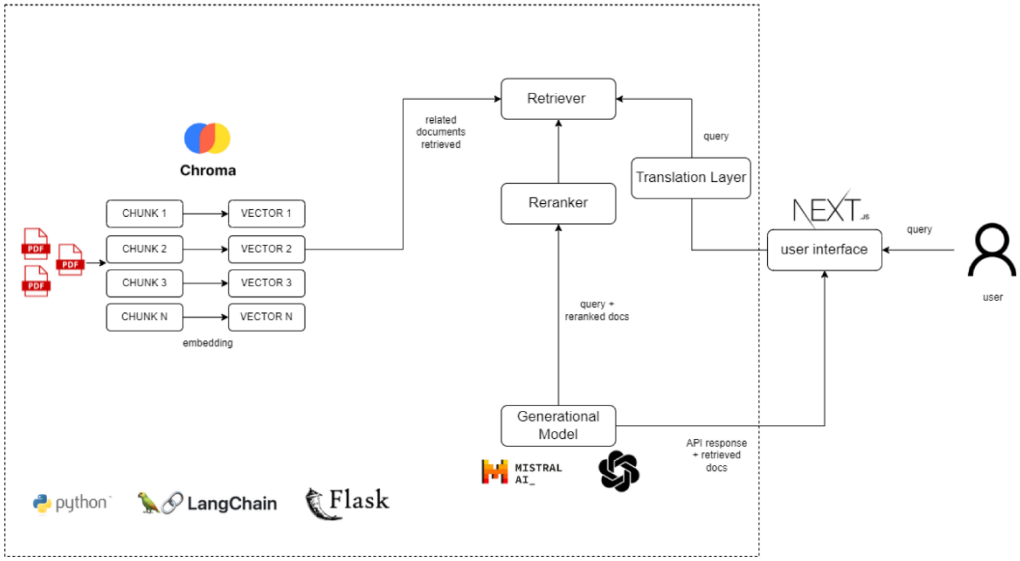


Fig. 1. Architectural diagram of the RAG system.

The user starts by selecting the desired LLM and then types his question on the user interface. LangChain is selected as the LLM framework because of its ability to provide context aware outputs. It also provides all the necessary tools and structures to facilitate the integration and management of large language models into applications. This causes the frontend to call the Flask API to pass the question to the backend. The query then passes through a translation layer to convert the text in English (if it was in French), otherwise it is directly converted into a vector form. The query is then used by the vector store retriever and searches for a number (K) of related documents in the vector database using the cosine similarity search.

The retrieved documents also go through a reranker, where they are processed again to select the top N documents. The top K documents in the retrieved documents may not all be as relevant to the question. For instance, it is possible that a relevant document was given a lower score. Thus, a reranker has been used to circumvent this issue. The reranker assigns a new score to each document and retrieve the top 3 documents by default. However, in our case, we increased this threshold to 10 by modifying the source file in the Flash ReRank library. This helps the system in having a larger context to generate an answer. These documents are then sent alongside the query to the generational model so that an appropriate response can be obtained. The final response alongside the sources of answer is sent back to the frontend and displayed on the screen. In contrast with ChatGPT or other similar software, the queries and answers will be specific to the documents made available to the application that is in the context of the legal sector of Mauritius only.

The user interface is made up of the following: a home screen, a chat panel, a sidebar to select the LLM for the conversation, a text box, a send button, and a source button to check which documents are considered for the answer. The user interface is shown in Fig. 2. The libraries used in the backend are shown in Table I.

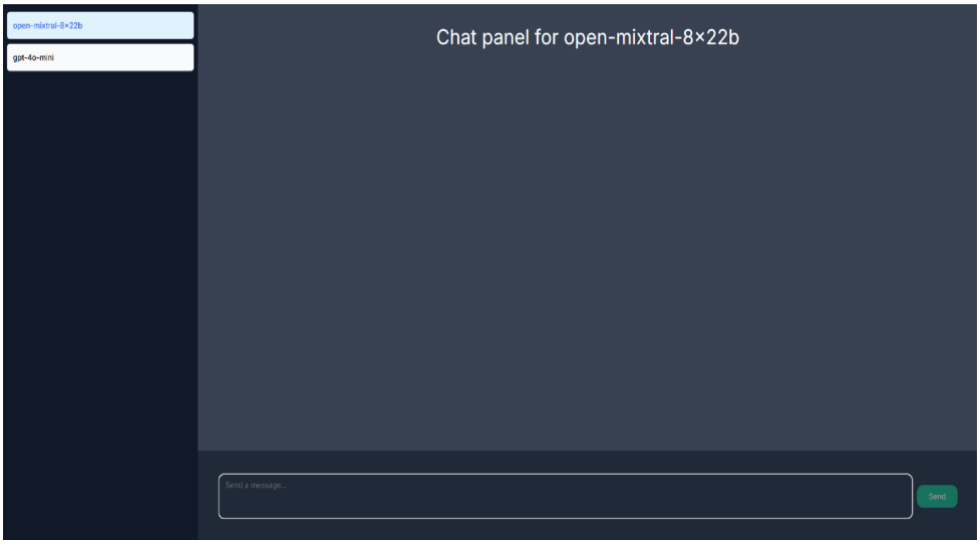


Fig. 2. Chat panel for the system.

TABLE I. BACKEND LIBRARIES

Library Name	Description
mistralai	Mistral AI Python library
openai	Open AI Python library
chromadb	Chromadb Python library
langchain	Bare minimum requirements of LangChain for Python
Langchain_mistralai	LangChain integrations for Mistral AI
Langchain_Open AI	LangChain integrations for Open AI
Langchain_community	Contains third-party integrations
Pypdf	Free and open-source pure Python PDF library to parse PDF files
Flask	Lightweight WSGI web application framework
Flask-cors	Flask extension for handling Cross Origin Resource Sharing (CORS)
Python-dovenv	Enable use of the .env file
FlashRank	Flash ReRanker library
Langdetect	Library to detect the language used

The libraries used in the frontend are shown in Table II.

TABLE II. FRONTEND LIBRARIES

Library Name	Description
Next	Next js react library
React	React library for JavaScript
React-dom	Entry point to the DOM and server renderers for React
React-markdown	React component to render markdown
Next-ui	Next JS library having beautiful components
Tailwindcss	CSS framework
Github-markdown-css	The minimal amount of CSS to replicate the GitHub Markdown style

For the frontend, the React framework and Next JS have been used. The following components were created to build the user interface: (a) Source Modal: This is a pop-up that can appear when the source button is pressed to view sources, (b) nav-links: This contains the navigation links to switch between LLMs, (c) Message: This is the component used to display messages sent by the user and received by the chatbot, (d) ChatSideBar: This is the Side Bar component where the user can switch LLMs and (e) Chat: This is the main chat component that displays the textbox, sidebar and chat components.

V. RESULTS AND EVALUATION

To perform the evaluation of the RAG system, a User Acceptance Test (UAT) was performed by a legal professional, a third year LLB student and a non-technical user. They asked questions to the chatbot and evaluated if the received answers match the expected answer. Initial testing showed that many questions were not properly answered. This made us to fine-tune the chunking overlap parameter to 200 and re-embed the document so that each chunk provides more context to the chatbot.

Both systems were able to provide relevant responses in most cases. However, it was observed that the model from Mistral AI (open-mixtral-8x22b) performed better when the query was in French, and the responses were located in French documents. The system was evaluated by three users: a legal professional (LP), a final year LLB student (LLB) and a Mauritian citizen (MC) without any formal legal background. The legal professional extended valuable help in improving the system by sharing the relevant documents for this system. He found that the system was satisfactory and will indeed help to improve access to justice in our country.

The LLB student also mentioned that both systems met his expectations for both basic and more specific legal questions sections, but he was of the opinion that the model from Mistral AI consistently provided more detailed, structured, and comprehensive answers than the model from Open AI. Notably, the model from Mistral AI excelled in explaining the duties of a company secretary, the procedure for incorporation, and employee protections under the Worker's Rights Act, offering greater depth and alignment with the relevant laws. As regards the third evaluator, he found the system to be very user-friendly and helpful.

The list of questions and answers from the various testers, together with their expected answers and remarks on the system, are provided in Appendix 2. Table III summarises the list of questions and answers from the three evaluators.

TABLE III. SUMMARY OF QUESTIONS

#	Question	User	gpt-4o-mini	open-mixtral-8x22b	Comments
1	I have been in employment with ABC since January 2024. Am I entitled to special leave?	LP	correct	correct	
2	What is a labour contractor?	LP	correct	correct	
3	Je suis convoqué à un comité disciplinaire dans une semaine. Comment savoir si le comité est valablement constitué?	LP	incomplete	incomplete	No mention of right to be accompanied.
4	Mon employeur ne respecte pas ma vie privée. Quels sont les moyens dont je dispose pour me protéger ou pour le dissuader?	LP	incomplete	incomplete	No mention of Code Civil and DPA.
5	I intend to cease business as a Company. How can I terminate the employment of staff?	LP	incomplete	incomplete	Mistral AI is better.
6	Quel est le salaire minimum à ce jour dans la République de Maurice?	LP	correct	correct	
7	What are my obligations as an employee?	LP	correct	correct	Implied obligations not mentioned.
8	I would like to hire the services of my nephew who is 15 years old. Can I do so?	LP	correct	correct	

9	I would like to transfer shares I hold in Company FEDER to my brother. What do I need to do?	LP	correct	correct	Registrar office not mentioned.
10	What does section 3 of the Constitution of Mauritius, which guarantees the right to life, liberty, and security of a person, balance individual freedoms with the state's responsibility to maintain public order and safety?	LLB	correct	correct	Model from Mistral AI is better.
11	Which section and chapter of the Constitution caters for the freedom of expression?	LLB	correct	correct	
12	What does the Constitution states about the powers of the President of Mauritius?	LLB	correct	incomplete	
13	What are they key duties of a company secretary under the Mauritius Companies Act 2001?	LLB	wrong	correct	
14	Can you explain the procedure for the incorporation process of a company under the Companies Act?	LLB	correct	correct	Model from Mistral AI is better.
15	What are the legal requirements for a company to hold an Annual General Meeting in Mauritius?	LLB	wrong	correct	
16	What are the key provisions under the Mauritian Workers' Rights Act for protecting employees against unfair dismissal?	LLB	correct	correct	Model from Mistral AI is better.
17	Can you explain the legal requirements for paid leave entitlements under the Workers' Rights Act?	LLB	correct	correct	
18	What does the Workers' Rights Act say about the right to severance allowance in cases of redundancy in Mauritius	LLB	correct	correct	Model from Open AI is better.
19	What are the lawful risks of creating a fake Instagram profile to profit from others?	MC	correct	correct	
20	Puis-je forcer mon fils a travailler sans le payer si j'ai besoin d'aide dans mon business?	MC	correct	correct	

VI. Conclusion

Developing a chatbot tailored for the legal sector of Mauritius is a significant step in making legal information more accessible to both legal professionals and non-legal professionals. The dynamic and hybrid nature of the Mauritian legal system makes it challenging for anyone to have easy access to legal information. The large number of acts and regulations available on the Supreme Court website also adds up to this difficulty. And yet, ignorance of the law is not an excuse. There is currently no AI-enhanced tool which provides the population with fast and easy access to Mauritian laws. In an optic to tackle this shortcoming, existing AI-related solutions in other countries were reviewed and a RAG system, which leverages the use of LLMs to generate answers and real time retrieval from relevant documents, has been proposed and provided as a novel solution for the Republic of Mauritius. In our solution, we proposed using two different LLMs, gpt-4o-mini from OpenAI and open-mixtral-8x22b from Mistral AI. This enabled us to compare the answers obtained from each model. A re-ranking algorithm was also used to enhance the RAG system by analysing retrieved documents and to decide on the most appropriate documents to be used in the answer. The solution aims at addressing several issues, such as, the difficulty in accessing legal information in Mauritius for non-legal professionals, given how online answers can be too generic and how time-consuming legal research can be. Mauritius being a bilingual country, our system can also converse in both French and English. To further assess how the system performs, a legal professional provided assistance by asking relevant questions to the system and evaluating the answers obtained. While the system is very promising and improves access to legal justice in Mauritius, there are limitations that were observed, such as the heavy reliance on the quality and completeness of legal documents made available to the system. The system can be enhanced in the future by providing accurate and updated legal texts. Judicial cases can be added to the system to provide more context. This study showcased how it is possible to bring forward an innovative AI solution to bridge the gap between the public and legal information, by making it more accessible to them.

References

[1] C. Stryker and E. Kavlakoglu, 2024. "What Is Artificial Intelligence (AI)? | IBM." Accessed: Jan. 04, 2025. [Online]. Available: https://www.ibm.com/think/topics/artificial-intelligence

[2] European Parliament, 2023, "EU AI Act: first regulation on artificial intelligence," Topics | European Parliament. Accessed: Jan. 04, 2025. [Online]. Available: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[3] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter and H. M. Blau, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.

[4] E. Maksuti and I. Erbas, "The Impact of Artificial Intelligence on Education," Multidisciplinary Joint Akseprin Journal, vol. 2, no. 2, pp. 11–20, 2024.

[5] C. Sansone and G. Sperlí, "Legal Information Retrieval systems: State-of-the-art and open issues," Information Systems, vol. 106, p. 101967, May 2022, doi: 10.1016/j.is.2021.101967.

[6] F. Amato, M. Fonisto, M. Giacalone, and C. Sansone, "An Intelligent Conversational Agent for the Legal Domain," Information, vol. 14, no. 6, p. 307, 2023, doi: 10.3390/stats3030023

[7] M. Queudot, É. Charton, and M.-J. Meurs, "Improving Access to Justice with Legal Chatbots," Stats, vol. 3, no. 3, Art. no. 3, Sep. 2020, doi: 10.3390/stats3030023.

[8] J. Morgan, A. Paiement, M. Seisenberger, J. Williams, and A. Wyner, "A Chatbot Framework for the Children's Legal Centre," in Legal Knowledge and Information Systems, IOS Press, 2018, pp. 205–209. doi: 10.3233/978-1-61499-935-5-205.

[9] S. Alam, R. Pande, M. S. Ayub, and M. A. Khan, "Towards Developing an Automated Chatbot for Predicting Legal Case Outcomes: A Deep Learning Approach," in Intelligent Information and Database Systems, N. T. Nguyen, S. Boonsang, H. Fujita, B. Hnatkowska, T.-P. Hong, K. Pasupa, and A. Selamat, Eds., Singapore: Springer Nature, 2023, pp. 163–174. doi: 10.1007/978-981-99-5834-4_13.

[10] V. A. H. Firdaus, P. Y. Saputra, and D. Suprianto, "Intelligence chatbot for Indonesian law on electronic information and transaction," IOP

Conf. Ser.: Mater. Sci. Eng., vol. 830, no. 2, p. 022089, Apr. 2020, doi: 10.1088/1757-899X/830/2/022089.

[11] P. N. Devaraj, R. T. P. V, A. Gangrade, and M. K. R, "Development of a Legal Document AI-Chatbot," Nov. 21, 2023, arXiv: arXiv:2311.12719. doi: 10.48550/arXiv.2311.12719.

[12] J. Ioannidis, J. Harper, M. S. Quah, and D. Hunter, "Gracenote.ai: Legal Generative AI for Regulatory Compliance," Jun. 19, 2023, Social Science Research Network, Rochester, NY: 4494272. doi: 10.2139/ssrn.4494272.

[13] L. Cao, "DiagGPT: An LLM-based and Multi-agent Dialogue System with Automatic Topic Management for Flexible Task-Oriented Dialogue," Apr. 15, 2024, arXiv: arXiv:2308.08043. doi: 10.48550/arXiv.2308.08043.

[14] H.-T. Nguyen, "A Brief Report on LawGPT 1.0: A Virtual Legal Assistant Based on GPT-3," Feb. 14, 2023, arXiv: arXiv:2302.05729. doi: 10.48550/arXiv.2302.05729.

[15] R. P. Gunputh, 2022, "The Mauritian Legal System and Research," GlobaLex | Foreign and International Law Research. Accessed: Jan. 04, 2025. [Online]. Available:
https://www.nyulawglobal.org/globalex/Mauritius.html

[16] "Mauritius National Assembly," 2024, Acts. Accessed: Nov 20, 2024. [Online]. Available:
https://mauritiusassembly.govmu.org/mauritiusassembly/index.php

[17] S. Pudaruth, R. P. Gunputh, K. M. S. Soyjaudah, and P. Domun, "A Question Answer System for the Mauritian Judiciary," in 2016 3rd International Conference on Soft Computing & Machine Intelligence (ISCMI), Nov. 2016, pp. 201–205. doi: 10.1109/ISCMI.2016.47.

[18] S. Pudaruth, S. Soyjaudah, and R. Gunputh, "Classification of Legislations using Deep Learning", International Arab Journal of Information Technology, vol. 18, no. 5, pp. 651–662. doi 10.34028/iajit/18/5/4

[19] "History | The Supreme Court of Mauritius." Accessed: May 13, 2024. [Online]. Available: https://supremecourt.govmu.org/history

[20] "The Constitution," Mauritius National Assembly, 2018. Accessed: Jan. 04, 2025. [Online]. Available:
https://mauritiusassembly.govmu.org/mauritiusassembly/index.php/the-constitution/

[21] "Annual Report of the Judiciary 2022," Supreme Court of Mauritius, 2022. Accessed: Nov 1, 2024. [Online]. Available: https://supremecourt.govmu.org/sites/default/files/annual_reports/2023-07-17/annual-report-2022-of-the-judiciary.pdf

[22] "Office of the DPP," 2019, E-newsletter- Issue 88 Accessed: Oct 29, 2024, [Online]. Available:
https://dpp.govmu.org/Documents/Newsletter/Issue88Jan2019.pdf

[23] "MCCI," Mauritius Chamber of Commerce and Industry, 2015, Mediation in Mauritius. Accessed: Nov 1, 2024. [Online]. Available: https://www.mcci.org/media/36602/mediation-in-mauritius.pdf

[24] T. Medeiros, M. Medeiros, M. Azevedo, M. Silva, I. Silva, and D. G. Costa, "Analysis of Language-Model-Powered Chatbots for Query Resolution in PDF-Based Automotive Manuals," Vehicles, vol. 5, no. 4, Art. no. 4, Dec. 2023, doi: 10.3390/vehicles5040076.

APPENDIX 1

A Consolidated Version of the Employment Relations Act 2008 as at 27 July 2024.pdf

A Consolidated Version of the Workers' Rights Act 2019 as at 27 July 2024.pdf

Attorneys' and Notaries' Workers (Remuneration) Regulations 2019.pd

Baking Industry (Remuneration) Regulations 2019.pdf

Banking-act-2004.pdf

Banks Fishermen and Frigo-workers (Remuneration) Regulations 2019.pdf

Blockmaking, Construction, Stone Crushing and Related Industries (Remuneration) Regulations 2019_.pdf

Catering and Tourism Industries (Remuneration) Regulations 2019.pdf

Cinema Employees (Remuneration) Regulations 2019.pdf

Cleaning Enterprises (Remuneration) Regulations 2019.pdf

Code-civil-mauricien.pdf

Companies Act.pdf

Constitution.pdf

Cybersecurity-and-cybercrime-act-2021.pdf

Data-protection-act-2017.pdf

Distributive Trades (Remuneration) Regulations 2019.pdf

Domestic Workers (Remuneration) Regulations 2019.pdf

Education Act.pdf

Electrical, Engineering and Mechanical Workshops (Remuneration) Regulations 2019.pdf

Employment-relations-act-2008.pdf

Equal Opportunities Act.pdf

Export Enterprises (Remuneration) Regulations 2019 (Amended).pdf

Factory Employees (Remuneration) Regulations 2019.pdf

Field-crop and Orchard Workers (Remuneration) Regulations 2019.pdf

GN15WRA.pdf (workers' rights act 2019 legal supplement)

Immigration-act-2022.pdf

Industrial-court-act-1973.pdf

Land-duties-and-taxes-act-1984_19.pdf

Light Metal and Wooden Furniture Workshops (Remuneration) Regulations 2019.pdf

Livestock Workers (Remuneration) Regulations 2019.pdf

Newspapers and Periodicals Employees (Remuneration) Regulations 2019.pdf

Non-citizens-employment-restriction-act-1970.pdf

Occupational-safety-and-health-act-2005.pdf

Office Attendants (Remuneration) Regulations 2019.pdf

Pre-Primary School Employees (Remuneration) Regulations 2019.pdf

Printing Industry (Remuneration) Regulations 2019.pdf

Private Hospitals and Other Related Health Services (Remuneration) Regulations 2019.pdf

Private Secondary Schools Employees (Remuneration) Regulations 2019.pdf

Private Security Services Employees (Remuneration) Regulations 2019.pdf

Private-recruitment-agencies-act-2023.pdf

Public Transport (Buses) Workers (Remuneration) Regulations 2019.pdf

Registration-duty-act-1804_6.pdf

Road Haulage Industry (Remuneration) Regulations 2019.pdf

Salt-Manufacturing Industry (Remuneration) Regulations 2019.pdf

Social-contribution-and-social-benefits-act-2021.pdf

Tailoring Trade (Remuneration) Regulations 2019.pdf

The-National-Minimum-Wage-Amd-Reg-24.pdf

Travel Agents and Tour Operators Workers (Remuneration) Regulations 2019.pdf

APPENDIX 2

1. I have been in employment with ABC since January 2024. Am I entitled to special leave?

Open AI gpt-4o-mini:

Based on the previous context, you would not be entitled to special leave yet. According to the regulations, an employee must remain in continuous employment with the same employer for a period of 12 consecutive months to be eligible for special leave. Since you have been employed since January 2024, you need to wait until January 2025 to qualify for special leave.

Mistral AI open-mixtral-8x22b:

Based on the previous content, if you have been continuously employed with ABC for a period of 12 consecutive months, then you are entitled to special leave. However, as you started your employment in January 2024, you would need to wait until January 2025 to be eligible for special leave. The specific entitlements for special leave are outlined in the context, which include 6 working days' special leave on fill pay for the occasion of the celebration of your first civil or religious marriage, 3 working days' special leave on full pay for the occasion of the first civil or religious marriage of your son or daughter, and 3 working days' special leave on full pay on the death of your spouse, child, father, mother, brother, or sister.

Expected answer:

According to the legal provisions currently in force, special leave is granted to a worker who has been in continuous employment for 12 consecutive months. A worker who has been in employment for a period of less than 12 consecutive months shall not be entitled to special leave.

2. What is a labour contractor?

Open AI gpt-4o-mini:

A labour contractor is defined as a person who is engaged in the recruitment and supply of migrant worker to work with a hirer employer.

Mistral AI open-mixtral-8x22b:

A labour contractor is a person who is engaged in the recruitment and supply of a migrant worker to work with a hirer employer.

Expected answer:

A labour contractor is defined as "a person who is engaged in the recruitment and supply of a migrant worker to work with a hirer employer". A labour contractor must be registered with the Ministry of Labour. A labour contractor may supply a migrant worker to work with a hirer employer in any approved sector of economy. The Labour contractor shall remain the Employer of the migrant worker after the latter is placed at the seat of a hirer employer. Also, the labour contractor is compelled under law to keep a register of the placement of a migrant worker with each hirer employer retaining its services.

3. Je suis convoqué à un comité disciplinaire dans une semaine. Comment savoir si le comité est valablement constitué ?

Open AI gpt-4o-mini:

Pour savoir si le comité disciplinaire est valablement constitué, vous devez vérifier les éléments suivants:

Composition du Comité:

Le comité doit être composé d'au moins une personne indépendante.

Quorum :

Il doit y avoir un quorum, ce qui signifie que le nombre minimum de membres requis pour que le comité puisse prendre des décisions doit être présent.

Objections :

Si vous ou votre syndicat avez des objections concernant la composition du comité, vous devez les soumettre à l'employeur dans les 24 heures suivant l'information sur les membres nommés.

Si ces conditions sont remplies, le comité peut être considéré comme valablement constitué.

Mistral AI open-mixtral-8x22b:

Pour savoir si le comité disciplinaire est valablement constitué, vous devez vous assurer qu'il est conforme à l'article 14 de la législation en vigueur. Selon cet article, le comité doit inclure au moins une personne indépendante et être guidé par les principes de justice naturelle. En outre, en tant que travailleur, vous avez le droit de contester toute personne faisant partie du comité disciplinaire et de soumettre vos objections, ainsi que les motifs de ces objections, à l'employeur au plus tard 24 heures après avoir été informé du nom des personnes nommées pour entendre l'affaire.

Expected answer:

Il faut que le comité soit présidé par une personne indépendante. La personne convoquée devant le comité doit répondre aux charges énoncées dans la lettre de convocation et peut se faire accompagner par un représentant syndical, un homme de loi (avocat) ou les deux ou par un officier du Ministère du Travail, des ressources humaines et de la Formation. Il doit s'écouler au moins sept jours entre la notification/convocation et la tenue du comité disciplinaire.

4. Mon employeur ne respecte pas ma vie privée. Quels sont les moyens dont je dispose pour me protéger ou pour le dissuader ?

Open AI gpt-4o-mini:

Vous avez le droit de demander réparation pour les griefs liés à votre emploi. Voici quelques moyens que vous pouvez envisager :

Faire un signalement:

Vous pouvez signaler le non-respect de votre vie privée à votre supérieur hiérarchique ou à un responsable des ressources humaines.

Recours à un syndicat:

Si vous êtes membre d'un syndicat, vous avez le droit d'être représenté par un représentant syndical qui peut vous aider à défendre vos droits.

Procédures formelles:

L'employeur doit établir des arrangements avec les syndicats pour permettre aux travailleurs de soulever des griefs et de les résoudre de manière équitable et rapide. Vous pouvez demander à connaître ces procédures.

Demande d'information:

Si vous avez besoin d'informations concernant vos droits ou les politiques de l'entreprise relatives à la vie privée, vous pouvez demander ces informations à votre employeur.

Saisir le Tribunal:

Si votre employeur refuse de fournir des informations ou ne respecte pas vos droits, vous pouvez faire une demande auprès du Tribunal pour obtenir une ordonnance.

Il est important de documenter tous les incidents et de conserver des preuves de toute violation de votre vie privée.

Mistral AI open-mixtral-8x22b:

Selon le contexte fourni, il est interdit de divulguer des informations qui peuvent causer un préjudice aux intérêts de l'entreprise ou d'un travailleur, ou qui sont des informations personnelles liées à la vie privée d'un travailleur, à moins que le travailleur ne consente à la divulgation de ces informations.

Si votre employeur ne respecte pas votre vie privée, vous pouvez demander à votre employeur de cesser cette pratique. Si l'employeur refuse de cesser cette pratique, vous pouvez alors déposer une plainte auprès de l'autorité compétente, telle que le Tribunal. La loi prévoit également que le Tribunal peut ordonner à l'employeur de fournir des informations si cela est nécessaire pour les négociations collectives, à condition que les informations demandées ne relèvent pas des exceptions prévues par la loi.

Enfin, si vous estimez que votre employeur a violé vos droits en matière de confidentialité, vous pouvez également engager des poursuites judiciaires contre lui.

Expected answer:

Tout justiciable est titulaire du droit à la vie privée. Ce droit est garanti par la Constitution. De plus, le droit a la vie est reconnue dans le code civil. En cas de non-respect de la part de l'employeur, il est possible de saisir le tribunal

compétent. Il est aussi possible de rapporter le cas au ministère du Travail ou encore le Data Protection Office.

5. I intend to cease business as a Company. How can I terminate the employment of staff?

Open AI gpt-4o-mini:

To terminate the employment of staff when ceasing business as a company, you must follow these steps:

Written Notice: You must give written notice to the Director-General specifying the date on which the cessation will take effect.

Notice to Employees: You must inform your employees about the termination of their employment. According to the provisions, you must state the reason for the termination at the time of notifying them.

Notice Period: Ensure that the notice period is not less than 30 days, unless you opt to pay the employees the amount of remuneration they would have earned during the notice period.

Severance Allowance: Pay severance allowance to workers who have been in continuous employment for a period of not less than 12 months, as specified in the relevant sections.

Consultations: Consult with trade unions and provide reasonable notice to them as well, as well as to the Ministry responsible for labour.

Assistance in Job Search: Consider offering help to workers in finding other jobs in cooperation with the Ministry responsible for employment.

By adhering to these steps, you can ensure that the termination of employment is carried out in compliance with the relevant legal provisions.

Mistral AI open-mixtral-8x22b:

To terminate the employment of staff due to the cessation of business, you must follow the guidelines outlined in the provided context. Here are the key steps:

Provide reasonable or prescribed notice to the workers, to the Ministry responsible for labour and industrial relations, and to the trade unions concerned.

Consider introducing schemes for voluntary redundancy, retirement, transfer to other departments within the undertaking, and a phased rundown of employment.

Establish which workers are to be made redundant and the order of discharge.

Offer help to workers in finding other jobs, in cooperation with the Ministry responsible for employment.

Decide how and when to make the facts public, ensuring no announcements is made before the ministry, the workers, and their trade unions have been informed.

Please note that termination of employment without justification, where the ground of termination is other than economic, technological, or structural grounds, is deemed unjustified. Therefore, it's crucial to follow proper procedures and provide adequate notice to all relevant parties.

Lastly, it's important to issue a certificate of employment to each workers whose employment has been terminated or who has been resigned. The certificate should be provided within 7 days of the termination of employment and should include the specified particulars.

Expected answer:

The cessation of business is governed by the Companies Act 2001 and the Insolvency Act 2009. The legal framework regarding the termination of employment on grounds of redundancy is set in the Workers' Rights Act 2019. An employer is obliged to refer the matter to the Redundancy Board if it employs 15 persons or more. The Supervising Officer must be informed of the intended cessation of business if there are 14 persons or less.

6. Quel est le salaire minimum à ce jour dans la république de Maurice ?

Open AI gpt-4o-mini:

Le salaire minimum à ce jour dans la République de Maurice est de 16,500 roupies par mois, applicable à partir du 1er janvier 2024.

Mistral AI open-mixtral-8x22b:

Le salaire minimum à ce jour dans la république de Maurice est de 16,500 roupies.

Expected answer:

Le salaire minimum est aujourd'hui fixé à Rs. 15,000. Cependant, toute personne étant salarié se voit percevoir un salaire minimum de Rs. 16,500, ce qui inclut le Cost of Living Allowance fixe par voie d'arrêté ministériel en Janvier 2024.

7. What are my obligations as an employee?

Open AI gpt-4o-mini:

As an employee, your obligations include:

Taking Reasonable Care:

You must take reasonable care for your own safety and health, as well as for the safety and health of others who may be affected by your actions or omissions at work.

Understanding Your Contract:

You should ensure that you understand the terms of your contract of employment and abide by them.

Familiarizing with Grievance Procedures:

You need to make yourself familiar with any arrangements for dealing with grievances and other questions that may arise from your contract of employment and utilize them when necessary.

Special obligations:

If you belong to a profession, you may have additional obligations that arise from your membership, which you must fulfill.

Membership Obligations:

If you are a member of a trade union, you must fulfill the obligations you voluntarily undertook upon joining the union.

These obligations are essential for maintaining a good working relationship and ensuring a safe and productive work environment.

Mistral AI open-mixtral-8x22b:

As an employee, you have several obligations. These include:

Satisfying yourself that you understand the terms of your contract of employment and abiding by them.

Making yourself familiar with any arrangements for dealing with grievances and other questions which may arise out of your contract of employment, and making use of them as and when the need arises.

If you belong to a trade union, fulfilling the obligations which you have voluntarily undertaken on joining the trade union.

Not taking any actions in your professional capacity, which would conflict with the standard of work or conduct laid down for your profession, particularly if that action would endanger public health and safety, the health of an individual needing medical or other treatments, or the well-being of an individual needing care through personal social services.

Taking reasonable care for the safety and health of yourself and other persons who may be affected by your acts or omissions at work.

Abiding by any special obligations arising from membership of a profession, which may include obligations with regard to health, safety, and welfare, over and above those which are shared by the community as a whole.

Expected answer:

The obligations of an employee are set in a contract of employment whose terms must be in line with the provisions to be found in the Workers' Rights Act 2019. There are implied obligations which stem from the provisions of the Mauritian Civil Code- these include the duty of loyalty, the duty of

confidentiality, the obligation to perform work personally, the duty of diligence while performing his or her assignment.

8. I would like to hire the services of my nephew who is 15 years old. Can I do so?

Open AI gpt-4o-mini:

According to the provided context, you cannot employ a young person who is under the age of 16. Therefore, you cannot hire your nephew who is 15 years old.

Mistral AI open-mixtral-8x22b:

According to the provided context, a person who is of the age 16 years or more shall be competent to enter into an agreement and shall, in relation to the agreement and to its enforcement, to be deemed to be a full age and capacity. Therefore, as your nephew is 15 years old, he cannot legally enter into an employment agreement. Additionally, no person shall employ, to continue to employ, a young person in work which is likely to jeopardize his health, safety, or physical, mental, moral or social development.

Expected answer:

No. The law provides that any person having attained the age of 16 may be employed.

9. I would like to transfer shares I hold in Company FEDER to my brother. What do I need to do?

Open AI gpt-4o-mini:

To transfer shares, you hold in company FEDER to your brother, you need to follow these steps:

Prepare a Transfer Document that includes:

Your full name and address as the transferor.

Your brother's full name and address as the transferee.

The name of the Company FEDER and its registration number.

The number and description of shares being transferred.

The consideration for the transfer.

Sign the Document: Both you and your brother must sign the transfer document, certifying the transfer of shares.

Conditions of Transfer: Ensure that the transfer is accepted subject to the same conditions attached to the shares at the time of transfer.

Notify the Company: Submit the signed transfer document to Company FEDER's registered office as part of the transfer process.

Compliance with Regulations: Make sure to comply with any other regulatory specific requirements specific to share transfers in Mauritius.

It's advisable to consult with a legal professional to ensure all necessary steps and legalities are adhered to.

Mistral AI open-mixtral-8x22b:

In order to transfer shares, you hold in Company FEDER to your brother, you will need to follow the process outlined in the provided context. However, please note that the context does not provide a comprehensive guide to the share transfer process. Based on the available information, here are the general steps you need to follow:

Complete a transfer of shares form, which should include the following details:

Full Name of Transferor (you) and your full address

Full Name of Transferee (your brother) and his full address

Name of the company in which the shares are held (Company FEDER)

Company No.: (you will need to provide the company registration number)

Number and description of shares transferred (you will need to specify the number of shares and their type)

Consideration (if applicable)

Both you (the transferor) and your brother (the transferee) will need to sign the transfer of shares form.

The transfer or issue of shares is subject to the same conditions as attached to such shares. Please ensure that you understand these conditions before proceeding with the transfer.

Depending on the company's requirements, you may also need to provide additional documentation or meet certain conditions. It is recommended that you contact Company FEDER for further guidance on their specific requirements for transferring shares.

Once the transfer of shares has been completed and registered, your brother will become the new shareholder in Company FEDER.

Expected answer:

You should sign the statutory share transfer form. Then, the transfer shall be registered with the Registrar General Department. Once approved by the Registrar General Department, a certified copy of the transfer form shall be submitted to the Corporate Business Registration Department to have the records amended.

10. What does section 3 of the Constitution of Mauritius, which guarantees the right to life, liberty and security of a person, balance individual freedoms with the state's responsibility to maintain public order and safety?

Open AI gpt-4o-mini:

Section 3 of the Constitution of Mauritius recognizes and declares that fundamental rights and freedoms that exist without discrimination, but it also emphasizes that these rights are subject to respect for the rights and freedoms and for the public interest. This means that while individuals have the right to life, liberty, and security, these rights must be balanced against the state's responsibility to maintain public order and safety. Therefore, any limitations on these rights must be reasonable and justifiable in a democratic society, ensuring that the enjoyment of individual rights does not prejudice the rights and freedoms of others or the public interest.

Mistral AI open-mixtral-8x22b:

Section 3 of the Constitution of Mauritius guarantees the right to life, liberty, and security of a person, but this protection is subject to respect for the rights and freedoms of others and for the public interest. This means that while individuals have certain fundamental rights and freedoms, these rights are not absolute and may be limited to ensure that the enjoyment of these rights by an individual does not prejudice the rights and freedoms of others or the public interest.

The Constitution further provides that no person shall be deprived of his life intentionally, save in execution from a sentence in a court in respect of a criminal offence of which he has been convicted. However, a person shall not be regarded as having deprived of his life if he dies as a result of the use of such force as is reasonable justifiable for the defence of any person from violence or for the defence of property, in order to effect a lawful arrest or to prevent the escape of a person lawfully detained, for the purpose of suppressing a riot, insurrection or mutiny, or in order to prevent the commission by that person of a criminal offence, or if he dies as a result of a lawful act of war.

Therefore, the Constitution of Mauritius balances individual freedoms with the state's responsibility to maintain public order and safety by allowing for limitation on individual rights and freedoms in certain circumstances, such as defence, public safety, public order, public morality, or public health. These limitations are designed to ensure the enjoyment of individual rights and freedoms does not prejudice the rights and freedoms of others or the public interest.

11. Which section and chapter of the constitution caters for the freedom of expression?

Open AI gpt-4o-mini:

The freedom of expression is catered for in Section 12 of Chapter II – Protection of Fundamental Rights and Freedoms of the Individual of the Constitution.

Mistral AI open-mixtral-8x22b:

The freedom of expression is catered for in Section 12 of Chapter II of the Constitution.

12. What does the constitution states about the power of the President of Mauritius?

Open AI gpt-4o-mini:

The Constitution states that:

The executive authority of Mauritius is vested in the President, who may exercise this authority directly or through subordinate officers.

The President acts in accordance with the advice of the Cabinet or a Minister, except in cases where the Constitution requires the President to act according to the advice of someone else or in his own judgement.

The President has the power to appoint the Prime Minister, Deputy Prime Minister, and other Ministers, as well as Junior Ministers, acting on the advice of the Prime Minister.

The President is also responsible for keeping informed about the general conduct of the government and may request information from the Prime Minister regarding government matters.

Overall, the President plays a crucial role in the governance and administration of the country, guided by the advice of the Prime Minister and the Cabinet.

Mistral AI open-mixtral-8x22b:

The constitution states that the executive authority of Mauritius is vested in the President. Unless otherwise provided in the Constitution, that authority may be exercised by the President either directly or through officers subordinate to him. However, this section does not preclude persons or authorities, other than the President, from exercising such functions as may be conferred upon them by the law. Additionally, the President, acting in accordance with the advice of the Prime Minister, may appoint Junior Ministers from among the members of the Assembly to assist Ministers in the performance of their duties.

13. What are they key duties of a company secretary under the Mauritius Companies Act 2001?

Open AI gpt-4o-mini:

I do not have an answer.

Mistral AI open-mixtral-8x22b:

Under the Mauritius Companies Act 2001, the key duties of a company secretary are:

To ensure that the company complies with the requirement of the Act and its constitution.

To keep and maintain the minutes of all meetings of the Board and the shareholders.

To keep and maintain the register of directors and officers and the register of shareholders.

To ensure that the company's accounting records are kept as required by the Act.

To ensure that the company's annual returns are submitted to the Registrar.

To ensure that the company's name is properly displayed at its registered office and at all other places where it carries on business.

To ensure that the company's common seal is kept in safe custody.

To carry out all other duties as may be required by the Act or the company's constitution.

The company secretary may also be required to act as a director or liquidator of the company in certain circumstances.

14. Can you explain the procedure for the incorporation process of a company under the Companies Act?

Open AI gpt-4o-mini:

The procedure for the incorporation process of a company under the Companies Act is as follows:

Right to Apply: Any person may apply for the incorporation of a company, subject to the provisions of the Act.

Application submission: The application for incorporation must be sent or delivered to the Registrar.

Required Form: The application must be in a form approved by the Registrar.

Signatures: Each applicant must sign the application.

Accompanying Documents: The application must be accompanied by:

A document, in a form approved by the Registrar, signed by every person named as a director or Secretary, containing their consent to be director or Secretary.

A certificate confirming that the person is not qualified from being appointed or holding office as a director or Secretary of a company.

Once these steps are completed and the Registrar is satisfied with the application, the company will be incorporated under the Act.

Mistral AI open-mixtral-8x22b:

Incorporation Process of a Company under the Companies Act in Mauritius:

Application: The first step in the incorporation process is to file an application with the Registrar of Companies. The application should include the following:

The proposed name of the company

The company's objects of purposes

The company's registered office address

The name, addresses, and occupations of the company's directors and secretary

The name and address of the company's auditor

Payment of fees: Along with the application, the applicant must pay the prescribed fee for incorporation.

Issuance of Certificate of Incorporation: If the Registrar is satisfied that the application meets the requirements of the Act, he or she will issue a Certificate of Incorporation, which signifies that the company is duly incorporated and registered under the Act.

Post-incorporation obligations: After incorporation, the company must comply with certain post-incorporation obligations, such as holding its first board meeting, appointing an auditor, and filing an annual return with the Registrar.

Exemptions: Certain companies, such as small private companies and companies registered under the Small and Medium Enterprise Act, are exempt from filing financial summaries with the Registrar for a specified period.

15. What are the legal requirements for a company to hold an Annual General Meeting in Mauritius?

Open AI gpt-4o-mini:

I do not have an answer.

Mistral AI open-mixtral-8x22b:

Question: What are the legal requirements for a company to hold an Annual General Meeting in Mauritius?

According to the Companies Act of Mauritius, a company is required to hold an Annual General Meeting (AGM) within 18 months of its incorporation and subsequently, once at least every calendar year and not more than 15 months after the last AGM.

For a company incorporated outside Mauritius, if it is not required by the law of its place of incorporation or origin to hold an annual meeting of shareholders and prepare a balance sheet, it shall prepare and file with the Registrar a balance sheet within such period, in such form, and containing such particulars and annex thereto such documents as directors of the company would have been required to prepare or obtain if the company were a public company.

Additionally, a company is required to file an annual return with the Registrar within 28 days of the date of the annual meeting of the company or within 8 weeks after the date of the annual meeting for a company that keeps a branch register outside Mauritius. The annual return should be completed and signed by a director or Secretary.

The annual return should contain the matters specified in the Tenth Schedule provided that where matters required to be stated are in each case unchanged from the last preceding annual return, the annual return may simply refer to the last preceding annual return.

The company must also send a copy of the annual report to every shareholder of the company not less than 14 days before the date fixed for holding the annual general meeting of the shareholders. However, the Board of a company shall not be required to send an annual report to a shareholder where the shareholder has given notice in writing to the company waiving the right to be sent a copy of the annual report or copies of annual reports of the company generally, and the shareholder has not revoked that notice, and a copy of the report is available for inspection by the shareholder in the manner specified in section 227.

A public company shall deliver a copy of its annual report to the Registrar for registration at the same time as it delivers its financial statements to the Registrar under section 215.

16. What are the key provisions under the Mauritian Workers' Rights Act for protecting employees against unfair dismissal?

Open AI gpt-4o-mini:

The key provisions under the Mauritian Workers' Right Act for protecting employees against unfair dismissal include:

Assessment of Performance: Employers must assess workers on their workmanship, and this assessment should be recorded.

Reprimanding Procedure: If a worker cannot fulfil their obligations, the employer may reprimand them initially before initiating disciplinary actions. This can include a verbal warning, followed by a documented verbal warning and a written warning.

Disciplinary Actions: If no improvement is noted after a period of 6 months, the employer may take disciplinary action, which should be commensurate with the gravity of the case.

Lapse of Sanctions: Any sanctions imposed for minor misconduct shall lapse after one year if there has been no recurrence.

Protection against Dismissal for Minor Misconduct: Workers cannot be dismissed for reasons related to minor misconduct.

Complaint Procedure: Workers may make complaints against their employer regarding any employment-related issues, and the supervising officers is required to investigate these complaints.

These procedures aim to ensure their fair treatment and due process for workers facing disciplinary actions or dismissal.

Mistral AI open-mixtral-8x22b:

The key provisions under the Mauritian Workers' Right Act for protecting employees against unfair dismissal are as follow:

Now workers shall be dismissed for reasons related to minor misconduct.

Every supervising officer is responsible for providing guidance to workers for their performance at work.

Where a worker's work is reported to be unsatisfactory, the supervising offer shall inform the worker of this weaknesses and expected standards of work required of him.

The employer shall assess the worker on this workmanship and the assessment shall be recorded.

Where the worker cannot fulfil his obligations within the terms of this contract of employment, the employer may reprimand the worker in the first instance before initiating disciplinary actions in accordance with the Workers' Right Act 2019.

The worker may initially be sanctioned by way of a verbal warning followed by a documented verbal warning and a written warning.

Where no improvement is noted after a period of 6 months, the employer may, following disciplinary proceedings against the worker, take such disciplinary action which may be commensurate with the degree of gravity of the case.

Any sanction imposed under paragraphs (4) and (5) shall lapse after a period of one year where there has been no recurrence.

A worker shall be entitled to his basic wage during the period of suspension.

The worker shall be reinstated without prejudice and without loss of pay or other privileges if the alleged act of misconduct is not proved.

Any disciplinary committee set up to provide an opportunity to a worker charged with an act of misconduct to give his explanations shall, in his proceedings, be guided by the principles of natural justice.

Where the employer considers that the worker may have to answer a charge of misconduct which may lead to his summary dismissal, the employer shall appoint a disciplinary committee, consisting of at least one independent person, to hear the worker and make its recommendations.

The worker or the union may contest any person forming part of the disciplinary committee and shall submit his or her objections together with the reasons for the objections.

These provisions ensure that employees are protected against unfair dismissal and that any disciplinary action taken against them is fair and just.

17. Can you explain the legal requirements for paid leave entitlements under the Workers' Rights Act?

Open AI gpt-4o-mini:

Under the Workers' Right Act in Mauritius, the legal requirements for paid leave entitlements are as follows:

Annual Leave: Every employee who has worked for the same employer for 12 consecutive months is entitled to 20 working days of annual leave during each subsequent 12-month period of continuous employment. Employees who have worked for 6 consecutive months are entitled to one day of annual leave for each subsequent month of continuous employment, up to the twelfth month.

Sick Leave: Employees who remain in continuous employment for 12 consecutive months are entitled to 21 days of sick leave on full pay during the following 12 months. Employees who have worked for 6 consecutive months and have been present on all working days during that period are also entitled to sick leave.

Maternity Leave: Female workers are entitled to 16 weeks of leave on full pay upon providing a certified copy of the relevant Court order and a copy of the child's birth act.

Leave for Family Care: Workers can be granted leave with pay to care for their child, parents, or grandparents with healthcare-related issues, subject to certain conditions.

Notice for Leave: Employees must give at least 9 days' notice when applying for leave exceeding one day.

Payment for Unused Leave: If an employee does not take or is not granted all the leave they are entitled to, they shall be paid a normal day's wage for each day of unused annual leave at the end of the 12-month period.

Restrictions on Leave: No leave shall be granted during the period from December 20 to January 7, except at the employer's discretion.

These provisions ensure that employees have the right to take necessary leave while also providing guidelines for employers regarding the management of such leave.

Mistral AI open-mixtral-8x22b:

Under the Workers' Rights Act of 2019, every worker, other than a part-time worker, who remains in continuous employment with the same employer for a period of 12 consecutive months shall be entitled, during each subsequent period of 12 months while they remain in continuous employment, to 20 working days' annual leave.

Additionally, every worker shall be entitled to 2 days' leave in every year, in addition to the leave specified in subsection (1) or any other enactment, or any agreement, collective agreement or award of the Tribunal or an arbitrator.

The leave specified in subsection (1) or (2) may be taken on a full day, half day, or such shorter period as agreed between an employer and a worker.

Furthermore, a worker who wishes to take more than one day's annual leave consequently shall, except where reasonable cause is shown, give their employer at least 48 hours' advance written notice.

Lastly, an employer shall pay to a worker a normal day's wage in respect of each day's leave still due to them at the end of the period 12 consecutive months, where the worker has not taken or has not been granted by the employer, all the leave the worker is entitled to under the subsection (1), (2) or (5). However, during any period of 12 consecutive months, a worker has not requested for annual leave or where the request has not been granted the worker may opt to have the remaining leave accumulated and inform their employer in writing of their option

18. What does the Workers' Rights Act say about the right to severance allowance in cases of redundancy in Mauritius

Open AI gpt-4o-mini:

The Workers' Right Act 2019 provide specific provisions regarding severance allowance in cases of redundancy. Here the key points:

Justified Reasons for Redundancy: If the Board find that the reasons for the reduction of the workforce or closing down are justified, it may not order the employer to pay severance allowance.

Unjustified Reasons for Redundancy: If the reasons are found to be unjustified, the Board shall order the employer to pay the worker severance allowance at the rate specified in section 70(1).

Calculation of Severance Allowance:

For every period of 12 months of continuous employment, a worker is entitled to a sum equivalent to 3 months' remuneration.

For any additional period of less than 12 months, the worker is entitled to a sum equal to one twelfth of the amount calculated for the 12-month period multiplied by the number of months of continuous employment.

Deductions from Severance Allowance: Employers may deduct contributors payable to the Portable Retirement Gratuity Fund from the severance paid to the worker.

Interest on Severance Allowance: The Court may order an employer to pay interest on the severance allowance at a rate not exceeding 12 percent per year from the date of termination to the date of payment.

These provisions ensure that workers are compensated fairly in the event of redundancy, depending on the justification of the employer's actions.

Mistral AI open-mixtral-8x22b:

The Workers' Right Act 2019 in Mauritius provides for the payment of severance allowance in cases of redundancy. According to section 70(1), a worker is entitled to severance allowance if:

(a) the worker has been continuously employed for not less than 12 months with the employer on a contract of indeterminate duration and that employer terminates the agreement in circumstances specified in section 70(1); or

(b) the worker has been employed under on more fixed-term contracts with the same employer for a total period of not less than 12 months and the employer terminated the agreement in circumstances specified in section 70(1).

The amount of severance allowance payable is calculated based on the last complete month of the worker's employment on a full-time basis or, where applicable, a monthly rate circulated over a period of 12 months before the termination of the agreement.

However, no severance allowance shall be payable to a migrant worker or a non-citizen employed under one or more contracts of fixed duration at the expiry of their contracts. Additionally, no severance allowance shall be payable where a worker and an employer enter into an agreement under section 13(1) and the agreement comes to an end, unless otherwise agreed by the parties.

In cases where the employer intends to reduce the workforce or close down, the employer must give notice to the Board. If the Board finds that the reasons for the reduction of the workforce or the closing down are unjustified, it shall order the employer to pay severance allowance to the worker at the rate specified in section 70(1). However, if the Board finds that the reasons for the reduction of the workforce or the closing down are justified, it may order the worker to proceed to leave without pay for a specified period, instead of terminating their employment.

19. What are the risks of creating a fake Instagram profile to profit from others?

Open AI gpt-4o-mini:

Creating a fake Instagram profile to profit from others can lead to several risks, including:

Misuse of Fake Profile: Under Act 2021584, making use of a fake profile to cause harm is an offence. If convicted, the individual could face a fine not exceeding one million rupees or penal servitude for a term not exceeding 20 years.

Cyberbullying: If the actions associated with the fake profile involve harassment or intimidation, it could be classified as cyberbullying, which also carries a penalty of a fine not exceeding one million rupees and penal servitude not exceeding 20 years.

Cyber Extortion: If the profile is used to demand money or goods by threatening them, it constitutes cyber extortion, leading to similar penalties.

Fraudulent Activities: Engaging in fraudulent activities, such as deceiving others for personal gain, can result in charge under laws related to electronic fraud, which carry significant penalties.

Unauthorized Disclosure: If the profile is used to disclose personal information without consent, it could lead to legal repercussions related to unauthorized disclosure of information.

Overall, the creation and use of a fake profile can expose individuals to severe legal consequences under various cybercrime laws in Mauritius

Mistral AI open-mixtral-8x22b:

Creating a fake Instagram profile to profit from others falls under the offence of "Misuse of fake profile" as per the provided context. The risks associated with this offence are:

On conviction, being liable to a fine not exceeding one million rupees.

On conviction, being liable to a penal servitude for a term not exceeding 20 years.

20. Puis-je forcer mon fils à travailler sans le payer si j'ai besoin d'aide dans mon business?

Open AI gpt-4o-mini:

Non, vous ne pouvez pas forcer votre fils à travailler sans le payer. Selon le contexte fourni, un enfant ne doit pas être considéré comme employé pour un travail qui pourrait nuire à sa santé ou à son développement. De plus, toute forme de travail doit être appropriée et ne pas interférer avec son éducation.

Mistral AI open-mixtral-8x22b:

Non, vous ne pouvez pas forcer votre fils à travailler sans le payer. Le droit du travail en vigueur à Maurice stipule que tout travailleur doit être payé pour son travail. De plus, le salaire doit être payé à intervalles mensuels, sauf si les parties conviennent d'un paiement à des intervalles plus courts.

Model for Training and Predicting the Occurrence of Potato Late Blight Based on an Analysis of Future Weather Conditions

Daniel Damyanov, Ivaylo Donchev
Department of Information Technologies, Veliko Tarnovo University, Bulgaria

***Abstract*—Plant diseases pose a significant challenge to agriculture, leading to serious economic losses and a risk to food security. Predicting and managing diseases such as potato blight requires an analysis of key environmental factors, including temperature, dew point, and humidity, that influence the development of pathogens. The current study uses machine learning to integrate this data for the purpose of early detection of diseases. The use of local weather data from sensors, combined with forecast data from public weather API servers, is a prerequisite for accurate short-term forecasting of adverse events. The results highlight the potential of predictive models to optimize prevention strategies, reduce losses and support sustainable crop management. Machine learning provides powerful tools for analyzing and predicting data related to plant diseases. Combining different approaches allows the creation of more precise and adaptive models for disease management.**

***Keywords*—Machine learning; potato late blight; data analysis; forecast; prediction models**

I. Introduction

Vegetable diseases represent one of the most serious challenges facing modern agriculture. They affect yields, production quality and economic sustainability of farms [21]. Caused by various pathogens, such as fungi, bacteria, viruses and nematodes, these diseases not only reduce the amount of food produced but also increase the cost of its production due to the need for treatment, prevention and control [1].

A. Economic Importance

Vegetable diseases lead to significant financial losses for both small producers and large agricultural companies [19]. They can:

- reduce yields by up to 30-50% in serious epidemics.
- oblige farmers to use expensive fungicides and pesticides, which increases production costs.
- reduce the quality of production, making it unfit for sale or export.

B. Agriculture Importance

From a business management perspective, vegetable diseases can lead to:

- Loss of competitiveness in the market due to a decline in the quality of production.
- Increased demand for resistant vegetable varieties, which are often more expensive to grow.
- Growing needs for the implementation of modern technologies for disease monitoring and control.

An example of a disease of high economic and agriculture importance is potato blight, which affects both tomatoes and potatoes [16]. In the absence of appropriate measures, it can destroy entire crops in a short time, leading to serious socio-economic consequences.

The development of effective disease management strategies, including risk forecasting through meteorological and agrotechnical data, is essential to ensure a sustainable and productive vegetable growing system. So, the purpose of this study is to improve contemporary models by adding hourly monitoring of the parameters which are important for the development of late blight and to issue timely warnings on this basis.

II. Methodology

A. Objectives of the Current Work

The current study uses environmental and meteorological data to create predictive models for the occurrence of plant diseases, with an emphasis on potato blight in tomatoes. Many models do not consider changes in real-time weather and early warning of disease occurrence but mainly use historical data [18]. Others analyze already appeared plant diseases based on machine learning from photos [11]. Thus, these models may not respond to short-term changes in conditions that can affect the development of the disease and will not trigger the preventive measures that can be taken to minimize losses.

The aim of the current work is to create a dynamic model that is constantly updated with new meteorological data (such as temperature, humidity, wind, etc.) in order to give forecasts for a short-term interval and to predict possible dangerous agronomic events.

Many of the existing models do not include the effects of carrying pathogens over distances, which can be important for widespread diseases such as potato late blight. In addition to basic data, it must take into account winds and other factors that contribute to the spread of spores over certain distances. This approach can combine weather data with diffusion patterns to predict the risk of new infections in different parts of the region or even on neighboring farms. The model should

be able to predict the development of blight in the short term, for example, in the next 24-48 hours, using different sensors for a specific location, rather than global weather data.

The methodology includes the following stages:

- **Data collection**: Collection of historical meteorological data (temperature, humidity, dew point, precipitation) and disease records from open and controlled farms.
- **Preliminary data preprocessing**: Filling in missing values, normalization and identification of critical variables for disease development.
- **Modeling**: Using machine learning algorithms, such as logistic regression, decision trees, and neural networks, to predict favorable conditions for infection.
- **Model Evaluation**: The accuracy and performance of models are evaluated through metrics such as accuracy, sensitivity, and Area Under the Curve (AUC).
- **Validation and deployment**: The results are validated on field data, then the models are integrated into practical tools for agricultural applications.

This methodology aims at early warning and effective management of risks associated with plant diseases.

B. Potato Late Blight

Potato late blight (Phytophthora infestans) is one of the most significant diseases of tomatoes and potatoes, which annually causes serious damage to agriculture. There are many scientific studies in the field of agriculture that describe in detail the situations of the onset of the disease [10]. Favorable conditions are temperatures in the range of 10℃ to about 22℃, humidity above 75%, cloudy and/or rainy, foggy weather [24]. Also, there are side factors that have no less influence on development: light, dew point, wind speed, sunshine, etc., having a strong direct relationship with the pathogen of potato late blight and its development [6], [7].

1) Temperature conditions for occurrence: The temperature ranges in which the appearance of sporangia is favored are between 12-15℃, and for the growth and spread of infection – 20-22℃. At lower temperatures (about 5-6℃), zoospores remain mobile for up to 22 hours, which increases the time for possible infection. Conversely, at higher temperatures (e.g. 15-16℃), the period of mobility is significantly shorter, since accelerated metabolic activity leads to faster germination or depletion of energy, while high temperature, i.e. 24-25℃, reduces their mobility [12]. The optimum temperature for the development of sporangia is 16-24℃. For the reproduction of sporangia, a temperature of 19-22℃ is required. At temperatures above 26℃, the disease stops its development.

Sporangia are formed at high humidity and dispersed at high temperature and low relative humidity. The release of spores is mainly due to changes in humidity. Epidemic conditions are favored mainly by humidity, that is, the prolonged survival of sporangia requires high relative humidity. The development of the disease also depends on the presence of drops of water on the surface of the leaves. Retention of water droplets on the leaves for several hours [23]. In the absence of water vapor, air spores lose their viability. At higher air speeds, multiple spores are formed at 100% relative humidity [7]. Sporulation is also increased by high humidity associated with the dew point [15]. Sporulation in the presence of strong daylight is inhibited during the day. Spores are formed only at night, when temperature and humidity favor sporulation. Sporangia germinate by releasing zoospores at a low temperature, i.e. 12-14℃, while at high temperature (20-26℃) direct germination takes place. Sporulation is stimulated by high humidity near the leaves, which is also associated with surface soil moisture.

C. Analysis Models

There are many well-studied models for predicting the occurrence of potato late blight. They are mainly divided into two groups: empirical and analytical. Of the empirical ones, the most famous are that of Van Everdingen or the so-called "Dutch rules" – based on a quantitative assessment of temperature and humidity to predict initial infections, Wallin's model focuses on the degree of danger through daylight hours during which humidity remains high. Boork's system or also known as "Irish Rules" (IR) considers the duration of humid conditions and the presence of free water in the air to be decisive factors for the development of the pathogen [3]. On the analytical side are those of Fry's model [5] – analyzes critical temperatures and moisture conditions, including an assessment of the genetic resistance of plants, Hartill-Young [8] – it is based on mathematical dependencies that assess the development of the pathogen under specific weather conditions, Baker-Lake [2] – uses modern statistical techniques to determine the time for maximum risk, taking into account historical and current weather data, etc.

One of the important aspects of using machine learning is its ability to adapt to different real-world conditions. For example, models can adapt to changing weather conditions and the specific needs of different regions, considering local climatic differences. In addition, machine learning can combine various data sources, such as satellite imagery, weather forecasts, and agricultural sensor data, allowing for a more comprehensive and accurate assessment of potato blight development conditions.

Most models include an assessment of factors such as relative humidity, length of period of high humidity, temperatures and the likelihood of rain. For example, SimCast is another established model that includes weather data such as temperature, humidity, and genetic resistance of the host to assess the risk of infection [9].

The advantage of empirical models is that they are easily understandable, applicable and require a small amount of data. They are practice-oriented. Analytical ones can be said to be highly accurate, adaptable and predictive. The negatives, on the other hand, in the former are limited precision, lack of universality, sensitivity to local data, empirical models often do not include the biological aspects of infection, such as the

latent periods of the pathogen, while in the analytical ones are complexity of algorithms, large volume of data, over-calibration. A third conclusion can be drawn, which applies to both approaches to data analysis, namely: neither category sufficiently integrates economic or environmental aspects that are important for sustainable disease management, many of the models do not take into account climate changes that can affect the prevalence and intensity of potato late blight.

D. Data Collection and Analysis

For the purposes of the ongoing development, data were collected from a local station made with Arduino architecture and temperature and humidity sensors DHT22, barometric pressure sensor BMP280**,** precipitation sensor, wind force and wind direction sensor Wind Speed Sensor RS485. Data collection, hardware architecture is borrowed from [20]. The data were collected for four months from the beginning of May to September 2024, and were collected every hour, and the placement of the sensors was in Central Northern Bulgaria, as vegetables that are susceptible to potato blight were also grown there. The goal is to track all the data and based on the selected prediction algorithm, to give an immediate signal of potential danger. The collected data is summarized in a file and distributed by columns, which are presented in Fig. 1.

date	temp	dew	humidity	precip	windspeed	winddir	pressure	cloudcover	results
2024-05-01 00:00:00	10	10.3	92.41	0	3.9	27	1025	72.4	0
2024-05-01 01:00:00	11	9.9	86.98	0	5	42.8	1024	93.2	0
2024-05-01 02:00:00	11.7	10	89.31	0	3.2	7.4	1024	96.3	0
2024-05-01 03:00:00	11.5	10.6	94.2	0	2.2	18.7	1024	93.1	0
2024-05-01 04:00:00	11.2	10.8	97.38	0	2.5	355.9	1023	95.2	0
2024-05-01 05:00:00	10.9	10.8	99.34	0	2.2	347.5	1023	94.1	0
2024-05-01 06:00:00	10.8	10.8	100	0	2.5	354.4	1022	90.7	0
2024-05-01 07:00:00	10.8	10.8	100	0	1.1	300.4	1022	85	0
2024-05-01 08:00:00	12.1	12	99.34	0	2.2	292.1	1022	83	1

Fig. 1. Collected data.

The last column, results, reflects the different states of development or stagnation of the sporangia. The calculations in it are made with logical checks according to the most described meteorological prerequisites for the development of potato late blight. They are divided into four categories:

- 0 – No blight conditions exist.
- 1 – Appearance of blight (favorable conditions). It is assumed that the temperature for the development of sporangia is between 12-15℃, the relative humidity is about 90%, the retention of water droplets on the plants is at least one hour or more, and this is monitored by dew. The formula by which the dew point is calculated is:

$$T_{dew} = T - \left(\frac{100 - RH}{5}\right) \quad (1)$$

where T_{dew} is the dew point in degrees Celsius.

T is the current air temperature (in degrees Celsius).

RH is the relative humidity of the air in percentage (from 0% to 100%). Another important condition is that these parameters are meaningful during the dark part of the day between 10 p.m. and 6 a.m.

Case favorable disease conditions:

$$FDC(t) = \begin{cases} 1, \ \sum_{i=n-10}^{n} a_i \ IF(IF \ (temp \geq 12 \ AND \\ temp \leq 15 \ AND \\ humidity \geq 85 \\ AND \ temp - dev < 2 \\ AND \ date \geq 20 \\ AND \ date \leq 7)) \geq 2 \ OR \ \leq 10 \\ 0, \ else \end{cases} \quad (2)$$

- 2 – Late blight development (development conditions are met). The air temperature should be between 15-20°C, high relative humidity above 80% and a difference between humidity and dew point less than 4 and the higher the relative humidity, the higher the dew point – hence we have water vapor condensation [14].

  Case develops disease conditions:

$$DDC(t) = \begin{cases} 1, \ IF(temp \geq 15 \ AND \\ temp \leq 20 \ AND \\ humidity \geq 80 \\ AND \ temp - dew \leq 4) \\ 0, \ else \end{cases} \quad (3)$$

- 3 – Stopping development (temperature >26°C or low humidity).

  Case block disease conditions:

$$\text{BDC(t)} = \begin{cases} 1, \ IF(temp \geq 26 \ OR \ humidity \leq 60) \\ 0, \ else \end{cases} \quad (4)$$

E. Model for Analysis and Learning

The considered way of analyzing input data and predicting future data involves several points. As it is known from the pathology of diseases, data are needed for appropriate temperature range, development time and moment of day, but also very important are the wind and moisture retention on the plants, combined with high humidity for the spread of spores. Chosen machine learning method is LightGBM [4] which is optimized to be very fast when processing large amounts of data. It uses a unique Leaf-wise tree growth technique, which allows it to achieve high accuracy and quickly process data, even with millions of samples. Although standard Gradient Boosting can be slow and resource-intensive, LightGBM is much more efficient and economical. The model includes multiple iterations, in which decision trees are built sequentially to improve prediction accuracy. LightGBM uses a gradient-boosting framework that combines weak learners (usually decision trees) into one strong predictive model. The basic concept relies on minimizing the loss function, which quantifies the error in the predictions.

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{j=1}^{K} \Omega(f_j) \quad (5)$$

L(θ) is the total loss, $l(y_i, \hat{y}_i)$ is the loss incurred by forecasting instead of the true value y_i, n is the number of samples, K is the number of trees, and Ωf_j is a normalization term that penalizes the complexity of the model to avoid refitting [13]. In each iteration, LightGBM creates a new tree

based on the negative gradient of the loss function, which shows how to adjust the predictions to minimize error. The gradient g_i for each i is calculated as

$$g_i = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i} \tag{6}$$

The new tree is trained on these gradients. This allows the model to focus on the areas where the most significant mistakes are made. LightGBM is a new way of innovative approach mostly known as leaf-wise tree growth. This way contrasts with the traditional level-wise growth of decision trees. Leaf-wise approach algorithm increases the tree by selecting the leaf with the maximum delta loss and expanding that leaf, which approach allows deeper and more informative trees [13].

Many of the models for predicting plant diseases, such as potato late blight, include various weather and environmental factors (temperature, humidity, wind, dew point, etc.). Using LightGBM in current development has the following advantages:

- It can handle high-dimensional and unstructured data well.
- It provides high-precision forecasts that can help farmers make informed decisions about plant protection methods.
- It processes large amounts of data for different weather conditions, such as those from sensors or weather forecasts.

The initial setup used the Macro Accuracy approach for evaluating the performance of multiclass classification models [22] and Area Under the Curve [17]. They assume that the closer the learning outcome is to 1, the more accurate predictive results the model will give us. If the results are below 0.5, it means that the training does not have the ability to correctly predict and analyze the input-output data.

The rows analysis is done over the column “results” considering “date”, “temp”, “dew”, “humidity”, “precip”, “windspeed”, “cloudcover”. The duration of the training, which is set, is five minutes, and can be changed depending on the results achieved. After training the model, a result of 0.9873 success rate was achieved, which is an excellent result.

III. Results

On using machine learning models (such as LightGBM and other classification algorithms), the results show high accuracy in predicting the risk of occurrence and development of late blight. After training the model and submitting sample input data for analysis, the model proposed here generates the results shown in Fig. 2. Working on the sample input data from Fig. 2, the model shows a 74% probability of situation 1 to occur – prerequisites for the development of late blight, which is a reliable result from the initial conditions for development. Thus, the data with which the model was tested can be replaced with the current input data from the sensors and those from an external API that provides information about the development of the weather conditions in the next hour. Input data can be very easily summarized and analyzed, returning the final information about the development of late blight.

```
Date: 2024-05-01 01:00:00
Temp: 15
Dew: 14
Humidity: 90
Precip: 0
Preciptype:
Windspeed: 5
Winddir: 42,8
Sealevelpressure: 1024
Cloudcover: 93,2
Results: 0
Class                                   Score
-----                                   -----
1                                       0,74
0                                       0,26
3                                       0,00
2                                       0,01
```

Fig. 2. Model results.

These models can process large volumes of data, including data on temperature, humidity, dew point, wind, and other weather conditions. By analyzing these factors, machine learning models are able to identify complex dependencies and predict with high accuracy when potato late blight is most likely to occur or develop. With the subsequent integration of the model, real-time data will be taken from the weather sensors, and an analysis will be made according to the current data, as well as those from the external API for forecast values in the coming hours. The collection of data from local measuring sensors will be monitored in the last hours to see if there are conditions for the development of the disease, as well as from the expected next few hours for the development of weather conditions. When collecting and summarizing all this data, the system will automatically make a guess and issue a warning about a possible occurrence. The created analytical model from the data with which it is trained will have to make the most accurate assumption about whether there are prerequisites for the development of the pathogen.

In order to check the effectiveness, during the construction of the model, an experimental field was planted in April 2024, divided into two groups – experimental and control. The experimental group consists of tomato plantations, on which agronomic measures will be applied in case of alarm from the model – spraying with a copper-based fungicide – "Carial Star" – Syngenta. The choice of such a fungicide is justified by the desire to minimize the harmful impact of modern fungicides on the environment, as well as on plants and human health, especially those with systemic and translaminar action. The second group includes the control plants, on which no treatment will be applied. For even more precise results, several varieties of tomatoes have been used. In both groups, tomatoes of the variety "Ideal", "Pink Magic", "Rugby" were planted, 10 of each variety. The aim is to check how different varieties react to environmental changes and whether there is a difference between them. The planting of both groups is done at a distance of 3 meters, in order not to transfer the aerosol from the fungicide to the border plants of the control group, and at the same time to keep the development conditions of both groups as similar as possible. The treatment strategy is to take measures only if an alarm is received. The tomatoes were planted on April 29, and they were not treated until a signal

appeared. The first alarm was received on 2024-05-02 between 4 and 8 a.m. The conditions were suitable, with an average temperature of 13.1℃ and humidity of 90.4%. Before treatment, the plants are in good condition, no changes, necrosis or plaques are noticeable on the leaves. According to the manufacturer's recommendations, the treatment interval is between 7-10 days, and we decided that the next spraying should be on the seventh day, if the weather forecasts remain unfavorable. On May 9, the code "1" for favorable development is again received, and treatment with the same fungicide is applied again. In the control group, it was noticed that a coating appeared on the leaves of several tomato roots (of all types). After two days, necrotic brown spots were noticeable on the petioles of the leaves on the plants of the control group. Several of the plants have watery spots on the leaves themselves, and several others on the stems themselves. In the next week, the weather conditions remain unfavorable, and there is precipitation, which further complicates the situation. On May 17, 8 plants from the control group are visibly sick, with a lot of necrosis in all parts, the flower that has sprouted falls off. In the plants treated, weak concentric spots have appeared on several of them, and this is probably due to the poor coating with the contact fungicide. On May 25, these spots were visibly calcified, and the spread was limited only to the original area. In the control group, things continue to develop in an unfavorable direction. There are apparently plants that are not alive, and this implies their removal, in others the development of the pathogen has continued. Of all the plants of each variety, there are 3-4 that do not have very serious damage. It is impossible to say whether there is a variety that is more tolerant of potato blight, but certainly the damage from lack of treatment in the risk periods has drastically changed the situation in both groups.

IV. DISCUSSION

Despite significant progress in the development of models for predicting plant diseases, such as potato late blight, there are still many uncertainties and challenges related to the accuracy and adaptability of these models to different agronomic conditions. The models commonly used to predict the development of potato late blight can be categorized as empirical and analytical. While they provide useful information, there are also some limitations that need to be considered when developing new and improved forecasts.

Empirical models, such as those of Van Everdingen, Wallin and Boork, rely on pre-accumulated observations and are mainly based on specific climatic and agronomic conditions related to temperature, humidity and wind. They provide quick and easy-to-use forecasts, but their main weakness is that they do not sufficiently take into account the differences in the microclimate of different regions. For example, area-specific conditions (such as wind, latitude, moisture levels and other factors) can affect the results of forecasts, making them less accurate in extreme weather conditions. Such models also cannot easily adapt to changes in climatic conditions or to emerging pathogens that have not been predicted in historical data.

Analytical models such as Fry, Hartill-Young, and Baker-Lake are based on mathematical and statistical approaches that allow for more accurate forecasts when new climate data is added. These models try to simulate the mechanism of disease development based on a deeper understanding of the biology of the pathogen and its interaction with the environment. While these models provide better accuracy, they require significantly more input and complex computational time, which can be a barrier to their mass deployment and practical application.

Some of the main challenges in predicting diseases such as potato late blight include dynamic microclimate conditions, frequently changing weather conditions, and interactions between various factors, such as temperature, humidity, wind, and precipitation. In addition, the dew point, which is essential for the development of late blight, is not always correctly calculated in existing models, which can lead to errors in forecasts.

Using machine learning (ML) offers new opportunities to address these challenges. With its adaptive learning and ability to train with large volumes of data involving multiple parameters, machine learning can extract data dependencies that traditional models cannot capture. The ability to analyze historical data as well as real-time weather forecasts gives machine learning models a significant advantage over traditional forecasting methods. Predictions that are based on such algorithms can be more accurate and adaptable, being able to take into account multiple external factors simultaneously and be updated in real time.

V. CONCLUSION

Predicting the development of potato late blight is crucial for sustainable agriculture, as it allows farmers to make informed decisions about disease prevention and treatment. Existing models, including empirical and analytical, have provided valuable guidance for determining the conditions under which the disease can occur or develop. However, these models often have limitations in terms of accuracy and adaptability to changing climatic and agronomic conditions.

Machine learning models that use large volumes of data and can adapt to different conditions offer new opportunities to improve predictions of the development of potato late blight. They provide more accurate and dynamic forecasts while taking into account multiple factors such as temperature, humidity, wind, and dew point. However, the success of these models depends on the quality and reliability of the data used, as well as on their proper integration into agronomic practice.

Despite these advantages, it is important to note that the use of machine learning requires a high degree of trust in the quality and accuracy of the data. Incorrect or incomplete data can lead to errors in forecasts and reduce the effectiveness of models. Therefore, for the successful application of machine learning in potato late blight forecasting, it is necessary to work with reliable and diverse data sources and ensure continuous updating and updating of models.

Existing models for predicting potato late blight have their advantages, but they are also limited in their accuracy and adaptability to changing conditions. Machine learning offers significant opportunities to improve these predictions, allowing for more accurate prediction of the conditions for the onset and development of the disease. However, to achieve reliable and

practical results, it is important to combine different approaches and technologies and to ensure the quality of the data used to train the models.

In the future, in order to achieve better accuracy and efficiency, different approaches must be combined, including traditional models and machine learning technologies. This will allow farmers not only to predict the development of late blight, but also to optimize their efforts to prevent and control the disease, which will ultimately lead to better results in agricultural production and a reduction in disease losses.

REFERENCES

[1] Agrios, G. N., Plant Pathology, 5th edition, Elsevier, 2004, 952 pages

[2] Baker, K.M., Lake, T., Benston, S.F., Trenary, R., Wharton, P., Duynslager, L., Kirk, W., Improved weather-based late blight risk management: comparing models with a ten year forecast archive, The Journal of Agricultural Science, Volume 153, Issue 2, 2014, pp. 245-256

[3] Cucak, M., Sparks, A., Moral, R.d.A., Kildea, S., Lambkin, K., Fealy, R., Evaluation of the 'Irish Rules': The potato late blight forecasting model and its operational use in the Republic of Ireland. Agronomy 2019, 9, 515. https://doi.org/10.3390/agronomy9090515

[4] Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X., Zeng, W., Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data, Agricultural Water Management, Volume 225, 2019, 105758, https://doi.org/10.1016/j.agwat.2019.105758

[5] Fry W.E., Apple A.E., Bruhn J.A., Evaluation of potato late blight forecasts modified to incorporate host resistance and fungicide weathering. Phytopathology vol 73 No. 7, 1983, pp. 1054–1059, DOI: 10.1094/Phyto-73-1054

[6] Harrison, J.G., Lowe, R., Effects of humidity and air speed on sporulation of Phytophthora infestans on potato leaves, Plant Pathology vol 38 (4), 1989, pp. 585-591

[7] Harrison, J.G., Effects of the aerial environment on late blight of potato foliage - A review, Plant Pathology vol. 41 (4), 2007, pp. 384 - 416

[8] Hartill, W.F.T., Young, K., Recent New Zealand studies on the chemical control of late blight of potatoes. In: Hill GD, Wratt GS ed. Potato growing - a changing scene. Agronomy Society of New Zealand, 1985, pp. 55-60

[9] Henderson, D., Williams, Chr.J. , Miller, J., Forecasting late blight in potato crops of Southern Idaho using logistic regression analysis, Plant disease, vol 91 (8), 2007, pp. 951-956, DOI: 10.1094/PDIS-91-8-0951

[10] Hjelkrem, A., Eikemo, H., Le, V.H., Hermansen, A., Nærstad, R., A process-based model to forecast risk of potato late blight in Norway (The Nærstad model): model development, sensitivity analysis and Bayesian calibration, Ecological Modelling, vol 450, 2021, 109565

[11] Javidan, S., Banakar, A., Rahnama, K., Vakilian, K., Ampatzidis, Y., Feature engineering to identify plant diseases using image processing and artificial intelligence: A comprehensive review, Smart Agricultural Technology, Volume 8, August 2024, 100480, https://doi.org/10.1016/j.atech.2024.100480

[12] Johnson, S. B., Potato Late Blight, Bulletin #2441, University of Maine Cooperative Extension Publications, 2020, https://extension.umaine.edu/publications/2441e/

[13] Kumar, S., Sohail, M., Jadhav, S., Gupta, R., Light gradient boosting machine for optimizing crop maintenance and yield prediction in agriculture, ICTACT Journal on Soft Computing, vol 15 (2), 2024, pp. 3551-3555

[14] Lawrence, M., The relationship between relative humidity and the dewpoint temperature in moist air: A simple conversion and applications, Bulletin of the American Meteorological Society, vol 86(2), 2005, pp. 225–234

[15] Napper, M., Observations on potato blight (Phytophthora Infestans) in relation to weather conditions, Journal of Pomology and Horticultural Science, Vol 11 (3), 1933, pp. 177-184

[16] Nowicki, M., Foolad, M., Nowakowska, M., Kozik, E., Potato and tomato late blight caused by phytophthora infestans: An overview of pathology and resistance breeding, Plant Disease, vol 96(1), 2012, pp. 4-17

[17] Powers, D., The problem of Area under the curve, IEEE International Conference on Information Science and Technology, Wuhan, China, 2012, pp. 567-573, doi: 10.1109/ICIST.2012.6221710

[18] Rahman, M., Yeasmin, T., Islam, J., Mahmud, D., Rahman, Mahb., Potato leaf disease prediction: A machine learning perspective, Journal of Scientific and Technological Research, Vol.5 (1), 2023, pp. 27-35

[19] Strange, R., Scott, P., Plant Disease: A Threat to Global Food Security, Annual Review of Phytopathology, vol. 43, 2005, pp. 83-116

[20] Saraswathi, V., Sridharani, J., Saranya, Ch., Nikhil, K, Sri, H., Mahanth, S., Smart farming: The IoT based future agriculture, 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2022, pp. 150-155

[21] Savary, S., Willocquet, L., Pethybridge, S., Esker, P, McRoberts, N., Nelson, Andy., The global burden of pathogens and pests on major food crops, Nature Ecology & Evolution 3(3), 2019, pp. 430–439

[22] Suhaimi, N., Othman, Z., Yaakub, M., Comparative analysis between macro and micro-accuracy in imbalance dataset for movie review classification, In: Yang, XS., Sherratt, S., Dey, N., Joshi, A. (eds) Proceedings of Seventh International Congress on Information and Communication Technology. Lecture Notes in Networks and Systems, vol 464. Springer, Singapore, 2023, pp.83-93

[23] Wallin, J. R., The production and survival of sporangia of Phytophthora infestans on tomato and potato plants in the field, Phytopathology, Vol. 43, No. 9, 1953, pp. 505-508

[24] Weille, G. A., Forecasting crop infection by the potato blight fungus: a fundamental approach to the ecology of a parasite-host relationship, Staatsdrukkerij- en Uitgeverijbedrijf, 's-Gravenhage, 1964, 144 pages

Lung Parenchyma Segmentation Using Mask R-CNN in COVID-19 Chest CT Scans

Wilmer Alberto Pacheco Llacho, Eveling Castro-Gutierrez, Luis David Huallpa Tapia
Universidad Nacional de San Agustín de Arequipa, Peru

***Abstract*—During the COVID-19 pandemic, the precise evaluation of lung impairments using computed tomography (CT) scans became critical for understanding and managing the disease; however, specialists faced a high workload and the urgent need to deliver fast and accurate results. To address this, deep learning models offered a promising solution by automating lung identification and lesion localization associated with COVID-19. This study employs the semantic segmentation technique Mask R-CNN, integrated with a ResNet-50 backbone, to analyze CT scans of COVID-19 patients. The model was trained using an annotated dataset, enhancing its ability to accurately segment and delineate the lung parenchyma in CT images. The results showed that Mask R-CNN achieved a Dice Similarity Coefficient (DSC) of 93.4%, demonstrating high concordance between the segmented areas and clinically relevant regions. These findings highlight the effectiveness of the proposed approach for precise lung tissue segmentation in CT scans, enabling quantitative assessments of lung impairments and providing valuable insights for diagnosis and patient monitoring.**

***Keywords*—Mask R-CNN; ResNet-50; computed tomography; lung parenchyma; COVID-19**

I. Introduction

The COVID-19 pandemic posed significant challenges in the diagnosis and management of respiratory diseases, particularly in the accurate assessment of lung conditions. In this context, medical imaging has become a fundamental tool, as its primary objective is to generate meaningful data to analyze the physiology and anatomy of various organs or areas of the human body. Among the most prominent modalities addressing these needs is simple computed tomography (CT). Based on X-ray emission, this technique, considered the oldest within medical imaging, enables non-invasive analysis of the internal structures of the human body with high precision and accuracy [1], [2], [3].

Medical imaging emerged as an indispensable tool to address these challenges. Unlike previous viral pandemics, where its use was more limited, imaging modalities have taken on a central role by enabling the rapid and accurate identification of lung patterns characteristic of the disease [4]. Furthermore, medical imaging is essential for detailed follow-up in cases where the disease becomes complicated, allowing for monitoring of progression and evaluation of the response to clinical interventions [5].

X-rays were one of the tools that emerged during the COVID-19 pandemic due to the limitations in the sensitivity of PCR tests and their lack of immediate availability. Their use enabled early diagnoses. Meanwhile, computed tomography (CT), widely available in developed countries, proved to be highly effective in precisely detailing the condition of the lung parenchyma [6]. Particularly, CT has been fundamental in evaluating lung changes such as ground-glass opacities (GGO), consolidations, and pleural thickening. These features significantly enhance diagnostic accuracy when combined with RT-PCR tests [7] and also allow for monitoring the progression of the disease [7], [8].

However, the manual inspection of these images requires a high level of expertise, specialized human resources, and considerable time, which poses a significant challenge during high-demand situations. In this context, deep learning techniques, particularly convolutional neural networks (CNNs), have emerged as highly effective tools for automating disease detection through the analysis of medical images [9].

The main objective of this study is to segment the pulmonary parenchyma in COVID-19 patients using the Mask R-CNN technique, evaluating the performance of the ResNet50 backbone under different learning rate hyperparameter configurations. Additionally, the impact of data augmentation on the model's accuracy is analyzed, implementing various augmentation techniques to optimize results. This work aims to provide an accurate and efficient tool to assist specialist doctors in identifying and assessing affected areas in lung tissue, thus contributing to a more precise diagnosis and better clinical care.

This article is organized as follows: Section II - Literature Review, Section III - Methodology, Section IV – Results and Discussion, and the final Section V presents conclusions and future work.

II. Literature Review

The authors in study [10] used Mask R-CNN for the diagnosis of COVID-19 through chest X-ray images. This deep learning approach achieved an accuracy of 96.98%, standing out for its efficiency and superiority compared to other artificial intelligence techniques, proving to be accurate and robust in identifying the disease from these images.

D. Suganya and R. Kalpana (2023) [11] employed Mask R-CNN to classify chest CT images and differentiate between COVID-19 positive and negative patients. This model, designed to identify infection regions in the lungs, achieved a mean average precision (mAP) of 91.52% and a classification accuracy of 98.60%, making it an effective solution for the accurate diagnosis of COVID-19.

S. Aparna, et al. (2021) [12], introduced a Mask R-CNN-based model with the ResNet50 architecture to analyze dental

X-rays and estimate the level of filling performed on teeth. Using a dataset of different types of fillings, they trained the model, which performs pixel-level classification, improving the accuracy in diagnosing dental treatments. This approach enables machines to perform automated dental procedures by understanding in more detail the exact region and position of the treatment.

The study by Al Masarweh and colleagues [13] proposed a Mask R-CNN-based method to autonomously generate reference data in MRI images through instance segmentation. Their model achieved a mean average precision (mAP) of 98% for locating and identifying discs, along with a 70% accuracy in classifying regions of interest. This approach allows radiologists to automate the detection of relevant areas in MRI images, improving efficiency and reliability in medical diagnoses, as well as advancing automatic medical image segmentation with cutting-edge neural network technologies.

In study [14] used the ResNet-50 deep learning architecture to classify and detect human sperm heads, achieving an accuracy of 96.66%. This proposed model proved to be efficient in identifying healthy sperm, which are used in the process of intracytoplasmic sperm injection (ICSI). The automation of the process allows for faster and more accurate results, minimizing human errors and improving success rates in infertility treatment.

S. Suriyavarman and A. X. A. R. Annie proposed an algorithm based on the combination of U-Net and Efficient-Net neural networks for the segmentation and classification of lung nodules in CT images (2023) [15]. Using semi-supervised learning with a feature pyramid network (FPN) and the ResNet-50 model for feature extraction, they were able to predict unlabeled nodules. The U-Net technique, with its skip connections, allows for precise nodule localization, while Efficient-Net optimizes the scaling of depth, width, and resolution. Evaluated on the LIDC-IDRI dataset, the model achieved an accuracy of 91.67%, outperforming most existing methods and addressing issues such as high false positive rates and variability in longitudinal data.

In study [16], an improved Mask R-CNN model is proposed, specifically tailored for multi-organ segmentation in the medical field. This model introduces two major enhancements to the original framework: a multi-scale region of interest (ROI) generation method within the region proposal network (RPN) and a pre-background classification subnetwork to enhance segmentation accuracy. Experimental results on an esophageal cancer dataset demonstrated the model's effectiveness, achieving accurate segmentation of organs such as the heart, lungs, and clinically relevant volumes.

The study by E. Dandıl and M. S. Yıldırım (2021) [17] highlights the significance of computer-aided tools for automatic lung segmentation in diagnosing lung diseases. Manual segmentation by experts can introduce errors and inefficiencies. Their research proposed a Mask R-CNN-based approach, leveraging publicly available datasets like HUG-ILD and VESSEL12. The method demonstrated high performance, achieving a Dice similarity coefficient of 95.95% and a volumetric overlap error of 7.65% for the HUG-ILD dataset, and 96.80% and 6.12% for the VESSEL12 dataset. These results validate the effectiveness of the proposed method for precise lung segmentation (Dandıl & Yıldırım, 2021).

Mask R-CNN is a deep learning model that extends Faster R-CNN to perform instance segmentation, classifying objects and generating pixel-level masks for each detected instance. It introduces the RoIAlign method to improve accuracy when processing regions of interest. Its design, which combines detection and segmentation, makes it ideal for applications such as medical computer vision and autonomous driving [18]. Fig. 1 shows the Mask R-CNN framework for instance segmentation

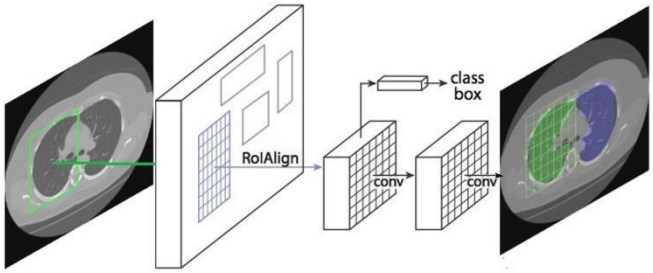


Fig. 1. The Mask R-CNN framework for instance segmentation.

III. METHODOLOGY

A. Database

A total of 20 computed tomography (CT) scans from patients, comprising 3,520 slices, were used for the training and testing phases. Of these, 10 patients were diagnosed with Covid-19 infection pathologies, and 10 patients were diagnosed as healthy or showed no evidence of pulmonary disease.

The number of CT scans used for the training phase included 18 patients, representing 90% of the patients, with a total of 3,126 slices. The number of CT scans for the testing phase included two patients, with a total of 394 slices per image, as shown in Table I.

TABLE I. DATASET FOR TRAINING AND TESTING

Data Set	*%*	*Nro Patient*	*Slides*	*Health*	*Covid*
Training	90	18	3126	9	9
Test	10	2	394	1	1

For the Testing phase, a total of eight computed tomography (CT) scans were used, comprising 259 chest images, as shown in Table II.

TABLE II. DATASET OF HEALTHY AND COVID-19 PATIENTS

Data Set	*%*	*Nro Patient*	*Slides*
Covid-19	75	6	167
Healthy	25	2	92

All CTs for the training and testing phases were obtained from the Zenodo repository in NIFTI format [19]. The CTs for the evaluation phase were obtained from the Research, Technology Transfer, and Software Development Center (CiTeSoft) at the National University of San Agustín.

B. Preprocessing

The images for training, validation, and testing were obtained from medical files in NIFTI format. These slices were converted to PNG format with three channels to preserve the information, scaled to a size of 512 x 512 pixels, and their values were normalized to a range of 0 to 256.

C. Data Augmentation

To avoid overfitting and lack of information, four data augmentation techniques were applied: rotation, horizontal flipping, grid distortion, and elastic transformation. These techniques increased the dataset by 50%, reaching a total of 1760 images distributed between training and testing data, as shown in Table III.

TABLE III. DATASET WITH DATA AUGMENTATION

Data Set	*+%*	*Data Augmentation*
Training	50	1563
Testing	50	197

D. Mask RCNN Architecture

1) Pre-training: In this study, a transfer learning strategy has been adopted by using pre-trained weights from the ImageNet database in the training process of Mask R-CNN. The ImageNet database contains a wide variety of images from different categories, allowing the pre-trained weights to capture general and meaningful visual features. Leveraging these pre-trained weights from ImageNet provides our model with weight initializations that already have a deep understanding of visual patterns, textures, and details in images. This significantly accelerates and improves the training process of Mask R-CNN for our specific task of instance segmentation in medical images of chest CT scans from COVID-19 patients

2) Learning rate: Within the Mask R-CNN architecture, the learning rate is an important hyperparameter that controls the size of the steps the SGD (Stochastic Gradient Descent) optimization algorithm takes to adjust the weights in the model during the training process. This hyperparameter determines how quickly the model converges to a local minimum in the loss function and is responsible for finding a balance between fast and stable model fitting.

3) Implementation: For the implementation of the Mask R-CNN architecture, this study relied on the implementation available in the 'Mask_RCNN' repository by Matterport [9]. This repository provides a Python implementation using TensorFlow, which allowed us to develop and train our instance segmentation model on medical images of chest CT scans from COVID-19 patients.

To carry out the experiments and training of the proposed model, the Mask R-CNN architecture was used with an NVIDIA TITAN RTX graphics card with 24 GB of GDDR6 memory.

The Mask R-CNN parameter configuration was adapted to the specific characteristics of our medical CT image dataset from chest scans of COVID-19 patients. A total of three classes were defined: a) Background and the class corresponding to the areas of interest in the images of the lung regions, b) left, and c) right. For the neural network, 'resnet50' was chosen as the backbone, which leverages the ResNet architecture for feature extraction. The input image dimensions were defined, setting both the minimum and maximum size at 512x512 pixels to ensure consistency in processing. Anchor selection for the Region Proposal Network (RPN) was done using anchor scales of (32, 64, 128, 256, 512) to address different object sizes in our images. During training, a maximum of 200 regions of interest (ROIs) per image and a maximum of 5 true instances for instance detection were established.

For the inference phase, the maximum number of ROIs after the non-maximum suppression (NMS) process was set to 1000, and for training, it was set to 2000. Additionally, a minimum detection confidence threshold of 0.7 was established to ensure the accuracy of the predictions. To mitigate overlap, an NMS threshold of 0.3 was applied to filter redundant detections and improve the consistency of the resulting instance segmentation.

E. Evaluation

For each of the different pre-trained models with a Resnet50 backbone, using different learning rates (0.001, 0.0001, 0.00001) and different epochs, they were evaluated with the following metrics: a) Jaccard Index (1) and b) Dice Coefficient (2).

The equations corresponding to these metrics are presented below.

$$J(A,B) = \frac{D(A,B)}{2+D(A,B)} \quad (1)$$

$$D(A,B) = \frac{2.J(A,B)}{1+J(A,B)} \quad (2)$$

IV. RESULTS AND DISCUSSION

Results were evaluated under two scenarios: with and without Data Augmentation. The best performance was achieved with a learning rate of 0.001. Metrics such as Jaccard Index and Dice Coefficient highlighted the model's segmentation capabilities, which are shown in Table IV.

TABLE IV. SUMMARY OF THE BEST RESULTS

Metric	*With Augmentation*	*Without Augmentation*
Jaccard Index	0.8794	0.8901
Dice Coefficient	0.9266	0.9340

Pixel-level analysis showed high sensitivity (93.39%), specificity (99.56%), and accuracy (98.84%) with augmentation, underscoring the model's reliability in identifying lung regions and lesions.

A. Training with Data Augmentation

In relation to the results obtained for the Jaccard Index metric, better results were achieved with a learning rate of 0.001, with maximum and minimum values of 0.8600 and 0.8794, respectively. Regarding the Dice Coefficient metric, minimum and maximum values of 0.9105 and 0.9266 were

recorded, as shown in Fig. 2. Table V details the results corresponding to each data subset used for training (Fold).

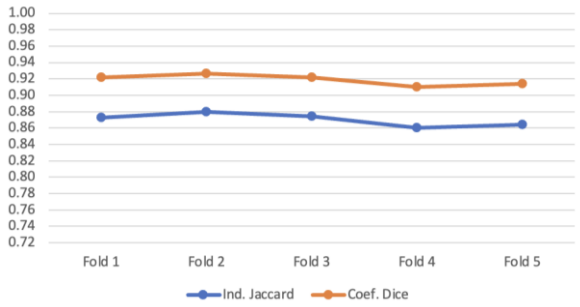


Fig. 2. Learning rate with Jaccard metrics and dice coefficient with data augmentation technique and learning rate of 0.001.

TABLE V. JACCARD METRICS AND DICE COEFFICIENT WITH DATA AUGMENTATION AND LEARNING RATE OF 0.001

Folds	*Jaccard*	*Coef. Dice*
Fold 1	0.8730	0.9215
Fold 2	0.8794	0.9266
Fold 3	0.8742	0.9218
Fold 4	0.8600	0.9105
Fold 5	0.8639	0.9142

In the context of pixel analysis, the metrics of Accuracy, Sensitivity, Specificity, and Precision were evaluated. These metrics provide a detailed understanding of how the model performs in the classification and precise segmentation of pixels in the images. The results corresponding to this evaluation are presented in Fig. 3, Table VI, offering a comprehensive view of the model's effectiveness at the pixel level.

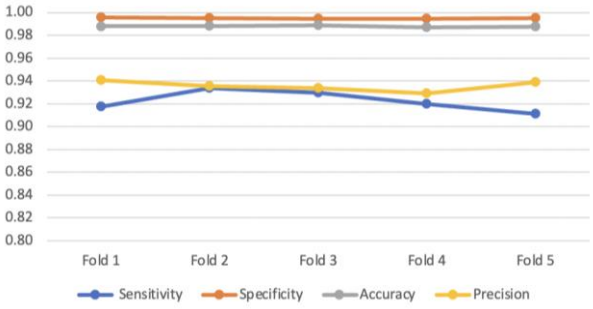


Fig. 3. Learning rate with pixel-level metrics with data augmentation and learning rate of 0.001.

TABLE VI. PIXEL-LEVEL METRICS WITH DATA AUGMENTATION AND LEARNING RATE OF 0.001

Folds	*Sensitivity*	*Specificity*	*Accuracy*	*Precision*
Fold 1	0.9175	0.9956	0.9879	0.9406
Fold 2	0.9339	0.9950	0.9884	0.9357
Fold 3	0.9298	0.9947	0.9885	0.9337
Fold 4	0.9196	0.9946	0.9872	0.9289
Fold 5	0.9113	0.9949	0.9877	0.9390

B. Training without Data Augmentation

In relation to the results obtained for the Jaccard Index metric, better results were achieved with a learning rate of 0.001, with minimum and maximum values of 0.8787 and 0.8901, respectively. Regarding the Dice Coefficient metric, minimum and maximum values of 0.9226 and 0.9339 were recorded, as shown in Fig. 4. Table VII details the results corresponding to each data subset used for training (Fold).

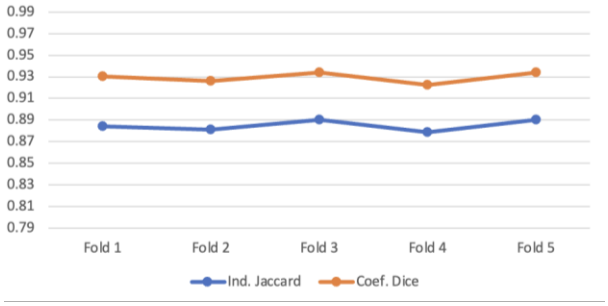


Fig. 4. Learning rate without Jaccard metrics and dice coefficient with technique without data augmentation and learning rate of 0.001.

TABLE VII. JACCARD METRICS AND DICE COEFFICIENT WITHOUT DATA AUGMENTATION AND LEARNING RATE OF 0.001

Folds	*Ind. Jaccard*	*Coef. Dice*
Fold 1	0.8839	0.9302
Fold 2	0.8808	0.9260
Fold 3	0.8901	0.9340
Fold 4	0.8787	0.9226
Fold 5	0.8901	0.9339

In the context of pixel analysis, the metrics of Accuracy, Sensitivity, Specificity, and Precision were evaluated. These metrics provide a detailed understanding of how the model performs in the classification and precise segmentation of pixels in the images. The results corresponding to this evaluation are presented in Fig. 5, Table VIII, offering a comprehensive view of the model's effectiveness at the pixel level.

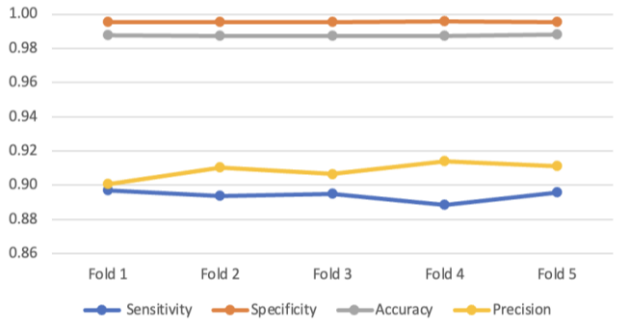


Fig. 5. Learning rate with pixel-level metrics without data augmentation and learning rate of 0.001.

TABLE VIII. PIXEL-LEVEL METRICS WITHOUT DATA AUGMENTATION AND LEARNING RATE OF 0.001

Folds	*Sensitivity*	*Specificity*	*Accuracy*	*Precision*
Fold 1	0.8969	0.9952	0.9877	0.9005
Fold 2	0.8936	0.9954	0.9874	0.9103
Fold 3	0.8947	0.9953	0.9874	0.9064
Fold 4	0.8885	0.9957	0.9872	0.9140
Fold 5	0.8956	0.9954	0.9879	0.9110

In Table XI and Fig. 6, the most notable results regarding the Jaccard Index and Dice Coefficient metrics are presented.

TABLE IX. PIXEL-LEVEL METRICS WITHOUT DATA AUGMENTATION AND LEARNING RATE OF 0.001

Learning Rate	*Ind. Jaccard*	*Coef. Dice*
0.001	0.8901(3)	0.9340(3)
0.0001	0.8777(5)	0.9234(5)
0.00001	0.8264(5)	0.8788(4)

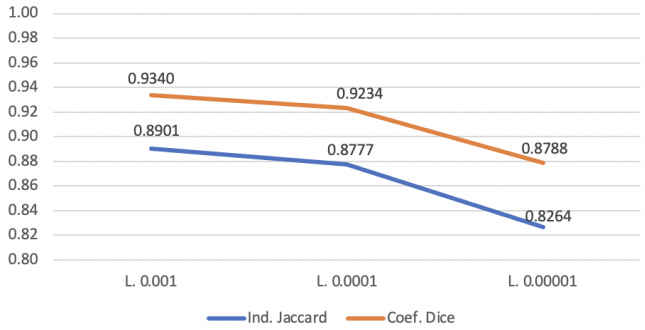


Fig. 6. Learning rate with pixel-level metrics without data augmentation and learning rate of 0.001.

Thorough tests were conducted to optimize the configuration of the semantic segmentation model. In order to explore different approaches, experiments were carried out with the approach a) With data augmentation and b) Without data augmentation, and learning rates were adjusted to 0.001, 0.0001, and 0.00001. After a thorough analysis, the most promising results were obtained by combining the absence of Data Augmentation with a learning rate of 0.001. These findings highlight the importance of precisely and custom tailoring the hyperparameters to maximize the performance of our model. Table IX and Fig. 6, presents the performance metrics under different learning rate configurations, demonstrating that a rate of 0.001 yielded the highest Dice Coefficient and Jaccard Index values.

One of the main limitations of this research was the need to work with all the computed tomography images due to the knowledge provided by the specialists. According to them, COVID-19 could manifest both in the upper and lower images, which complicated the analysis and reduced our detection rate, as the affected areas were very small and difficult to identify. Additionally, we faced the challenge of accurately segmenting the pulmonary parenchyma due to the pathologies caused by COVID-19 in the lungs, which added significant complexity to the analysis process.

The study by Shu, Nian, Yu, and Li (2020) [16] employed Mask R-CNN for lung segmentation, reporting high Dice similarity coefficients (98.1% for the right lung and 97.6% for the left lung). However, the images used do not specify whether the lungs had any pathologies, and the tests were conducted on a range of CT scan slices that included complete organs as well as the lungs, which could influence the model's accuracy when segmenting the lungs exclusively.

In contrast, our research addresses a more complex scenario by working exclusively with CT scan images from patients with COVID-19. This approach not only ensures that the images include clear cases of the disease but also utilizes all slices of the CT scans, including those with smaller or limited affected lung regions. Despite these additional challenges, our model achieved a Dice coefficient of 93.4%, demonstrating its effectiveness in real clinical scenarios related to COVID-19.

Additionally, the work by Dandıl and Yıldırım (2021) [17] also used Mask R-CNN but focused exclusively on lung images with interstitial lung diseases. This study reported Dice coefficients of 95.95% and 96.80% on the HUG-ILD and VESSEL12 datasets, respectively. However, these datasets do not include specific cases of COVID-19, which limits their direct applicability in diagnosing and managing this disease in clinical settings.

In summary, our research stands out by directly addressing lung segmentation in images from patients with COVID-19 and utilizing all the available slices in the CT scans. This ensures a more detailed and relevant approach to the specific challenges posed by this disease.

V. CONCLUSION

This study demonstrates the effectiveness of Mask R-CNN with a ResNet-50 backbone for segmenting lung parenchyma in COVID-19 chest CT scans. Optimal performance was achieved with a learning rate of 0.001 and without Data Augmentation, achieving a Dice Similarity Coefficient (DSC) of 93.4%. Future work will focus on expanding the dataset, exploring alternative backbone architectures, and enhancing segmentation in heterogeneous clinical settings.

Additionally, the use of pre-trained ImageNet weights significantly enhanced the model's performance. By capturing general visual features, these weights accelerated the training process and improved segmentation accuracy. This highlights the importance of transfer learning in specialized tasks, such as the segmentation of medical images from COVID-19 patients.

When evaluating the impact of data augmentation, the study found that while this technique achieved a maximum DSC of 92.7%, training without data augmentation outperformed it, yielding a higher DSC of 93.4%. These results suggest that, for this specific task, excluding data augmentation contributes to better segmentation performance, challenging common assumptions about the universal benefits of augmentation.

Extensive experiments were conducted using a public dataset for training and a custom dataset from Arequipa, Peru,

provided by the Research, Technology Transfer, and Software Development Center I+D+i - CiTeSoft. The findings validated the adaptability of the Mask R-CNN method, demonstrating its effectiveness even when applied to a regional population with potentially different characteristics from the training data.

This study opens new opportunities for improving and expanding the application of the proposed method. Future work will focus on experimenting with different backbone architectures, comparing models trained from scratch versus those using pre-trained weights, and determining the optimal number of epochs to achieve well-trained models without overfitting. Furthermore, testing will be extended to medical images of other anatomical regions to evaluate the method's adaptability and robustness across diverse clinical scenarios.

REFERENCES

[1] Y. Huérfano et al., "Imagenología médica: fundamentos y alcance," *Archivos Venezolanos de Farmacología y Terapéutica*, vol. 35, no. 3, pp. 71–76, 2016.

[2] Y. X. Tay, S. Kothan, S. Kada, S. Cai, and C. W. K. Lai, "Challenges and optimization strategies in medical imaging service delivery during COVID-19," *World Journal of Radiology*, vol. 13, no. 5, pp. 102–121, 2021.

[3] Z. Y. Zu et al., "Coronavirus disease 2019 (COVID-19): A perspective from China," *Radiology*, vol. 296, no. 2, pp. E15–E25, 2020.

[4] V. Varadarajan, M. Shabani, B. Ambale Venkatesh, and J. A. C. Lima, "Role of Imaging in Diagnosis and Management of COVID-19: A Multiorgan Multimodality Imaging Review," *Frontiers in Medicine*, vol. 8, p. 765975, 2021.

[5] N. Brandi and M. Renzulli, "The role of imaging in detecting and monitoring COVID-19 complications in the Intensive Care Unit (ICU) setting," *APS*, vol. 2, no. 3, p. 3, Jan. 2024.

[6] D. Gandhi et al., "Current role of imaging in COVID-19 infection with recent recommendations of point of care ultrasound in the contagion: a narrative review," *Annals of Translational Medicine*, vol. 8, no. 17, p. 1094, 2020.

[7] D. Dong *et al.*, "The Role of Imaging in the Detection and Management of COVID-19: A Review," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 16-29, 2021.

[8] M. Alhasan and M. Hasaneen, "Digital imaging, technologies and artificial intelligence applications during COVID-19 pandemic," *Computerized Medical Imaging and Graphics*, vol. 91, p. 101933, 2021.

[9] P. Kumar, D. Jayaswal, M. Khan, and B. Singh, "A relative analysis of different CNN based models for COVID-19 detection using CXR and CT images," *Procedia Computer Science*, vol. 235, pp. 3163–3173, 2024.

[10] S. Podder, S. Bhattacharjee, and A. Roy, "An efficient method of detection of COVID-19 using Mask R-CNN on chest X-ray images," *AIMS Biophysics*, vol. 8, no. 3, pp. 281–290, 2021.

[11] D. Suganya and R. Kalpana, "Prognosticating various acute COVID lung disorders from COVID-19 patient using chest CT images," *Engineering Applications of Artificial Intelligence*, vol. 119, p. 105820, 2023.

[12] S. Aparna, K. Muppavaram, C. C. V. Ramayanam, and K. S. Sai Ramani, "Mask RCNN with RESNET50 for Dental Filling Detection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 10, 2021.

[13] M. Al Masarweh, O. Oluseyi, A. Alkafri, H. Alsmadi, and T. Alwadan, "Automatic Detection of Lumbar Spine Disc Herniation," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 11, 2024.

[14] A. A. Mashaal, M. A. A. Eldosoky, L. N. Mahdy, and K. A. Ezzat, "Classification of Human Sperms using ResNet-50 Deep Neural Network," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 2, 2023.

[15] S. Suriyavarman and A. X. A. R. Annie, "Lung Nodule Segmentation and Classification using U-Net and Efficient-Net," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 7, 2023.

[16] J.-H. Shu, F.-D. Nian, M.-H. Yu, and X. Li, "An Improved Mask R-CNN Model for Multiorgan Segmentation," *Mathematical Problems in Engineering*, vol. 2020, Article ID 8351725, 11 pages, 2020.

[17] E. Dandıl and M. S. Yıldırım, "A Mask R-CNN based Approach for Automatic Lung Segmentation in Computed Tomography Scans," 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Kocaeli, Turkey, 2021, pp. 1-6.

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *arXiv*, 2018.

[19] "COVID-19 CT Lung and Infection Segmentation Dataset," Zenodo, Apr. 20, 2020.

Impact of the TikTok Algorithm on the Effectiveness of Marketing Strategies: A Study of Consumer Behavior and Content Preferences

Raquel Melgarejo-Espinoza, Mauricio Gonzales-Cruz, Juan Chavez-Perez, Orlando Iparraguirre-Villanueva*
Facultad de Ingeniería, Universidad Tecnológica del Perú, Chimbote, Perú

***Abstract*—TikTok has become one of the most widely used platforms, its innovative video format has allowed companies and users to increase their visibility, transforming the way brands communicate their strategies. This systematic literature review (SLR) explored how the TikTok algorithm influences marketing strategies during the period 2021 to 2024. For this purpose, research was conducted based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) method. Also, reliable and relevant research databases were consulted, specifically Springer, Science Direct and EBSCO, from which 64 studies aligned with the inclusion and exclusion criteria were extracted, all corresponding to academic articles. After compilation, it was determined that 2024 was the year with the highest number of publications, representing 50% of the total number of articles. Likewise, the country that stood out was China with 28.13% of the related documents. Regarding the research approach, quantitative research predominated, followed by qualitative and mixed research. Finally, the study helped to understand the positive impact of TikTok on marketing, showing how it improves the visibility of brands, as well as identifying trends in consumer preferences, which allows the creation of more accurate strategies that are closer to the public.**

***Keywords*—TikTok; algorithm; consumer behavior; marketing**

I. Introduction

In recent years, social media has significantly transformed the way brands engage with consumers and promote their products or services [1]. Globally, platforms such as Facebook, Instagram, Twitter and Tiktok have redefined the marketing landscape, providing companies with new tools to connect with their audiences in a more direct and personalized way [2].

Among all platforms, TikTok has stood out as one of the most influential apps, amassing over 272 million followers [3]. Its focus on short, engaging videos allows users to easily record, edit and share content, making it an attractive space for the digitally connected Generation Z audience [4]. This generation, born between 1997 and 2012, spends a considerable amount of time on social media, preferring to search for products on platforms such as TikTok rather than traditional search engines such as Google [5]. Recent data indicates that 43% of Generation Z members prefer to browse products on TikTok, revealing a significant shift in the purchasing patterns of young consumers [6].

New digital habits have exposed the platform's users to constant content, a key factor in the emergence of a new market, where content creators become essential allies for the presentation and promotion of products or services [7]. This trend, together with the rapid evolution of the platform, has consolidated TikTok as an indispensable tool for the market. The same platform has introduced innovative features such as integrated e-commerce and affiliate programmers, allowing companies to promote and sell products directly [8].

TikTok has positioned itself as one of the preferred entertainments platforms and has become a massive sales channel designed to create a buying environment [9]. Constant exposure to content has made consumers more susceptible to marketing strategies [10], and brands and content creators play an important role in the user experience [11]. Platforms such as TikTok leverage artificial intelligence (AI) to display relevant content that captures user interest, keeping them on the platform for longer [12]. Algorithms store knowledge of interactions, browsing behavior and content preference. This information is useful for businesses because it allows them to make decisions based on the behavior of their consumers [13].

TikTok has an advanced algorithm system, which stands out for its ability to encourage engagement and deliver relevant content [14]. Changes in the technological landscape have raised new adaptive scenarios for the business environment, brands have seen TikTok as a tool to generate business opportunities, interact with customers and generate personalized experiences [15].

The present research is justified by the need to comprehensively examine the impact of these strategies on consumer behavior. This research article focuses on analyzing how TikTok marketing, and the algorithm, influences brand awareness, purchase intentions and actual consumer behavior. This research aims to provide a greater understanding of the mechanisms behind the effectiveness of TikTok marketing and provide insight into applied trends to optimize advertising campaigns.

II. Literature Review

Several studies have looked at the impact of social media and consumer behavior on digital marketing strategies. An analysis by [16] noted that short-form platforms such as TikTok have rethought the interaction between brands and consumers in order to improve the connection and personalization of content, taking an approach based on big data analysis and applying the quantitative method, they concluded that TikTok has managed to transform marketing strategies by facilitating a level of personalization and

connection between brand and audience. Furthermore, the study [17] analyzed the importance of TikTok marketing on SKINTIC's brand image to determine its impact on user perception. Through a simple linear regression analysis, the study concludes that TikTok marketing has a 53% influence on SKINTIC's brand visibility.

The paper in [18] studied the marketing strategies employed in TikTok's e-commerce platform, using SWOT analysis to understand consumer motivation to improve the interaction between consumers and brands. Similarly, in [19] they investigated the implementation of the algorithm in promotion and marketing, with the aim of analyzing how AI has impacted technological and advertising interaction, using a descriptive approach, the study highlights AI as a fundamental tool to improve the user experience in the advertising field.

The study in [20] conducted an advertising analysis with the aim of understanding how creators, users and the recommendation algorithm influence responses and decisions to make purchases. Using a quantitative approach 2,000 posts on TikTok were analyzed, the results expressed the importance of generating valuable and relevant content to user interests, to increase the reliability and interaction of the brand. In this regard, [21] focused on examining TikTok's short-form video ads to determine their influence on purchase intentions by using the SOR (Stimulus-Organism-Response) model, the study posits that marketers can gradually drive consumers' purchase intentions through the interactivity of video advertising.

The research in [22] studied consumer behavior to understand how TikTok influences Semitic's purchase intent by conducting a survey of 403 people, the results of which underline the importance of observing variations in attitudes towards the product, pointing out how trends, along with reviews and recommendations, are key factors in Semitic's business success. Similarly, in study [23] they focused on identifying the key characteristics of short-form video ad content and how these attributes affect consumer purchase behavior, the study analyzed 2578 videos from which they concluded in identifying three key characteristics that influence the purchase process, trustworthiness, experience and appeal.

Table I shows the conclusions and gaps found in the reviewed works.

TABLE I. CONCLUSIONS AND GAPS FOUND IN THE REVIEWED WORKS

Reference	Key findings	Gaps/deficiencies
[16]	TikTok transforms marketing strategies through personalization and audience connection.	Specific conversion and ROI metrics are not explored.
[17]	TikTok marketing influences brand visibility by 53%.	Lack of data on impact on different consumer segments.
[18]	Marketing strategies on TikTok motivate interaction between consumers and brands.	Does not analyze the impact of competition on consumer motivation.
[19]	AI on TikTok enhances the advertising experience and facilitates technological interaction.	Limited to a descriptive view, with no quantitative measurement of impact.
[20]	Quality content aligned with user interests improves interaction and brand trust.	Does not address how strategies can be optimized for different consumer profiles.
[21]	Short-form ads can drive purchase intent if they have interactivity.	This work does not consider the impact of content saturation on ad effectiveness.
[22]	Trends, reviews and recommendations are key to consumer purchase intent.	Does not study the role of creator authenticity on consumer perception.
[23]	Content characteristics influence purchase behavior: trustworthiness, experience and appeal.	Lack of analysis on how the combination of these factors affects customer loyalty.

III. METHODOLOGY

This research initiative aimed to study the impact of the TikTok platform algorithm on digital marketing strategies by employing a comprehensive qualitative approach, coupled with the PRISMA systematization method [24].

This method was designed to help authors of systematic reviews to transparently document each research process. The recent update in 2020 has facilitated study selection, evaluation and systematization, thus ensuring a more efficient and structured approach to research [25].

To carry out the literature review, we systematized studies related to consumer behavior and content preferences on TikTok, considering the PRISMA method guidelines, which allowed for a detailed analysis of the TikTok platform algorithm and its impact on the interaction and effectiveness of marketing campaigns between 2021 and 2024. During the process, specific parameters determined by PRISMA were used, consisting of data collection and filtering. For this, identification, correlation and acceptance criteria were examined. These criteria aided in the understanding and analysis of the research. The inclusion criteria under consideration for the research process had to meet the following parameters:

- Studies must be from scientific databases, which guarantee reliability and quality, to deliver a solid and well-founded systematic review.
- Research should be conducted from 2021 to 2024, as this is the most relevant range for data collection.
- Studies should use terminology related to the research topic, such as marketing, TikTok algorithm and consumer behavior, to ensure the focus of the research.
- Research papers should be written in English, as this broadens the possibilities of finding relevant information.

For the exclusion criterion, the following requirements were determined:

- Any research outside the publication range date of 2021 to 2024 was excluded from the systematic review.
- Studies whose content was not fully aligned with the research objective, as it would not contribute to answering the questions of the review paper.
- When the information in the papers is far from the research topic, it will not be useful for the construction of the study.

In the research process, multiple databases recognized for their contribution to the scientific literature were used, such as Springer, Science Direct, IEEE Xplore and EBSCO.

In the first phase of the research, a search was conducted using key terms for the study, carried out in the English language, such as "Marketing", "TikTok AND Algorithm" and "Consumer Behavior". Subsequently, and to collect current and relevant information, criteria filters were applied to delimit the results obtained during the year 2021 to 2024. After this process, 64 articles with valuable content on the functioning of the TikTok platform and the effectiveness of the algorithm in the strategies used to impact consumers were obtained, thus aligning with the objective of the research.

Fig. 1 shows the process of extracting the documents and their relationship with the databases. This process was divided into three phases, to finally obtain the total number of documents relevant to the research.

From the beginning of the data collection, we proceeded to identify the selected documents considering their origin, which allowed us to carry out a more efficient and systematic analysis of the scientific literature. To identify duplicates, we took advantage of the organizational functionality of the software, Mendeley. We also sought the incorporation of another tool, which facilitated the efficient management of the information, as well as allowing export functions in XML format of the documents stored in Mendeley. Subsequently, the data collected was imported into Microsoft Excel. This process facilitated the identification of 24 duplicates and allowed the improvement in the structuring of the studies, for the improvement process aspects such as the year of publication, the type of publication, the country of origin and the methodological approach used have been considered.

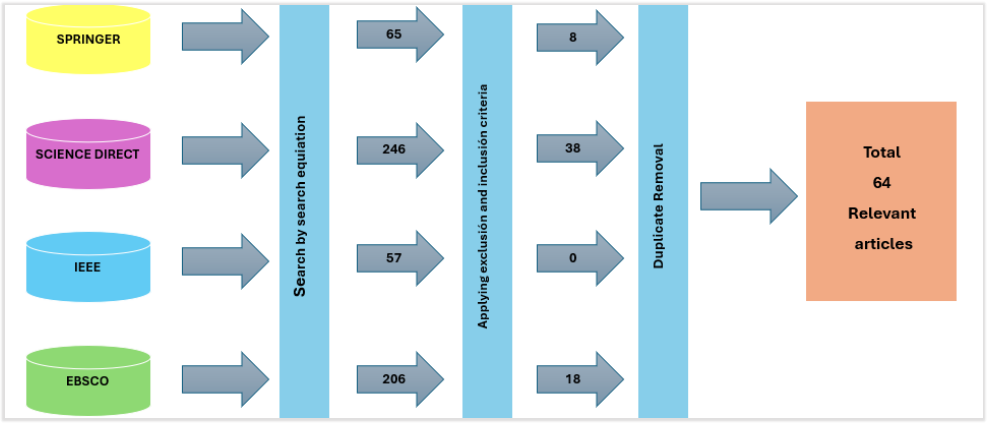


Fig. 1. Database distribution.

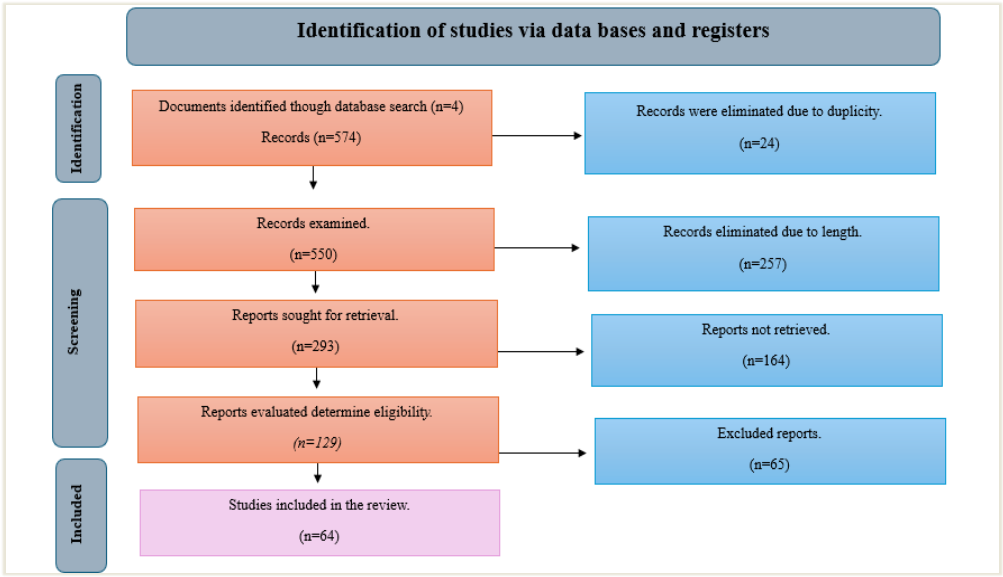


Fig. 2. PRISMA method.

In addition, Fig. 2 shows the total number of documents used during the search for information in all the databases used. This procedure was carried out following the framework of parameters established by the PRISMA method, thus ensuring the transparency and integrity of the literature review process.

IV. RESULTS

For the analysis of the results of the systematic review, the initial database containing 574 records had to be structured and organized. After a rigorous analysis and based on inclusion and exclusion criteria and the application of various filters, a final selection of 64 studies was made. In the initial stage, 574 papers were recognized associated with keywords such as "Marketing", "TikTok AND Algorithm" AND consumer Behavior. As a result, the number of studies found per database was as follows: Springer contributed 11.32%, Science Direct 42.85%, IEEE Xplore 9.93% and EBSCO 35.88%.

For the second stage, 293 records were examined considering the research year range 2021-2024, this process was called "screen viewing". As a result of this phase, it was determined that 13.93% of the documents came from Springer, 39.59% from ScienceDirect, 9.55% from IEEE Xplore and 36.86% from EBSCO.

In the third phase, the titles of the available documents were evaluated, prioritizing those that included keywords related to the focus of the study. As a result, 164 records were discarded, leaving 129 valid documents. These documents were distributed as follows: 11.62% from Springer, 34.88% from Science Direct, 0% from IEEE Xplore and 53.48% from EBSCO. This step was essential to ensure the relevance and quality of the selected studies, aligning them with the research objective.

For the final stage of the analysis, all studies were selected from the scientific literature, which were related to the research objectives. This process involved reviewing the abstracts to verify the validity and functionality of the access links to the documents, ensuring that the studies were available in their entirety and avoiding access problems. As a result, 64 scientific papers were included, while 65 studies were discarded during the final review as they did not meet the assessment criteria.

During the research, 64 articles were analyzed, and it was determined that from 2021 to 2024, several publications were made in recognized databases such as Springer, Science Direct and EBSCO, of which 8 articles belong to Springer, 38 to Science Direct and 18 to EBSCO. Fig. 3 shows the final distribution of the studies in their respective databases used for systematic review.

For the analysis of articles by year, it was considered that the publication period complies with the established range between 2021 and 2024. After this analysis it was determined that 7 (10.94%) publications were from the year 2021, 9 (14.06%) were made in the year 2022, while in 2023 16 (25.00%) articles were published and during 2024 32 (50.00%) articles related to the research topic were published. Fig. 4 shows the number of articles found according to their year of publication.

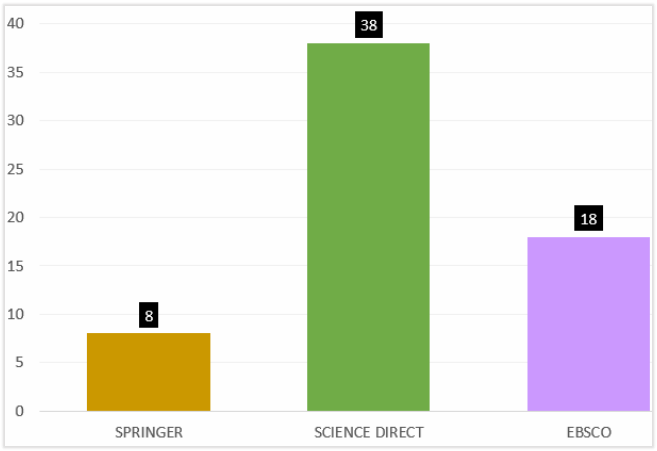


Fig. 3. Distribution of studies by database.

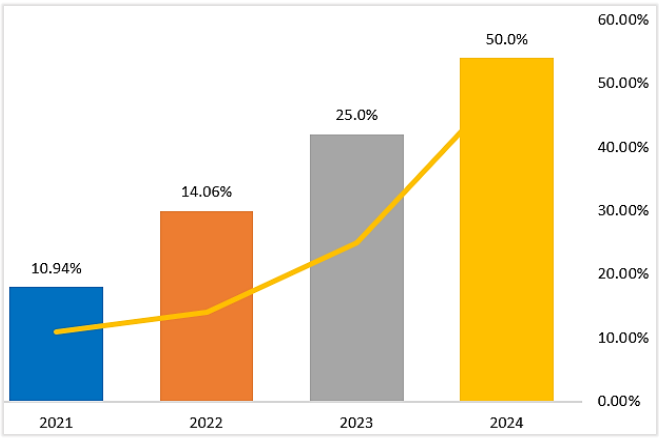


Fig. 4. Distribution of studies by year of publication.

Fig. 5 shows the number of articles published per year in the Springer, Science Direct and EBSCO databases. In 2021, a total of 7 articles were published: 1 from Springer accounting for 14.3% of documents, 4 from Science Direct accounting for 57.1% and 2 from EBSCO accounting for 28.6%. In 2022, 9 articles were registered: 1 from Springer with 11.1%, 7 from Science Direct equivalent to 77.8% and 1 from EBSCO 11.1%. During 2023, the total was 16 articles: 4 from Springer with 25%, 9 from Science Direct with 56.3% and 3 from EBSCO with 18.8%. Finally, in 2024, 32 articles were published: 2 from Springer representing 6.3%, 18 from Science Direct representing 56.3% and 12 from EBSCO equivalent to 37.5%. The following graph shows the visual representation of the studies by year and their origin in the database.

To study the origin of the research papers, the following graph was made. In it, it was identified that the largest number of studies come from China, the result represents 28.13%, followed by the United States with 15.63%, Spain represents 7.81% and the United Kingdom 6.25% of the collection. In addition, it should be noted that most of the documents were written in English. Fig. 6 shows the origin of research at country level.

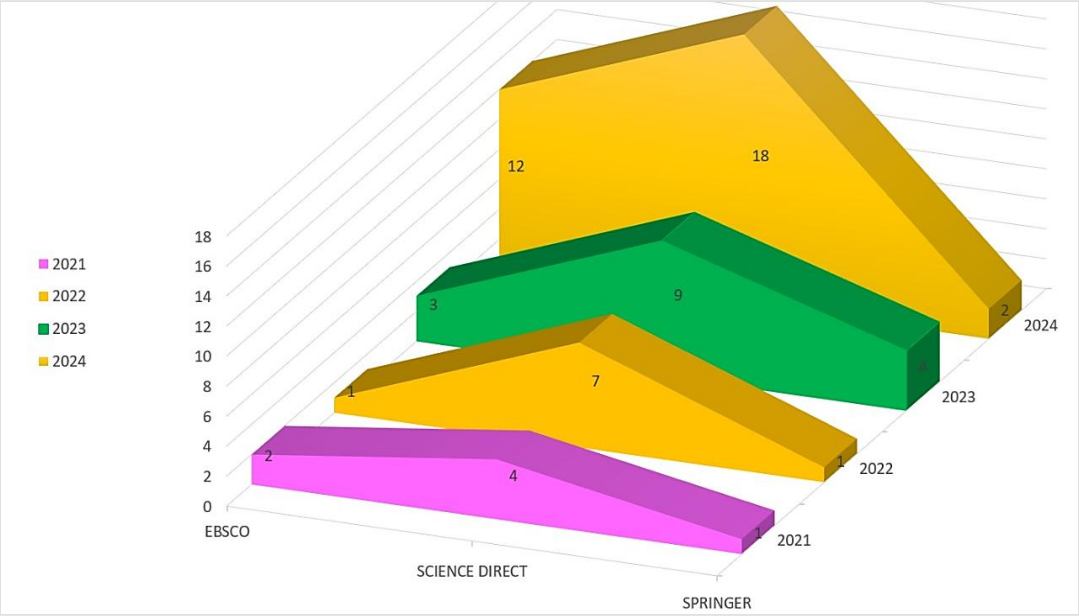


Fig. 5. Number of studies per year and databases.

Fig. 6. Scientific literature reviews by country.

Similarly, it is important to note the classification of the bibliography according to the type of research. To determine the type of research, an exhaustive analysis of the documents had to be carried out, for which purpose they were classified and counted; it was determined that 100% of the scientific publications collected correspond to journal articles, which is equivalent to a total of 64 documents.

On the other hand, Fig. 7 reflects the classification of three types of methodological research. The quantitative approach for Science Direct represented the largest amount with 23 documents, while EBSCO obtained 5 and Springer obtained 1. As for the qualitative approach, the EBSCO database addressed 9 documents, Science Direct contributed 6 and Springer 4 documents. Likewise, the largest amount contributed for the mixed approach was from Science Direct with 9 papers, EBSCO 4 and Springer accounted for 3 papers. Fig. 7 provides the distribution of previously mentioned studies.

For the final analysis, VosViewer was used, a tool recognized for its visualization and bibliometric analysis capacity. This tool made it possible to identify the most recurrent terms in the selected studies, providing a comprehensive perspective on the main trends in digital marketing. Essential terms such as "Social media" have enabled the development of "influencer marketing" on platforms known as "TikTok". Similarly, the analysis evidenced the relevance of the term "Digital content" and showed a strong preference for "short form video" content, this format is highly engaging because of its "persuasion knowledge", the algorithm of these platforms personalizes the content which makes users prone to "purchase intention" within the platform. All terms align significantly with the theoretical framework of the studies analyzed, reaffirming the centrality of social media and short form video in contemporary digital marketing strategies.

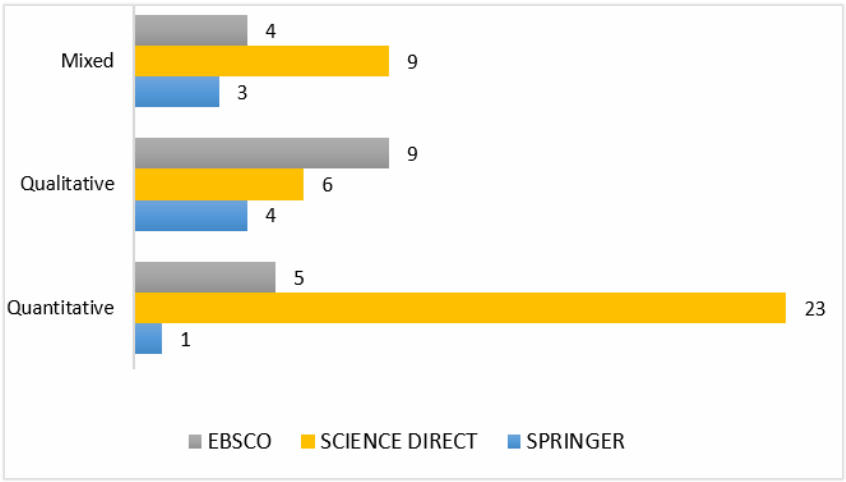


Fig. 7. Scientific literature incorporated in research by methodological approach.

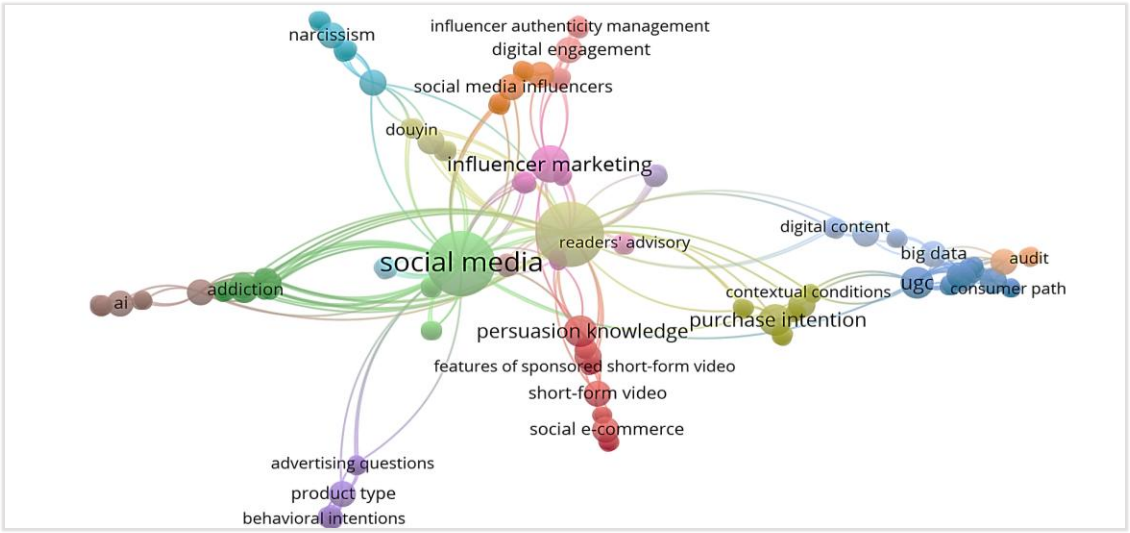


Fig. 8. Bibliometric exploration of the analyzed literature.

The analysis carried out provides an overview of the most relevant terms of the research and is visually represented in Fig. 8.

V. Discussion

A. *RQ1: What Patterns of Consumer Behavior can be Identified from Interaction with TikTok Content?*

Ongoing interaction has led to the identification of behavioral patterns of TikTok consumers. Users tend to show a greater preference for short, visually appealing content. Product presentation and quality is also a factor that users look for. These features are critical and have reflected increased consumer engagement with their preferred brands. The power of consumers has increased with the advent of TikTok, the available video information and brand recommendations abound within the platform, so the virality factor becomes a predominant pattern. In contrast, after an analysis of interactions, the study [16] supports that engaging content is a fundamental part of improving the connection between brands and consumers. On the other hand, for the study [18] interaction is the predominant pattern of consumer behavior on TikTok, the development of e-commerce and live streaming is some evidence of the relevance of today's consumer to create, recommend and comment on videos. It also relates to research [22] that recognizes the importance of interaction to Semitic's success, with the brand highlighting that creating short, quality videos help to enhance its brand image. These studies indicate part of the patterns found, however, to go deeper into the topic, a larger number of related documents were considered, as shown in Table II.

TABLE II. Patterns of Behaviors Found in the Research

#	Patterns	Quantity	Reference
1	Quality	5	[26], [27], [28], [29], [30]
2	Attractive	4	[31], [32], [33], [34]
3	Trends	8	[35], [36], [37], [38], [39], [40], [41], [42]
4	Participation	9	[40], [43], [44], [45], [46], [47], [48], [49], [50]
5	Brevity	3	[51], [52], [53]

The data collection highlights key consumption patterns to understand how companies can effectively connect with consumers. Among the most relevant findings, interactive content and current trends, followed by quality, visual appeal and brevity of content are identified as recurring consumption elements within the platform.

B. RQ2: Are there Differences in Users' Emotional Responses to Ads Targeted by TikTok's Algorithm?

There are a variety of emotional responses to ads targeted by the TikTok algorithm. It is shown that ads aligned with the user's personal interests tend to generate positive emotional responses, which results in greater consumer engagement and willingness to interact with the ad. In contrast, ads perceived as intrusive tend to result in negative user experiences, thus highlighting the importance of generating positive consumer experiences to ensure the success of ads. The various responses have been supported by [19] in their research on the implementation of the algorithm in promotion and marketing, the study has highlighted that AI helps to generate positive user experiences. It is also related to the study [20] that after its analysis of 2,000 publications in TikTok has highlighted that users have had a positive experience, against the content presented, in addition, it highlighted a greater participation of users against content of their interest. Whereas, for [21] interactivity is one of the frequent responses generated by TikTok's short-form video ads, the study highlights the algorithm as a tool that marketers can use to get favorable purchase responses from consumers through the interactivity of video advertising. To further explore the response, several studies have been considered that will broaden the picture and provide insight into the various responses generated by the algorithm-driven ad, as evidenced in Table III.

TABLE III. EMOTIONAL DIFFERENCES OF USERS

#	Differences	Quantity	References
1	Participation	9	[26], [31], [47], [51], [54], [55], [56], [57], [58], [59]
2	Trust	4	[27], [60], [61], [62]
3	Relevance	9	[45], [63], [64], [65], [66], [67], [68], [69], [70]
4	Experience	4	[71], [72], [73], [74]

After analysis of the collected findings, it can be noted that users' emotional responses are diverse. However, a significant difference is observed in algorithm-driven ads, as these generate higher audience engagement by displaying content relevant to users' interests. It is also considered that another key factor influencing positive or negative responses is the experience and trust that users may have with the advertisement shown.

C. RQ3: What Roles Do Consumer Content Preferences Play in the Effectiveness of Advertising Campaigns on TikTok?

Consumer content preferences play a crucial role in the effectiveness of advertising campaigns on TikTok, as they allow personalizing ads, improving viewability and optimizing campaigns according to consumer interests. In that line, the answers found are supported by [16], which highlights the importance of adapting content with consumer preferences. The study emphasizes that personalization plays a significant role in the reach of advertising campaigns. It is also linked to [17], which analyzed consumer trends, in order to optimize SKINTIC's visibility; it concludes that the marketing employed in TikTok campaigns helped improve brand visibility by 53%. While [23] has also endorsed the importance of content preference, after analyzing 2578 videos, 3 key characteristics for campaign optimization were revealed: trustworthiness, experience and attractiveness are key elements that consumers prefer when consuming content. To further detail the responses, a wider range of sources have been consulted to better understand the role of consumer preferences in advertising campaigns. The relationship between consumer preferences and the effectiveness of advertising campaigns is presented in Table IV.

TABLE IV. RELATIONSHIP BETWEEN CONSUMER PREFERENCES AND ADVERTISING CAMPAIGN EFFECTIVENESS

#	Role of the consumer	Quantity	References
1	Customization	8	[30], [52], [75], [76], [77], [78], [79], [80]
2	Optimization	4	[70], [81], [82], [83]
3	Visibility	4	[84], [85], [86], [87]
4	Targeting	3	[87], [88], [89]

Regarding the third question, there is a clear incidence of research that agrees that personalization of content allows brands to adapt their message to the preferences of consumers, which increases the effectiveness of campaigns. On the other hand, optimization is also shown to be a key factor, followed by visibility and targeting.

VI. CONCLUSION

To conduct this research, an exhaustive review of the scientific literature was carried out to analyze the impact of the TikTok algorithm on the effectiveness of marketing strategies during the period 2021 to 2024. At this stage, studies that were not aligned with the research topic were discarded and a total of 64 original articles were selected as the basis for the work.

To support the research, studies were used that not only supported the topic but also established significant precedents around scientific research, which provided relevant background for the development of the analysis. The selected articles were obtained from three recognized databases: Springer, where 8 documents representing 12.5% of the total were identified, Science Direct, which provided 38 documents equivalent to 59.38% and 18 documents from EBSCO, which represented 28.13% of the collection. All the studies analyzed provided a great contribution to the understanding and analysis of the TikTok platform and its impact on marketing strategies.

This RSL provides a detailed analysis of the studies related to TikTok, to understand the relevance that this platform is acquiring in the field of business marketing. The research seeks to understand how the platform's algorithm works and how this tool becomes a key ally in understanding the consumer and developing effective advertising strategies. The study presented is positioned as a valuable source for future scientific research in the field related to marketing.

It is important to note that the research faced a limitation related to the number of studies related to specific terminology, so it was necessary to carefully determine the keywords related to the research article; after correctly combining the terminology, it was possible to find documents that support the work.

REFERENCES

[1] S. Karataş and E. Karakoç, "The Virtual World Platform 'Tiktok': A Study On Generatıon Z," Erciyes İletişim Dergisi, vol. 11, no. 2, pp. 517–537, Jul. 2024, doi: 10.17680/erciyesiletisim.1440628.

[2] Y. Awanda, N. Harahap, W. Yoga, S. Fadhilah Siregar, R. Ananda, and N. Afifah, "The Influence of Tiktok Social Media on Teenagers' Lifestyles," vol. 3, no. 2, pp. 66–69, 2024, doi: 10.32832/amk.

[3] S. Maghraoui and L. Khrouf, "Cyberaddiction to TikTok during the COVID-19 pandemic," Spanish Journal of Marketing - ESIC, 2024, doi: 10.1108/SJME-01-2023-0023.

[4] H. Ananda Putri and A. Albari, "The Influence of TikTok Shop Service Quality on Consumer Loyalty Regarding Customer Satisfaction, Customer Trust, and Behavioral Intention," Jurnal Impresi Indonesia, vol. 3, no. 3, pp. 195–207, Mar. 2024, doi: 10.58344/jii.v3i3.4710.

[5] C. Su, "Research on the Relationship Between Marketing Strategy of Contemporary TikTok E-commerce Platform and Young People's Consumption Motivation," Journal of Education, Humanities and Social Sciences, vol. 35, pp. 385–390, Jul. 2024, doi: 10.54097/9rafpp17.

[6] F. Campines, "Impacto del mercadeo en Tik Tok en el comportamiento de compra del consumidor," Revista Colón Ciencias, Tecnología y Negocios, vol. 11, no. 1, pp. 20–33, Jan. 2024, doi: 10.48204/j.colonciencias.v11n1.a4655.

[7] C. Aidan, "Sustainable consumption Helping consumers make eco-friendly choices SUMMARY."

[8] E. N. Siregar, P. Pristiyono, and M. A. Al Ihsan, "Analysis of Using Tiktok as Live Marketing in Attracting Consumers' Interest in Buying," Quantitative Economics and Management Studies, vol. 4, no. 3, pp. 453–463, Mar. 2023, doi: 10.35877/454ri.qems1633.

[9] W. Zhang, W. Zhang, and T. U. Daim, "Investigating consumer purchase intention in online social media marketing: A case study of Tiktok," Technol Soc, vol. 74, Aug. 2023, doi: 10.1016/j.techsoc.2023.102289.

[10] S. Aiolfi, S. Bellini, and B. Grandi, "Using mobile while shopping in-store: a new model of impulse-buying behaviour," Journal of Consumer Marketing, vol. 39, no. 5, pp. 432–444, Jul. 2022, doi: 10.1108/JCM-05-2020-3823.

[11] Y. W. Handranata, M. G. Herlina, L. Soendoro, and Q. Kamiliya, "Beyond the swipe: Understanding the power of TikTok marketing-interaction, entertainment, and trendiness in shaping purchase intentions," International Journal of Data and Network Science, vol. 8, no. 4, pp. 2519–2526, 2024, doi: 10.5267/j.ijdns.2024.5.006.

[12] X. Wang and Y. Guo, "Motivations on TikTok addiction: The moderating role of algorithm awareness on young people," Profesional de la Informacion, vol. 32, no. 4, Jul. 2023, doi: 10.3145/epi.2023.jul.11.

[13] I. Aimé, F. Berger-Remy, and M. E. Laporte, "The brand, the persona and the algorithm: How datafication is reconfiguring marketing work☆," J Bus Res, vol. 145, pp. 814–827, Jun. 2022, doi: 10.1016/j.jbusres.2022.03.047.

[14] R. Pérez García and M. Pérez García, "Influencia y consecuencias del neuromarketing en el uso de TikTok por niños, niñas y adolescentes," Pediatría (Asunción), vol. 50, no. 3, pp. 151–153, Dec. 2023, doi: 10.31698/ped.50032023002.

[15] A. K. S. Ong et al., "Consumer Behavior Analysis and Open Innovation on Actual Purchase from Online Live Selling: A case study in the Philippines," Journal of Open Innovation: Technology, Market, and Complexity, vol. 10, no. 2, Jun. 2024, doi: 10.1016/j.joitmc.2024.100283.

[16] K. Deng, C. Meng, and Z. Zhao, "The Analysis of Content-type Short Video Platform TikToks Marketing Strategies in China," Advances in Economics, Management and Political Sciences, vol. 84, no. 1, pp. 231–236, May 2024, doi: 10.54254/2754-1169/84/20240814.

[17] K. A. Riwong and H. Y. Wono, "THE INFLUENCE OF TIKTOK MARKETING CONTENT ON SKINTIFIC BRAND IMAGE Volume: 5 Number: 4 Page: 699-707."

[18] C. Su, "Research on the Relationship Between Marketing Strategy of Contemporary TikTok E-commerce Platform and Young People's Consumption Motivation," 2024.

[19] B. Al Haj Bara, N. N. Pokrovskaia, M. Y. Ababkova, I. A. Brusakova, and A. A. Korban, "Artificial Intelligence for Advertising and Media: Machine Learning and Neural Networks," Proceedings of the 2022 Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2022, pp. 8–11, 2022, doi: 10.1109/ELCONRUS54750.2022.9755590.

[20] E. Agrawal, "Agrawal 1 Going Viral: An Analysis of Advertising of Technology Products on TikTok," 2023.

[21] X. Li, G. Dong, and Y. L. Xie, "Research on the influence of short video advertising Interactivity on consumers' Purchase Intention - based on S-O-R model," 2023 International Conference on Computer Applications Technology (CCAT), pp. 271–275, Sep. 2023, doi: 10.1109/CCAT59108.2023.00057.

[22] Indrawati, P. C. Putri Yones, and S. Muthaiyah, "eWOM via the TikTok application and its influence on the purchase intention of somethinc products," Asia Pacific Management Review, vol. 28, no. 2, pp. 174–184, 2023, doi: https://doi.org/10.1016/j.apmrv.2022.07.007.

[23] L. (Monroe) Meng, Y. Bie, S. Kou, and S. Duan, "The impact of content characteristics of Short-Form video ads on consumer purchase Behavior: Evidence from TikTok," J Bus Res, vol. 183, p. 114874, Oct. 2024, doi: 10.1016/J.JBUSRES.2024.114874.

[24] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," BMJ, vol. 372, Mar. 2021, doi: 10.1136/BMJ.N71.

[25] J. J. Yepes-Nuñez, G. Urrútia, M. Romero-García, and S. Alonso-Fernández, "Declaración PRISMA 2020: una guía actualizada para la publicación de revisiones sistemáticas," Rev Esp Cardiol, vol. 74, no. 9, pp. 790–799, Sep. 2021, doi: 10.1016/J.RECESP.2021.06.016.

[26] H. Jiang, J. Cai, Y. Lin, and Q. Wang, "Understanding the effect of TikTok marketing on user purchase behavior: a mixed-methods approach," Electronic Commerce Research, pp. 1–36, Aug. 2024, doi: 10.1007/S10660-024-09882-X/METRICS.

[27] J. D. Barquero Cabrero, B. Castillo-Abdul, J. A. Talamás-Carvajal, and L. M. Romero-Rodríguez, "Owned media, influencer marketing, and unofficial brand ambassadors: differences between narratives, types of prescribers, and effects on interactions on Instagram," Humanit Soc Sci Commun, vol. 10, no. 1, pp. 1–12, Dec. 2023, doi: 10.1057/S41599-023-01779-8/FIGURES/13.

[28] Q. Zhou, M. Sotiriadis, and S. Shen, "Using TikTok in tourism destination choice: A young Chinese tourists' perspective," Tour Manag Perspect, vol. 46, p. 101101, Mar. 2023, doi: 10.1016/J.TMP.2023.101101.

[29] M. Mustak, H. Hallikainen, T. Laukkanen, L. Plé, L. D. Hollebeek, and M. Aleem, "Using machine learning to develop customer insights from user-generated content," Journal of Retailing and Consumer Services, vol. 81, p. 104034, Nov. 2024, doi: 10.1016/J.JRETCONSER.2024.104034.

[30] S. Sherly and E. Ruswanti, "The Influence of EWOM Dimensions, Purchase Intention on Buying Behavior in Women's Clothing Products in Java Island," Eduvest - Journal of Universal Studies, vol. 4, no. 3, pp. 1322–1331, Mar. 2024, doi: 10.59188/EDUVEST.V4I3.1052

[31] J. A. F. Ortiz, M. De Los M. Santos Corrada, E. Lopez, V. Dones, and V. F. Lugo, "Don't make ads, make TikTok's: media and brand engagement through Gen Z's use of TikTok and its significance in purchase intent," Journal of Brand Management, vol. 30, no. 6, pp. 535–549, Nov. 2023, doi: 10.1057/S41262-023-00330-Z.

[32] X. Lv, C. Zhang, and C. Li, "Beyond image attributes: A new approach to destination positioning based on sensory preference," Tour Manag, vol. 100, p. 104819, Feb. 2024, doi: 10.1016/J.TOURMAN.2023.104819.

[33] M. D. ul Haq and C. M. Chiu, "Boosting online user engagement with short video endorsement content on TikTok via the image transfer mechanism," Electron Commer Res Appl, vol. 64, p. 101379, Mar. 2024, doi: 10.1016/J.ELERAP.2024.101379.

[34] D. He, Z. Yao, T. S. H. Teo, Y. Ma, and W. Xu, "How social learning drives customer engagement in short video commerce: An attitude transfer perspective," Information & Management, vol. 61, no. 6, p. 104018, Sep. 2024, doi: 10.1016/J.IM.2024.104018.

[35] B. Duivenvoorde and C. Goanta, "The regulation of digital advertising under the DSA: A critical assessment," Computer Law & Security Review, vol. 51, p. 105870, Nov. 2023, doi: 10.1016/J.CLSR.2023.105870.

[36] J. Sun, M. Sarfraz, L. Ivascu, H. Han, and I. Ozturk, "Live streaming and livelihoods: Decoding the creator Economy's influence on consumer attitude and digital behavior," Journal of Retailing and Consumer Services, vol. 78, p. 103753, May 2024, doi: 10.1016/J.JRETCONSER.2024.103753.

[37] F. I. Oktaviany, S. Senliana, and A. D. Lestari, "The Role of Digital Marketing in Culinary Business Dynamics on Tiktok @Jihannnpp Account.," Indonesian Journal of Advanced Research (IJAR), vol. 3, no. 7, pp. 1051–1062, Jul. 2024, doi: 10.55927/IJAR.V3I7.10299.

[38] A. B. Barcelona et al., "#Budolfinds: The Role of TikTok's Shopee Finds' Videos in the Impulsive Buying Behavior of Generation Z Consumers.," International Journal of Multidisciplinary: Applied Business & Education Research, vol. 3, no. 11, pp. 2316–2328, Nov. 2022, doi: 10.11594/IJMABER.03.11.18.

[39] R. A. Jamil, U. Qayyum, S. R. ul Hassan, and T. I. Khan, "Impact of social media influencers on consumers' well-being and purchase intention: a TikTok perspective," European Journal of Management and Business Economics, vol. 33, no. 3, pp. 366–385, Jun. 2024, doi: 10.1108/EJMBE-08-2022-0270.

[40] P. He, Q. Shang, W. Pedrycz, and Z. S. Chen, "Short video creation and traffic investment decision in social e-commerce platforms," Omega (Westport), vol. 128, p. 103129, Oct. 2024, doi: 10.1016/J.OMEGA.2024.103129.

[41] N. Paraskeva, S. Haywood, F. Hasan, D. Nicholls, M. B. Toledano, and P. C. Diedrichs, "An exploration of having social media influencers deliver a first-line digital intervention to improve body image among adolescent girls: A qualitative study," Body Image, vol. 51, p. 101753, Dec. 2024, doi: 10.1016/J.BODYIM.2024.101753.

[42] Y. Niu, Z. Huang, and Z. Fang, "Follower attraction in live streaming: Knowledge driven by PKM and data driven by EF-LSTM," Comput Human Behav, vol. 159, p. 108317, Oct. 2024, doi: 10.1016/J.CHB.2024.108317.

[43] A. Japutra, Y. Ekinci, and L. Simkin, "Discovering the dark side of brand attachment: Impulsive buying, obsessive-compulsive buying and trash talking," J Bus Res, vol. 145, pp. 442–453, Jun. 2022, doi: 10.1016/J.JBUSRES.2022.03.020.

[44] M. E. David and J. A. Roberts, "TikTok Brain: An Investigation of Short-Form Video Use, Self-Control, and Phubbing," Soc Sci Comput Rev, 2024, doi: 10.1177/08944393241279422.

[45] X. Wang and Y. Guo, "Motivations on TikTok addiction: The moderating role of algorithm awareness on young people.," El Profesional de la Información, vol. 32, no. 4, pp. 1–10, Jul. 2023, doi: 10.3145/EPI.2023.JUL.11.

[46] Q. Jiang and L. Ma, "Swiping More, Thinking Less: Using TikTok Hinders Analytic Thinking.," Cyberpsychology (Brno), vol. 18, no. 3, pp. 115–143, 2024, doi: 10.5817/CP2024-3-1.

[47] C. Elliott, E. Truman, and J. E. Black, "Tracking teen food marketing: Participatory research to examine persuasive power and platforms of exposure.," Appetite, vol. 186, p. N.PAG-N.PAG, Jul. 2023, doi: 10.1016/J.APPET.2023.106550.

[48] A. M. Izza, M. N. Ardiansyah, F. Barkah, and J. Romdonny, "SYNERGISTIC EFFECTS OF CONTENT MARKETING AND INFLUENCERS MARKETING ON THE FORMATION OF BRAND AWARENESS AND PURCHASE INTEREST OF TIKTOK SHOP USERS (CIREBON CITY CASE STUDY).," International Journal of Social Service & Research (IJSSR), vol. 4, no. 5, pp. 1339–1347, May 2024, doi: 10.46799/IJSSR.V4I05.781.

[49] J. Ren, J. Yang, M. Zhu, and S. Majeed, "Relationship between consumer participation behaviors and consumer stickiness on mobile short video social platform under the development of ICT: based on value co-creation theory perspective.," Inf Technol Dev, vol. 27, no. 4, pp. 697–717, 2021, doi: 10.1080/02681102.2021.1933882.

[50] L. Saulīte and D. Ščeulovs, "The Impact on Audience Media Brand Choice Using Media Brands Uniqueness Phenomenon," Journal of Open Innovation: Technology, Market, and Complexity, vol. 8, no. 3, p. 128, Sep. 2022, doi: 10.3390/JOITMC8030128.

[51] C. Ruiz-Viñals, M. Pretel Jiménez, and J. L. Del Olmo Arriaga, "THE IMPACT OF TIKTOK ON THE GENERATION OF ENGAGEMENT FOR FASHION BRANDS A profile analysis of Zara," VISUAL Review. International Visual Culture Review / Revista Internacional de Cultura, vol. 16, no. 4, pp. 167–181, Jul. 2024, doi: 10.62161/revvisual.v16.5289.

[52] S. Scherr and K. Wang, "Explaining the success of social media with gratification niches: Motivations behind daytime, nighttime, and active use of TikTok in China.," Comput Human Behav, vol. 124, p. N.PAG-N.PAG, Nov. 2021, doi: 10.1016/J.CHB.2021.106893.

[53] R. Yao, G. Qi, Z. Wu, H. Sun, and D. Sheng, "Digital human calls you dear: How do customers respond to virtual streamers' social-oriented language in e-commerce livestreaming? A stereotyping perspective," Journal of Retailing and Consumer Services, vol. 79, p. 103872, Jul. 2024, doi: 10.1016/J.JRETCONSER.2024.103872.

[54] P. Cillo and G. Rubera, "Generative AI in innovation and marketing processes: A roadmap of research opportunities," J Acad Mark Sci, 2024, doi: 10.1007/S11747-024-01044-7.

[55] L. Cox and T. Piatkowski, "Influencers and 'brain building' smart drugs: A content analysis of services and market activities of nootropic influencers over social media," Perform Enhanc Health, vol. 12, no. 4, p. 100289, Oct. 2024, doi: 10.1016/J.PEH.2024.100289.

[56] K. S. Meng and L. Leung, "Factors influencing TikTok engagement behaviors in China: An examination of gratifications sought, narcissism, and the Big Five personality traits," Telecomm Policy, vol. 45, no. 7, p. 102172, Aug. 2021, doi: 10.1016/J.TELPOL.2021.102172.

[57] R. Filieri, F. Acikgoz, and H. Du, "Electronic word-of-mouth from video bloggers: The role of content quality and source homophily across hedonic and utilitarian products," J Bus Res, vol. 160, p. 113774, May 2023, doi: 10.1016/J.JBUSRES.2023.113774.

[58] D. L. M. van der Bend, N. Gijsman, T. Bucher, V. A. Shrewsbury, H. van Trijp, and E. van Kleef, "Can I @handle it? The effects of sponsorship disclosure in TikTok influencer marketing videos with different product integration levels on adolescents' persuasion knowledge and brand outcomes.," Comput Human Behav, vol. 144, p. N.PAG-N.PAG, Jul. 2023, doi: 10.1016/J.CHB.2023.107723.

[59] C. Wang and Z. Li, "Unraveling the relationship between audience engagement and audiovisual characteristics of automotive green advertising on Chinese TikTok (Douyin)," PLoS One, vol. 19, no. 4 April, Apr. 2024, doi: 10.1371/journal.pone.0299496.

[60] E. (Emily) Ko, D. Kim, and G. Kim, "Influence of emojis on user engagement in brand-related user generated content," Comput Human Behav, vol. 136, p. 107387, Nov. 2022, doi: 10.1016/J.CHB.2022.107387.

[61] D. Menon, "Factors influencing Instagram Reels usage behaviours: An examination of motives, contextual age and narcissism," Telematics and Informatics Reports, vol. 5, p. 100007, Mar. 2022, doi: 10.1016/J.TELER.2022.100007.

[62] S. H. Liao, R. Widowati, and Y. C. Hsieh, "Investigating online social media users' behaviors for social commerce recommendations," Technol Soc, vol. 66, p. 101655, Aug. 2021, doi: 10.1016/J.TECHSOC.2021.101655.

[63] L. Sharakhina, I. Ilyina, D. Kaplun, T. Teor, and V. Kulibanova, "AI technologies in the analysis of visual advertising messages: survey and application," Journal of Marketing Analytics, vol. 12, no. 4, pp. 1066–1089, Dec. 2023, doi: 10.1057/S41270-023-00255-1/METRICS.

[64] Q. Han, C. Lucas, E. Aguiar, P. Macedo, and Z. Wu, "Towards privacy-preserving digital marketing: an integrated framework for user modeling using deep learning on a data monetization platform," Electronic

Commerce Research, vol. 23, no. 3, pp. 1701–1730, Sep. 2023, doi: 10.1007/S10660-023-09713-5.

[65] F. J. S. Lacárcel, R. Huete, and K. Zerva, "Decoding digital nomad destination decisions through user-generated content," Technol Forecast Soc Change, vol. 200, p. 123098, Mar. 2024, doi: 10.1016/J.TECHFORE.2023.123098.

[66] J. D. Brüns and M. Meißner, "Show me that you are advertising: Visual salience of products attenuates detrimental effects of persuasion knowledge activation in influencer advertising," Comput Human Behav, vol. 148, p. 107891, Nov. 2023, doi: 10.1016/J.CHB.2023.107891.

[67] J. Liao and J. Chen, "The authenticity advantage: How influencer authenticity management strategies shape digital engagement with sponsored videos," J Bus Res, vol. 185, p. 114937, Dec. 2024, doi: 10.1016/J.JBUSRES.2024.114937.

[68] S. Gallin and A. Portes, "Online shopping: How can algorithm performance expectancy enhance impulse buying?," Journal of Retailing and Consumer Services, vol. 81, p. 103988, Nov. 2024, doi: 10.1016/J.JRETCONSER.2024.103988.

[69] E. Golab-Andrzejak, "Measuring the effectiveness of digital communication – social media performance: an example of the role played by AI-assisted tools at a university," Procedia Comput Sci, vol. 225, pp. 3332–3341, Jan. 2023, doi: 10.1016/J.PROCS.2023.10.327.

[70] Dra. A. Huertas-Bailén, Dra. N. Quintas-Froufe, and Dra. A. González-Neira, "La Participación de la Audiencia en los Metadatos de TikTok.," Comunicar, vol. 32, no. 78, pp. 82–92, Jan. 2024, doi: 10.58262/V32I78.7.

[71] D. Shah, E. Webster, and G. Kour, "Consuming for content? Understanding social media-centric consumption," J Bus Res, vol. 155, p. 113408, Jan. 2023, doi: 10.1016/J.JBUSRES.2022.113408.

[72] J. M. Alcántara-Pilar, M. E. Rodriguez-López, Z. Kalinić, and F. Liébana-Cabanillas, "From likes to loyalty: Exploring the impact of influencer credibility on purchase intentions in TikTok," Journal of Retailing and Consumer Services, vol. 78, p. 103709, May 2024, doi: 10.1016/J.JRETCONSER.2024.103709.

[73] Q. Q. Liu, S. K. Yu, and Y. T. Yang, "The effects of sponsorship disclosure in short-form video: A moderated mediation model of sponsorship literacy and perceived features of sponsored short-form video.," Comput Human Behav, vol. 150, p. N.PAG-N.PAG, Jan. 2024, doi: 10.1016/J.CHB.2023.107969.

[74] K. M. Nguyen, N. T. Nguyen, N. T. Q. Ngo, N. T. H. Tran, and H. T. T. Nguyen, "Investigating Consumers' Purchase Resistance Behavior to AI-Based Content Recommendations on Short-Video Platforms: A Study of Greedy And Biased Recommendations.," Journal of Internet Commerce, vol. 23, no. 3, pp. 284–327, 2024, doi: 10.1080/15332861.2024.2375966.

[75] J. Grandinetti, "Examining embedded apparatuses of AI in Facebook and TikTok," AI Soc, vol. 38, no. 4, pp. 1273–1286, Aug. 2023, doi: 10.1007/S00146-021-01270-5.

[76] M. Zhang, P. Xu, and Y. Ye, "Trust in social media brands and perceived media values: A survey study in China," Comput Human Behav, vol. 127, Feb. 2022, doi: 10.1016/j.chb.2021.107024.

[77] J. Schwenzow, J. Hartmann, A. Schikowsky, and M. Heitmann, "Understanding videos at scale: How to extract insights for business research," J Bus Res, vol. 123, pp. 367–379, Feb. 2021, doi: 10.1016/J.JBUSRES.2020.09.059.

[78] G. Gao, H. Liu, and K. Zhao, "Live streaming recommendations based on dynamic representation learning," Decis Support Syst, vol. 169, p. 113957, Jun. 2023, doi: 10.1016/J.DSS.2023.113957.

[79] Q. Zhang, Y. Wang, and S. K. Ariffin, "Keep scrolling: An investigation of short video users' continuous watching behavior," Information & Management, vol. 61, no. 6, p. 104014, Sep. 2024, doi: 10.1016/J.IM.2024.104014.

[80] M. K. Merga, "How can Booktok on TikTok inform readers' advisory services for young people?," Libr Inf Sci Res, vol. 43, no. 2, p. 101091, Apr. 2021, doi: 10.1016/J.LISR.2021.101091.

[81] L. Chen, Y. Yan, and A. N. Smith, "What drives digital engagement with sponsored videos? An investigation of video influencers' authenticity management strategies," J Acad Mark Sci, vol. 51, no. 1, pp. 198–221, Jan. 2023, doi: 10.1007/S11747-022-00887-2.

[82] G. Cecere, C. Jean, F. Le Guel, and M. Manant, "Artificial intelligence and algorithmic bias? Field tests on social network with teens," Technol Forecast Soc Change, vol. 201, p. 123204, Apr. 2024, doi: 10.1016/J.TECHFORE.2023.123204.

[83] Indrawati, P. C. Putri Yones, and S. Muthaiyah, "eWOM via the TikTok application and its influence on the purchase intention of somethinc products," Asia Pacific Management Review, vol. 28, no. 2, pp. 174–184, Jun. 2023, doi: 10.1016/J.APMRV.2022.07.007.

[84] Y. Li, Y. Chang, and Z. Liang, "Attracting more meaningful interactions: The impact of question and product types on comments on social media advertisings," J Bus Res, vol. 150, pp. 89–101, Nov. 2022, doi: 10.1016/J.JBUSRES.2022.05.085.

[85] H. Zhang and X. Wang, "The 'backfire' effects of luxury advertising on TikTok: The moderating role of self-deprecating online reviews," Comput Human Behav, vol. 155, p. 108163, Jun. 2024, doi: 10.1016/J.CHB.2024.108163.

[86] S. H. Taylor and K. S. C. Brisini, "Parenting the TikTok algorithm: An algorithm awareness as process approach to online risks and opportunities," Comput Human Behav, vol. 150, p. 107975, Jan. 2024, doi: 10.1016/J.CHB.2023.107975

[87] D. Grewal, D. Herhausen, S. Ludwig, and F. Villarroel Ordenes, "The Future of Digital Communication Research: Considering Dynamics and Multimodality," Journal of Retailing, vol. 98, no. 2, pp. 224–240, Jun. 2022, doi: 10.1016/J.JRETAI.2021.01.007.

[88] S. Miranda, I. Trigo, R. Rodrigues, and M. Duarte, "Addiction to social networking sites: Motivations, flow, and sense of belonging at the root of addiction," Technol Forecast Soc Change, vol. 188, p. 122280, Mar. 2023, doi: 10.1016/J.TECHFORE.2022.122280.

[89] X. Yin, J. Li, H. Si, and P. Wu, "Attention marketing in fragmented entertainment: How advertising embedding influences purchase decision in short-form video apps," Journal of Retailing and Consumer Services, vol. 76, p. 103572, Jan. 2024, doi: 10.1016/J.JRETCONSER.2023.103572.

Data Mart Design to Increase Transactional Flow of Debit and Credit Card in Peruvian Bodegas

Juan Carlos Morales-Arevalo[1], Erick Manuel Aquise-Gonzales[2], William Yohani Carpio-Ore[3], Emmanuel Victor Mendoza Sáenz[4], Carlos Javier Mazzarri-Rodriguez[5], Erick Enrique Remotti-Becerra[6], Edison Humberto Medina-La Plata[7], Luis F. Luque-Vega[8]

Graduate School, Universidad Peruana de Ciencias Aplicadas, Lima, Perú[1,2,3,4,5,6,7]
Department of Technological and Industrial Processes, ITESO AC[8]

***Abstract*—The objective of this research is to design a Data Mart to identify tactical actions and increase the use of POS (points of sale) in the bodega business sector of Lima, Peru. A quantitative approach, using transaction history data, is applied using the Kimball methodology. This involves the ETL (Extract, Transform, Load) process to create a dimensional model and to develop a dashboard to visualize key indicators using Power BI. This solution is expected to improve the detection and analysis of transactional errors, categorized by geographic location and business sector while enhancing decision-making processes. This research improves the transactional flow and digital payment adoption in small businesses, fostering greater financial inclusion in the Peruvian market. Therefore, the methodology and tools to be applied in this research offer a framework as a model for similar contexts, especially in emerging markets, which will allow closing gaps in digital payment adoption and financial inclusion.**

***Keywords*—Business intelligence; Extract; Transform; Load (ETL); dashboard; data mart; Point of Sale (POS)**

I. Introduction

Every year, there is a higher penetration of digitalization in business areas, which allows not only for the increase in sales, but also makes it possible to reach financial inclusion, which could solve the growth necessities of businesses in Peru. A clear example of this is Yape [1], a digital wallet through which users can send and receive money with enough speed to acquire products in physical establishments, transfer money to family members, or pay for services. Bodegas are not foreign to this development, since, through Yape, they not only increase their sales (increasing their digital demand) but also have better odds of credit access, since banks can visualize their money flow.

While it is well known that Peruvians are increasingly making more purchases through digital means [2] (either digital wallets, credit or debit cards, use of QR codes, among others), the main use of these methods is still low when compared to the rest of Latin American countries. According to the financial inclusion index published by Credicorp, the use of cash in 2024 is still the main means through which citizens' money flows [3]. This represents a big opportunity not only for digitalization but also for bankization.

This is how companies in the digital payment sector aim to incentivize sales through points of sale (POS). POSs are tools that allow users to make payments with not only credit or debit cards, but also through digital wallets, and offer a range of value-added services which bring access to banking financial services, a key part of financial inclusion.

The specific objectives set for this paper are the following:

- Identifying the main transactional errors in the bodega sector associated with POS usage.
- Finding where the transactional errors are present in terms of geographic location and business sector.
- Comparing the bodega sector and the restaurant sector to measure the average ticket sales.
- Verifying if there is an increase or decrease in transactional errors every month.
- Validating if the number of active businesses using POS solutions is steady or if it suffers notable variations.

With this article, we introduce the reader to the use of digital payment methods, mainly points of sale (POS), in the context of Peru, and how the process of digitalization still has a long way to go in this country. We then examine the existing literature, which grants significant insights to the digital payment sector. Subsequently, we review relevant concepts regarding the methodology used, as well as important products, challenges and difficulties within the sector, and present our proposal based on Business Intelligence and the Data Mart architecture in order to create a model and, with it, a dashboard which will allow companies within the digital payment sector to make smart, data-based decisions in order to increase their presence and the transactional flow in Peruvian bodegas, where cash is still the main payment method. Finally, we discuss our findings, focusing on the advantages and benefits of our proposal.

II. Literature Review

In this section, we present a literature review of research relevant to the proposal described in this paper. It contains topics such as the use of digital payment methods and their benefits for clients, the challenges to adopting these methods, the differences between the users who are taking advantage of them and the ones who are not, and important aspects of the use of digital payment methods, among others.

According to Muhtasim, D. A., Yee Tan, S., Hassan, A., Pavel, M. I., and Susmit, S. (2022), security aspects such as authentication, encryption mechanisms and information

provided to clients are strictly related to the improvement of customer experience. Other security-related aspects that should not be excluded are the speed of transactions, software performance, and the privacy of the user's data [4]. It is important to keep all of this in mind to reduce the weak points of companies since maintaining good customer experience is a key task in doing so.

While security is a crucial factor for users to be more welcoming towards the use of digital payments, Alabdan, R. and Sulphey, M. M. (2020) indicate that there are key factors belonging to payment platforms which are important for their acceptance, such as ease of use, utility (how easy they are to set up, where they are accepted, etc.) and user awareness (how well they understand payment options or how to use the services). In addition to this, they found a small difference in the acceptance of mobile payments between men and women, obtaining an average value a bit higher in men [5]. This helps in focusing on the target sector that needs to be worked on.

We can complement this data to the findings of Aurazo, J., & Vega, M. (2021), who explain that the population using digital payments is usually between the age range of 25 to 40 years old, they're usually people with higher education, formal jobs, and are often located in urban areas with Internet access [6]. Their findings further help in narrowing down the potential target sector that digital payment companies should focus on.

Another study, conducted by Ünver, Ş., & Alkan, Ö. (2021), supports these characteristics and mentions that as users age, their odds of participating in e-commerce decrease. Similarly, as their income increases, they're more likely to participate in e-commerce. Furthermore, they found that users who are active in social media and online banking are more likely to use e-commerce [7]. This information represents a significant opportunity for financial inclusion within the digital payment sector, as a target segment has been identified that can be leveraged, not only to drive revenue growth, but also to benefit the broader communities. This is crucial, as noted by de Moraes, C. O., Roquete, R. M., and Gawryszewski, G. (2023), who found that access to finance, in general, has a direct relationship with income inequality in the population, but the impact is even greater regarding digital finance specifically. [8]

It has been demonstrated that e-commerce models can increase client acquisition, sale goals, and overall revenue, as indicated by Hafiz Yusoff, M., Alomari, M. A., Adilah, N., Latiff, A., and Alomari, M. S. (2019). [9] This means that expanding the reach of digital payment methods can generate a positive impact on the population, not to mention the effect that digital platforms have on businesses.

However, the challenges that users may face must be considered when wanting to implement digital payment platforms. Widayani, A., Fiernaningsih, N., and Herijanto, P. (2022) explain that there is a set of barriers that complicate the adoption of said platforms, such as the tradition barrier, which refers to innovative changes that clash with user's routines or traditions; the use barrier, which represents the incompatibility of user's habits with technological innovation or the psychological barrier, defined by user behavior such as mistrust, anxiety, lack of control of discomfort [10]. Identifying and knowing how to deal with these barriers is crucial to face the challenge of digital transformation and reach users more efficiently.

We can complement this with the research by Hermenegildo-Chávez, Martín-Ruiz, and Rondán-Cataluña (2023), who found that, in Peru, client loyalty on sale channels is influenced by the environment, more specifically, online and offline environments [11]. Clients have a better feeling of security for offline sales, and this is related to their trust and loyalty to the seller. This factor represents another of the challenges of implementing digital transformation, especially in the context of developing countries.

One way to counteract this is to improve the interoperability between electronic systems, as stated by Libaque-Saenz, C. F., Ortega, C., Rodriguez-Serra, M., Chong, M., and Lopez-Puente-de-la-Vega, S. (2024). Digital wallets from different providers need to communicate between themselves easily. Additionally, providers must focus on keeping scalable systems to generate future benefits for their customers in the long term [1]. This is important because any way to make customers feel comfortable using these platforms will help in dealing with the previously mentioned barriers.

The authors Alkan, O.; Küçükoğlu H. and Tutar, G., (2021) add that, considering age, gender, education level, and monthly income have an impact on online shopping, it would be beneficial if the providers of these services use this data to develop adequate marketing strategies to reach the necessary users. Additionally, they state that complaints, suggestions, or user requests may be received directly using field studies to determine the factors of user preferences and keep them in mind to personalize and improve their experience with the service [12]. These findings give us an idea of how companies in the digital payment sector can expand their reach and improve their relationship with their customers.

While these related works bring us valuable information and insights, they do not present practical proposals that take advantage of such insights. Our proposal consists of a solution for companies in the digital payment sector to visualize not only their general performance, but also to bring light to the gaps in this market, allowing them to focus on unattended parts of the sector, such as different types of businesses, as well as provinces or districts where digital payment methods are barely present, if at all.

This solution will also allow its users to find and manage errors, issues or problems with more efficiency and precision, so that they can work in improving their clients' experience and continue to provide them with a steady and reliable service.

III. Business Intelligence Methodology

Before determining the methodology to be used in this research, we will review the concepts of data warehouse, data mart, and the most known business intelligence methodologies today.

A data warehouse is defined as a collection of data that helps in the decision-making process of the entity in which it is used. A data warehouse stores copious amounts of data coming from diverse sources, which must have a coherent format for their

later exploitation (analyses, reports, etc.) by the organization [13].

It is also important to know how to differentiate a data warehouse from a data mart. While the first holds the entirety of the organization's data, data marts hold only a subgroup of it, focusing on a specific business area. They are both key components in the business intelligence architecture.

A. Inmon Methodology

Developed by Bill Inmon and his team in 1980, this methodology is centered on the creation of one centralized and complete data warehouse, which serves as a data source for the whole organization. The Inmon methodology involves designing the data warehouse based on the data sources and the entities and relationships present in them. The normalized data warehouse is then used to feed several data marts which adapt to the specific commercial needs and use cases [14] (Fig. 1).

Key aspects of this methodology include:

- The data warehouse is very flexible to change.
- Business processes can be understood very easily.
- Reports can be handled by different enterprises.
- The ETL process is not as error prone.

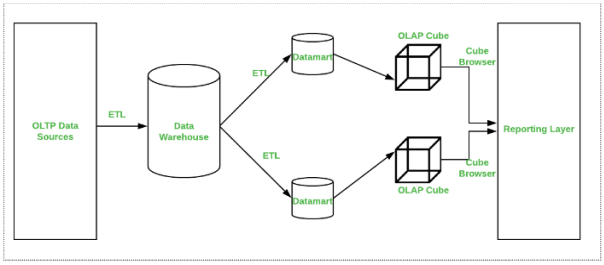


Fig. 1. Architecture according to the Inmon model. [14]

B. Kimball Methodology

A methodology developed by Ralph Kimball and his team in 1990, the Kimball methodology consists of delivering data to end users in the fastest and most efficient way possible. It implies designing the data warehouse based on business processes and key performance indicators [15].

Kimball's data model follows a down-up approach for the architecture of the data warehouse, in which data marts are formed first, according to the company's commercial requirements [16] (Fig. 2). Characteristics of this methodology include:

- Fast configuration and construction.
- When compared to the multiple-star schema, report generation is highly successful.
- Highly effective database operations.
- Takes up less space in the database.
- Easy to manage.

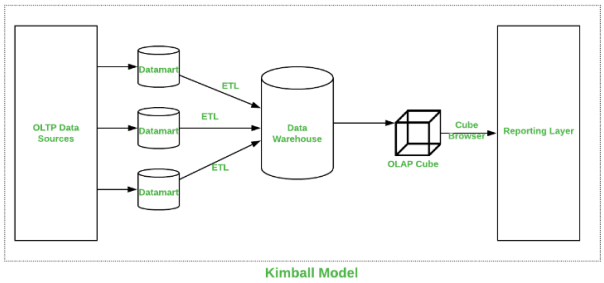


Fig. 2. Architecture according to Kimball's model. [15]

C. Data Vault Methodology

Created by Dan Linstedt in 2000, it is based on creating a scalable, agile, and resistant data warehouse. This methodology presents the design of the data warehouse around data history, auditability, and traceability. The data warehouse consists of a 3-layer data model which includes hubs, links, and satellites. This methodology is adequate for organizations that need flexible and adaptable data architecture, a high-performance data load process, and an auditable data history registry [14] (Fig. 3, 4).

Attributes of this methodology include:

- Scalability: Data vaults are highly scalable, which means they can manage large amounts of data and easily add new data sources. This makes it an ideal solution for fast-growing companies.
- Flexibility: The data vault is highly adaptable and can manage changes in data without having to restructure the entire model. This means that in the future, it can be easier and cheaper to make changes.
- Data history: Data vaults keep a complete history of the data, which allows for historic analyses and data comparisons.

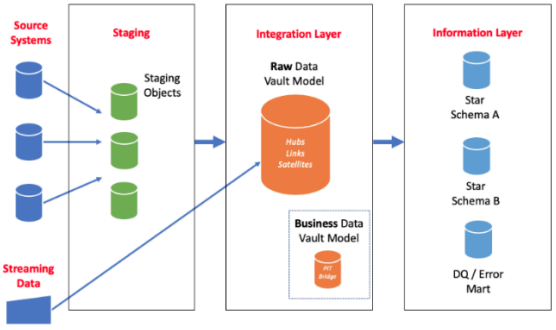


Fig. 3. Data Vault model architecture. [17]

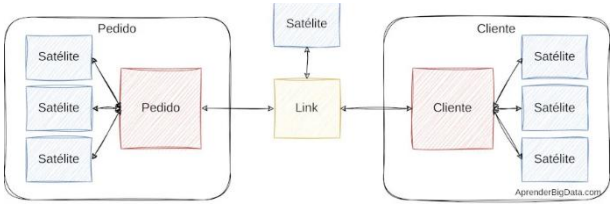


Fig. 4. Example of internal components of the Data Vault model. [18]

For this study, Ralph Kimball's methodology was chosen because the model is being proposed for the operative area of companies in the digital payment sector, for which a data mart would be used to measure the desired indicators, and in the future, more data marts or even a more robust data warehouse could be implemented as each company and/or their products grow.

IV. About the Digital Payment Sector

In Peru, the digital payment sector is playing a key role in the financial inclusion and growth of small businesses. This is what the companies in this sector strive to do – to offer new tools that assist businesses in moving to digital payment systems, thus assisting the businesses in managing their finances and surviving in the competitive market of today and tomorrow. Technology is not the focus; it is about empowering people and creating lasting opportunities [19]. The goals of the digital payment sector include this commitment:

- Putting customers first: Making sure businesses have the tools and support they need to succeed on their financial journey.
- Creating meaningful impact: Developing solutions that enable small businesses and enhance the economic development of communities.
- Fostering growth and adaptability: Always improving to overcome any given challenge and adapt to changing needs.
- Building trust through communication: Keeping open and honest conversations to build up the right connections with clients and stakeholders.
- Investing in talent: Attracting and nurturing talented people to lead the way to innovation and excellent service.
- Championing digital transformation: Using creativity and technology to turn problems into opportunities and provide useful, forward-thinking solutions.

A. Products in the Digital Payment Sector

The digital payment sector provides several innovative solutions for the needs of businesses and consumers. These include.

1) POS Systems: This first product offers a wireless POS, which has contactless payment technology and accepts all Visa and MasterCard debit and credit cards. To use this service, customers don't need to pay monthly, only the POS must be paid for and it has a regular price of S/.425. Commission for national cards is 3.89% + IGV and 4.99% + IGV for international cards.

2) Online payment platforms: This service allows for an easy and fast integration of plugins and APIs. This service also accepts payments from all credit and debit cards, with no monthly costs, which means only successful payments have a cost. It has the following commission values:

a) National cards: 4.20% + $0.30 + IGV

b) International cards: 5.49% + $0.30 + IGV

c) PagoEfectivo (a local online payment platform): 4.20% + $0.30 + IGV

B. Identified Challenges in the Digital Payment Sector

- Experience during the collecting payment process seems stagnant and does not take advantage of current technologies in procedures such as sales counting, order history, and efficiency in balancing.
- Service coverage is and always will be a weak point because payment methods always require an active internet connection for optimal functionality. This also depends on the location of the business. For example, when having a large concentration of customers in the same place, a weaker network hinders customer experience even more.
- Having a complex interface can make it difficult for customers and staff to navigate, operate and customize the software. It can also increase the risk of errors, confusion, and frustration.
- Limited functionality when the software does not have features or integration necessary for the business. For example, it is possible that a client needs software that accepts multiple payment methods.
- Security issues when the software is vulnerable to piracy, malware, or data violations that could compromise the business and/or the client's information.
- Bad customer service when the software provider does not offer support, orientation, or appropriate and timely problem solutions for the customer's issues or questions. Bad customer service can cause frustration, a feeling of helplessness, and dissatisfaction, and it can affect the performance and continuity of the business.

C. Opportunities and Improvements in the Digital Payment Sector

- Better POS payment experience: "Less is more" [20] defines the main needs of businesses in their day-to-day routine. Taking fewer steps to make a sale means more sales will be conducted. This is where the POS industry has a big opportunity: being able to bring more efficient experience in the payment process, reducing the number of steps, and streamlining the interactions between the business, the sale, and the client.
- Better financial inclusion: It is estimated that only 46.1% of the Peruvian population [3] is financially included. Meaning that they know, trust, and can access financial services (also known as credit). However, there's a social development barrier present, since a large part of society doesn't represent a financial benefit in the eyes of banking entities, which is why growth opportunities are lower or almost null for this sector. The insertion of digital payment methods aims to bring banking access to the income and expense flows, so that financial entities gain transparency and trust, and in this

way, make them more willing to bring their services to this population sector.

- Acceleration in digital transformation: The use of cash in Peruvian businesses is still relevant due to the low penetration of banking in society, in general. The increase in the use of digital wallets such as Yape or Plin impulses citizens to not only make digital transactions but also pushes business owners into accepting these payment methods in their establishments. There is a big challenge here since despite this increase in digital wallet use, there is still fear of using them among the population, mainly because of education (or lack thereof) reasons [3].
- Business analytics is a competitive advantage: While sales in the commerce sector increase, new opportunities for business growth arise, such as offering new products or services, making marketing campaigns, and hiring more staff, among others. It is here where competitors emerge and analytics gain relevance, because having metrics and more sector knowledge helps in making better decisions; for example, if a business owner knows that most of his clients pay using DINERS cards, he could make loyalty campaigns with this brand, increasing consumer recurrence.

V. Data Source Analysis (Dataset)

The dataset used for this research is based on information acquired from the transactional database (Data Lake) of a company in the digital payment sector, but it was slightly modified for confidentiality reasons. It was obtained in Excel format under the following structures:

TABLE I. Business Dataset

N°	Field	Description
1	Id_comercio	Business unique code
2	Tipo_doc_tributario	Tax document type (RUC20, RUC10, DNI)
3	Departamento	Department where the business is located
4	Provincia	Province where the business is located
5	Distrito	District where the business is located

The first dataset, shown in Table I, contains information regarding many businesses and it includes their unique ID, the document type corresponding to each of them (usually the ID number of the legal representative), and their location, which consists of three fields: department, province and district. The second dataset, shown in Table II, contains much more technical information. It has data related to the errors occurring during transactions, including the denial code, the description of said code and the sum corresponding to the transaction amounts for transactions with errors.

The third dataset is much smaller but still has valuable information. Shown in Table III, it stores the codes and descriptions for the business sectors, and the ID of every business, effectively linking each business to its corresponding sector.

Finally, the fourth dataset, shown in Table IV, is the most important. It holds the data corresponding to the details of the transactions made by each business. It stores the month, the number of transactions made by the business during said month, the values of these transactions, and how many of them were successful.

TABLE II. Detail of POS Errors During Authorization

N°	Field	Description
1	Id_comercio	Business unique code
2	Action_code	Denial code during transaction attempt
3	Desc_respuesta_trx	Error code description
4	GPV	Sum of transaction amounts for failed transactions

TABLE III. Commerce Sector Detail by Business

N°	Field	Description
1	Id_comercio	Business unique code
2	Cod_mcc	Commerce sector code
3	Desc_mcc	Description for the commerce sector

TABLE IV. Transaction Detail by Business

N°	Field	Description
1	Cod_mes	Month code
2	Id_comercio	Business unique code
3	trx	Number of transactions made by the business
4	GPV	Sum of transaction amounts for this business
5	Trx_exitosas	Number of successful transactions for this business

VI. Indicators or Business Metrics

The following indicators were identified as important for this research:

- Conversion of successful transactions in bodegas using POSs: This is essentially the main indicator for companies in the digital payment sector. Being able to track the number of successful transactions is of utmost importance in this area, as it will help companies assess their performance and key results, and it also facilitates the process of finding errors in any of the systems used throughout the payment processes.
- Use of digital payment methods in bodegas which use POSs: another key metric is how much digital payment methods are being used in bodegas. As stated in the literature review section of this paper, it has been proven that the use of digital payments has positive and noticeable impacts in businesses in general. Because of this, it's important that business owners (especially the older, sometimes more informal parts of the population) learn about these benefits to increase their income. Measuring this can lead to important insights for future works in this sector.

- Number of bodegas with POSs that also have access to financial services (credit, maintenance discounts): tracking this number can lead to insights on how the access to financial products impacts the number of transactions carried out by each business.
- Market share in POSs in the bodega sector in Peru: Keeping track of the market share for POSs can help determine which regions have the biggest growth potential and help companies direct their efforts into entering these regions.

VII. Business Intelligence Architecture

To apply Business Intelligence (BI) in any environment, an architecture capable of converting a company's operational information into useful information for the strategic and tactical areas of the organization is necessary [21]. The standard transactional model, called OLTP (online transaction processing) which companies use, does not allow adequately efficient access to the information needed to make business decisions. Because of this, BI solutions propose the use of the OLAP (online analytical processing) system (shown in Fig. 5), in which specialized data repositories, called data warehouses or data marts are used, as mentioned previously in this paper. These data repositories allow information queries to be much faster because of the way they structure data, and they are fed by the company's traditional sources of information, which could be databases or spreadsheets, among others.

The use of these repositories will allow users (usually staff in management positions), to visualize the information they need through dashboards adequately and clearly, according to the company's business sector and personal preferences. This will help them make key business decisions. Business Intelligence architectures are usually composed of the following elements/processes:

- Data sources: Any file system, databases, or other sources of information the company uses in its standard model to handle daily transactions. For example: plain text files, Excel files, MySQL or Oracle databases, among others.
- Extraction, transformation, and load (ETL) process: it's the process through which the company's information, which can have several origins, is combined into a single centralized repository, in which it's cleaned and organized according to the company's needs [22].
- Data storage (Data Warehouse/Data Mart): The data repositories that handle the storing of information provided from the data sources. They can store structured, unstructured, or semi-structured information.
- Data exploitation: The final stage of the process, where users can visualize the information they need easily and quickly, through one or more dashboards. This leads to a simpler business decision-making process.

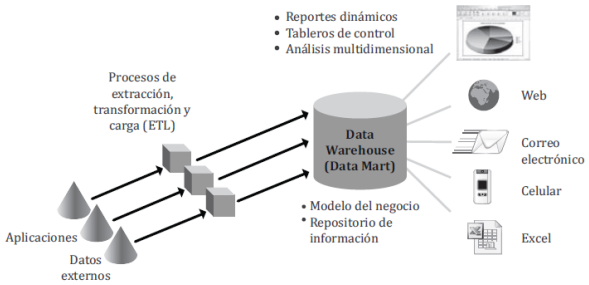


Fig. 5. Architecture of a Business Intelligence Solution [21].

VIII. ETL Process and Dimensional Model

The objective of the dimensional model is to facilitate information queries for users. For this case, the star model was chosen because, due to the simplicity of the data structures, it was not deemed necessary to go for the snowflake model, which usually consists of more complex relationships between tables. The star model, on the other hand, consists of a main fact table located in the center, which is related to several dimensional tables that are relevant to the business environment. The fact table contains information about transactions, which include the business code, the month and year of the transaction, the number of transactions, the value of the completed sales, error codes and descriptions (if any), and the code and description of the commerce sector. In addition to this table, the following dimensions were presented:

- Dim_Cod_Comercios: Contains information related to the business such as its code, document information, and geographic location, shown in Fig. 7.
- Dim_Cod_Rubros: Contains information related to the codes and names of the possible commerce sectors for each business. For example: transportation services, food sales, clothing sales, and medical services, among many others.
- Dim_Cod_Errores: Contains the codes and names for any errors that could occur when attempting to make a transaction.
- Dim_Tiempo: This dimension is usually considered necessary on any dimensional model. In this case, it contains the values for the transactions' months and years.

As we can see in Fig. 6, and according to what we've described above, the fact table stores the information from the fields of all dimensions. This allows queries to be made efficiently using dimensions as filters. For example, we could visualize all the transactions from August which generated a 116 error (insufficient funds). This ease of query is crucial for the business's continuous improvement, since it can be executed very quickly and allows data to be shown clearly and concisely, as explained in the results section.

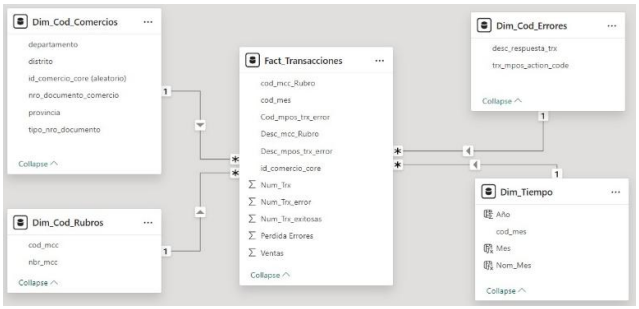


Fig. 6. Proposed dimensional model.

However, before starting with the dimensional modeling, the program Pentaho Data Integration was used as the main tool to conduct the ETL process. First, partial information was extracted from the transactional database to an Excel file (.xlsx). Then, for the transformation stage of the process, the data was ordered, filtered (to remove information with null values), and cleaned to remove any duplicate values, maintaining good data quality for the model. This process was repeated for all the dimensions proposed in the model.

Once the transformations were completed for each dimension, the corresponding transformation to generate the fact table was executed. This consisted of joining the data sources of the transactions, errors, and commerce sectors, as shown in Fig. 8. Then, the data was loaded to Power BI to start with the creation of the dashboard.

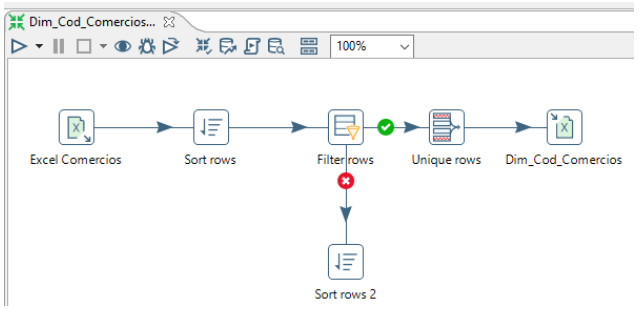


Fig. 7. ETL process carried out in Pentaho Data Integration for the Dim_Cod_Comercios dimension.

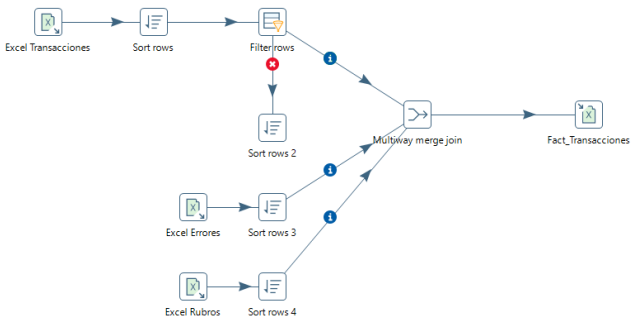


Fig. 8. Pentaho Data Integration view for the generation of the fact table, Fact_Transacciones.

IX. Results

Power BI was used as the tool to create a dashboard separated into pages to keep information organized. As shown in Fig. 9, the menu view was generated as the first page, which consists of a set of buttons that takes the user to the desired page. Each button is labeled with the information shown on each page. All these pages represent the most important metrics for the proposed solution and are the ones that generate the most value to the company's decision-making process.

The dashboard with transaction information shown in Fig. 10 allows users to visualize the business status quickly, as it presents a concise view of the number of transactions carried out to date, how many of them have been successful and how many have had errors, plus the value of sales. All of this is during the selected month in the filter located above.

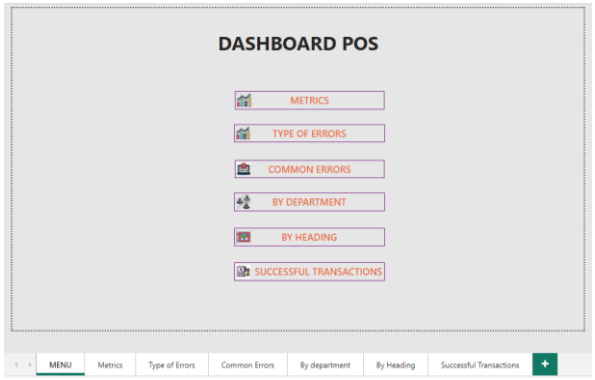


Fig. 9. Menu view.

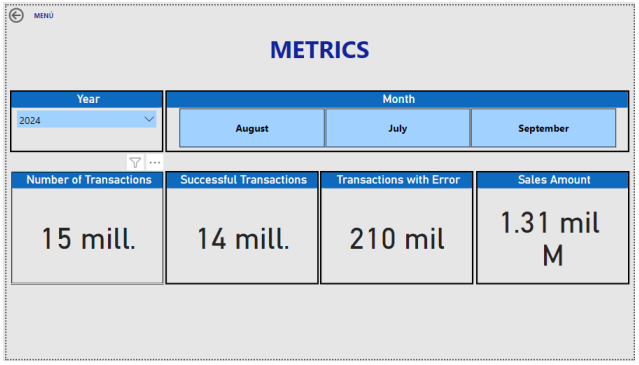


Fig. 10. Dashboard with transaction indicators.

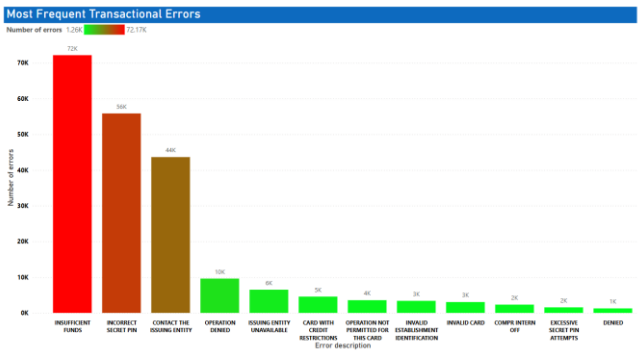


Fig. 11. Graphic showing the most frequent errors in transactions.

On the other hand, information about error types has been presented on several graphs. The most frequent errors (shown in Fig. 11), the number of transactions in Lima with errors and which of them, the number of failed transactions by commerce

sector (shown in Fig. 12), the distribution of departments with the highest number of errors (shown in Fig. 13), and a monthly comparison between the number of failed and successful transactions are shown in the dashboard's following pages.

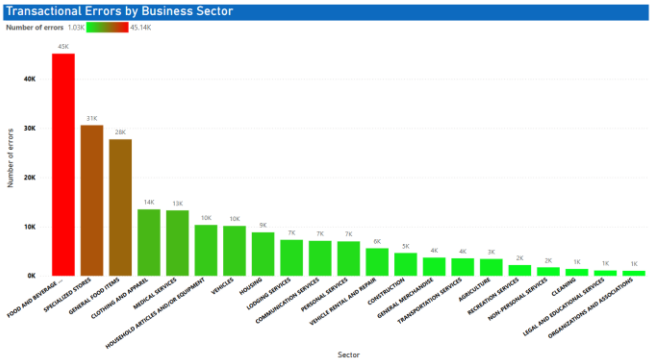


Fig. 12. Graph showing the number of errors by business sector.

These graphs are particularly important because, in the sector of e-commerce, there are many reasons why a transaction can fail, such as connection problems, internal errors in the bank or banks involved in the transaction, or even errors caused by the customer, such as having insufficient funds or incorrectly entering their PIN. Knowing which of these errors is the most common allows the management to determine quickly where to find problems in the business process and how to diagnose them effectively.

In addition to this, being able to visualize the distribution of errors by dimension is important to facilitate diagnostics and give an idea of where to focus efforts to prevent these errors in the future. The information shown must be treated as analytics that will contribute to increasing the company's value.

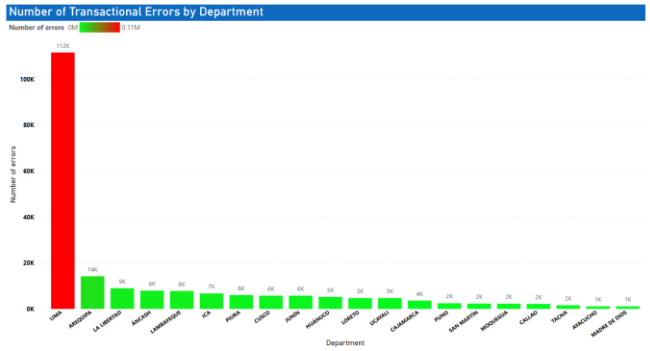


Fig. 13. Distribution of departments by their number of transactions with errors.

In Fig. 14, a comparison between the number of failed transactions and successful ones is shown by month. This information can provide benefits such as:

- Fraud detection: A high number of failed transactions in a certain month could represent fraudulent activity.
- User satisfaction: A high number of successful transactions can be a sign of good system performance and, therefore, a good customer experience. Similarly, a small number of errors can improve customer loyalty.
- Comparative analyses: Transaction data can be compared monthly to observe performance and allow the company to establish realistic goals for determined periods.
- Continuous improvement: Knowing the relation between failed and successful transactions allows management to determine the decisions they will make in the future.

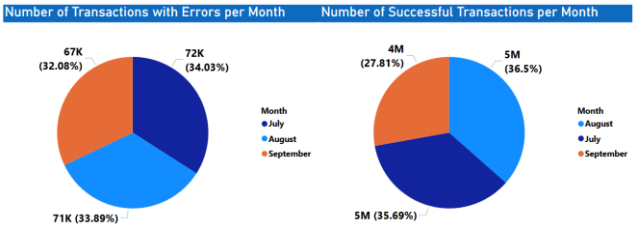


Fig. 14. Difference between the number of successful and failed transactions each month.

X. Discussion

This paper demonstrates the importance of the use of Business Intelligence in the sales sector. Through the identification of the problem, the setting of objectives, the application of the dimensional model, the ETL process, and finally the data visualization in dashboards, business management, and decision-making can be facilitated, since the company's weak points and their causes can be identified quickly.

The proposed Business Intelligence solution can help companies in the payment sector increase their income and improve the user's experience, if the dashboard is constantly monitored, and the proper measures are taken to address problems.

The objectives set at the beginning of the paper were accomplished, since the implementation of the dashboard allows for clear, quick, and concise visualization of the transactional errors, which commerce sectors and locations they occur in, and analyzing how many transactions are being carried out successfully or failing every month, which brings great analytical value to the company. In future work, it would be beneficial to analyze how viable the implementation of a data warehouse is for bigger companies in the digital payment sector, so that they can apply this model to a much larger and scalable volume of data.

Although the proposed solution demonstrates a clear improvement in the decision-making process for the digital payment sector, it is important to acknowledge certain limitations. One of these is the current scope of the Data Mart, which is restricted to specific transactional data from a single company. Expanding this scope to integrate additional data sources, such as customer feedback or competitor analysis, could provide a more comprehensive view of the business landscape. Moreover, while the Kimball methodology was effective in this case, it may not be suitable for all contexts. A future comparative study between Kimball and other methodologies, such as Data Vault or Inmon, could offer valuable insights into the best approach for companies with

similar needs. Additionally, the model's reliance on accurate data collection means that any issues in the ETL process, such as missing or incomplete data, could affect the quality of the analyses. Addressing these limitations and exploring opportunities for further optimization will be crucial to ensuring the long-term success and scalability of the solution. Finally, this study contributes to business intelligence by demonstrating how tactical solutions can drive digital transformation in small businesses, particularly in emerging markets where financial inclusion remains a significant challenge.

Therefore, Business Intelligence in transactional data analysis has proven to be a powerful tool for improving decision-making, particularly in sectors like e-commerce. Recent research has shown how implementing big data analysis models can optimize decision processes by extracting hidden patterns in transactions and enhancing the customer experience [23]. This approach applies to our solution for analyzing transactions in bodegas, where POS systems provide valuable insights into consumer behavior.

Furthermore, studies on the application of machine learning for customer segmentation suggest that combining these analytical approaches with Business Intelligence enables a more precise classification of customers based on their transactional behavior and demographic characteristics [24]. Similar to our research, combining these analytical approaches could lead to increased financial inclusion, allowing companies in the digital payment sector to better customize its services and offer solutions that more effectively meet the needs of bodegas.

Finally, as Bouchra et al. (2019) indicate, including the context in Data Mart design can further personalize BI solutions, adapting the information to the various user profiles [25]. This is particularly relevant to our study, where the bodega sector exhibits very specific characteristics that must be considered in the system's design.

XI. Conclusion

The design of the Data Mart presented in this work becomes a key tool to understand and improve the transactional flows in the bodegas of Lima, Peru. Thanks to the implementation of the Kimball methodology, it has been possible to accurately identify the most frequent errors in transactions and categorize them according to their geographic location and commercial sector, allowing more informed decisions to be made and optimizing digital payment operations.

The solution also has an important impact on financial inclusion, as it enables bodegas to adopt digital payment methods that, in addition to increasing their competitiveness, facilitate access to financial services such as credit. This development encourages the transition from cash to digital payments, driving the digitization of one of the most traditional sectors of local commerce.

One of the current limitations of the Data Mart is that it only deals with a single company's transactional data. Adding new data sources like customer feedback or competitive market analysis to a Data Warehouse would provide a more complete perspective and allow for the development of stronger strategies.

Using tools like Power BI has been useful in converting complex data into simple and easily understandable information. The dashboards created contain vital metrics; for example, transaction success and failure rates, and recurring error patterns, thus enabling accurate interventions and constant service enhancement.

The application of Business Intelligence solutions can make a difference in the management of small businesses in emerging markets by converting data into knowledge that creates value. It also provides a new opportunity to investigate technologies such as machine learning for predictive analytics and customer segmentation which may lead to better results.

Therefore, the proposed model is not only useful for improving transactional processes, but also for supporting the progress of the digital transformation of small businesses; this project has the potential to contribute to the development of the business environment and the reduction of barriers to financial inclusion particularly in the Peruvian market which needs such solutions.

References

[1] Libaque-Saenz, C. F., Ortega Ariza, C. P., Rodriguez-Serra, M., Chong, M., & Lopez-Puente-de-la-Vega , S. (2024). The role of interoperability and inter-side benefits on merchants' e-wallet adoption: The case of Peruvian nanostores. Industrial Management and Data Systems, 124(1), 64-84. https://doi.org/10.1108/IMDS-04-2023-0238

[2] Brito, J. B. G., Bucco, G. B., Heldt, R., Becker, J. L., Silveira, C. S., Luce, F. B., & Anzanello, M. J. (2024). A framework to improve churn prediction performance in retail banking. Financial Innovation, 10(1), 1-29. https://doi.org/10.1186/s40854-023-00558-3

[3] Grupo Crédito S.A. (2024). Índice de inclusión financiera de Credicorp 2024. Credicorp, Depósito legal: 202408884, 2024.

[4] Muhtasim, D. A., Yee Tan, S., Hassan, A., Pavel, M. I., & Susmit, S. (2022). Customer Satisfaction with Digital Wallet Services: An Analysis of Security Factors. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 13, Issue 1). https://doi.org/10.14569/IJACSA.2022.0130124

[5] Alabdan, R., & Sulphey, M. M. (2020). Understanding proximity mobile payment acceptance among Saudi individuals: An exploratory study. *International Journal of Advanced Computer Science and Applications*, *11*(4), 264–270. https://doi.org/10.14569/ijacsa.2020.0110436

[6] Aurazo, J., & Vega, M. (2021). Why people use digital payments: Evidence from micro data in Peru. *Latin American Journal of Central Banking*, *2*(4). https://doi.org/10.1016/j.latcb.2021.100044

[7] Ünver, Ş., & Alkan, Ö. (2021). Determinants of e-Commerce Use at Different Educational Levels: Empirical Evidence from Turkey e-Commerce Use at Different Educational Levels. *International Journal of Advanced Computer Science and Applications*, *12*(3), 40–49. https://doi.org/10.14569/IJACSA.2021.0120305

[8] de Moraes, C. O., Roquete, R. M., & Gawryszewski, G. (2023). Who needs cash? Digital finance and income inequality. *The Quarterly Review of Economics and Finance*, *91*, 84–93. https://doi.org/10.1016/J.QREF.2023.07.005

[9] Hafiz Yusoff, M., Alomari, M. A., Adilah, N., Latiff, A., & Alomari, M. S. (2019). Effect of e-Commerce Platforms towards Increasing Merchant's Income in Malaysia. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 10, Issue 8). https://doi.org/10.14569/ijacsa.2019.0100860

[10] Widayani, A., Fiernaningsih, N., & Herijanto, P. (2022). Barriers to digital payment adoption: micro, small and medium enterprises. *Mangament & Marketing. Challenges for the Knowledge Society*, *17*(4), 528–542. https://doi.org/10.2478/mmcks-2022- 0029

[11] Hermenegildo, M. Ruiz, D. Rondan, F. (2023). Point of sale loyalty analysis considering store environment (online and offline) and recency

of purchase. Revista Brasileira de Gestion de Negocios. https://doi.org/10.7819/rbgn.v25i4.4246

[12] Alkan, Ö., Küçükoğlu, H., & Tutar, G. (2021). Modeling of the Factors Affecting e-Commerce Use in Turkey by Categorical Data Analysis. *International Journal of Advanced Computer Science and Applications*, *12*(1), 95–105. https://doi.org/10.14569/IJACSA.2021.0120113

[13] Gascón, S. "Data Warehouse: ¿qué es el Data Warehouse y para qué sirve?" Hiberus Blog, 2021. Accesed: september 5th, 2024. [Online]. Available at: https://www.hiberus.com/crecemos-contigo/enfoques-de-data-warehousing/

[14] Geeksforgeeks.org, 2022. "Difference between Kimball and Inmon". Accesed: 5 de september de 2024. Available at: https://www.geeksforgeeks.org/difference-between-kimball-and-inmon/

[15] Arya, G., Beierschoder, M., Joshi, J. "¿Cuáles son las mejores metodologías de almacenamiento de datos?" LinkedIn, s.f. Accesed: 5 de september de 2024. [Online]. Available at: https://www.linkedin.com/advice/0/what-best-data-warehousing-methodologies-skills-data-warehousing?lang=es&originalSubdomain=es

[16] Naeem, T. "Conceptos de Data Warehouse: enfoque de Kimball vs. Inmon" astera.com. Accesed: september 6th, 2024. [Online]. Available at: https://www.astera.com/es/type/blog/data-warehouse-concepts/

[17] Srivastava, D. "Data vault builder" (01 Ago, 2022) LinkedIn.com, Accesed: september 5th, 2024. [Online]. Available at: https://www.linkedin.com/pulse/data-vault-builder-darshika-srivastava/

[18] Fernández, O. "Data Vault: Cómo estructurar tu Data Warehouse" (06 Ago, 2024) aprenderbigdata.com Accesed: 5 de setiembre de 2024. [Online]. Available at: https://aprenderbigdata.com/data-vault/

[19] Julião, J., Ayllon, T., Gaspar, M. (2023). Financial Inclusion Through Digital Banking: The case of Peru. In: Machado, J., et al. Innovations in Industrial Engineering II. icieng 2022. Lecture Notes in Mechanical Engineering. Springer, Cham. https://doi.org/10.1007/978-3-031-09360-9_24

[20] Zabalbeascoa, A. (2014). "Mies van der Rohe: Menos es más". España: El País.

[21] Medina, E. (2012). Business Intelligence: Una guía práctica, 2ª ed. Lima: Universidad Peruana de Ciencias Aplicadas.

[22] Amazon Web Services (AWS). ¿Qué es extracción, transformación y carga (ETL)? [Online]. Accessed september 4th, 2024. Available at: https://aws.amazon.com/what-is/etl/#seo-faq-pairs#what-is-etl

[23] El Falah, Z., Rafalia, N., & Abouchabaka, J. (2021). An intelligent approach for data analysis and decision making in big data: A case study on e-commerce industry. *International Journal of Advanced Computer Science and Applications, 12(7)*. https://doi.org/10.14569/IJACSA.2021.0120783

[24] Zineb, E. F., Rafalia, N., & Abouchabaka, J. (2021). Machine learning mini batch K-means and business intelligence utilization for credit card customer segmentation. *International Journal of Advanced Computer Science and Applications, 12(10)*, 220-225. https://doi.org/10.14569/IJACSA.2021.0121036

[25] Bouchra, A., Larbi, K., Abderrahim, A. W., & Sekkaki, A. (2019). Linking context to data warehouse design. *International Journal of Advanced Computer Science and Applications, 10(1)*. https://doi.org/10.14569/IJACSA.2019.0100102

Evaluation of Convolutional Neural Network Architectures for Detecting Drowsiness in Drivers

Mario Aquino Cruz, Bryan Hurtado Delgado, Marycielo Xiomara Oscco Guillen
Departamento Académico de Informática y Sistemas, Universidad Nacional Micaela Bastidas de Apurímac, Abancay, Perú

***Abstract*—Drowsiness in drivers is a condition that can manifest itself at any time, representing a constant challenge for road safety, especially in a context where artificial intelligence technologies are increasingly present in driver assistance systems. This paper presents a comparative evaluation of convolutional neural network (CNN) architectures for drowsiness detection, focusing on the identification of signals such as eye state and yawning. The research was of an applied type with a descriptive level, comparing the performance of LeNet, DenseNet121, InceptionV3 and MobileNet under challenging conditions, such as lighting and motion variations. A non-experimental design was used, with two datasets: a public dataset from Kaggle that included images classified into two categories (yawn and no yawn) and another created specifically for this study, which included images classified into three main categories (eyes open, eyes closed and undetected). The results indicated that, although all architectures performed well in controlled conditions, MobileNet stood out as the most accurate and consistent in challenging scenarios. DenseNet121 also showed good performance, while LeNet was effective in eye-state detection. This study provided a comprehensive assessment of the capabilities and limitations of CNNs for applications in drowsiness monitoring systems, and suggested future directions for improving accuracy in more challenging environments.**

Keywords*—Architectures; detection; drowsiness; neural networks

I. Introduction

Drowsiness at the wheel is a common problem that negatively impacts road safety worldwide. It is defined as the biological need for sleep, which can be caused by fatigue, sleep deprivation or medical conditions, affecting the driver's concentration and reaction time [1]. This condition poses a considerable danger to drivers, as it severely limits their ability to respond, increasing the risk of accidents, especially during long or night-time journeys. Moreover, it compromises not only the safety of the driver, but also that of passengers and other road users [2]. This problem is more serious in situations such as long working hours or lack of sleep.

According to the World Health Organization (WHO), road traffic accidents are responsible for approximately 1.19 million deaths each year, with men being three times more likely than women to lose their lives in these incidents [3]. An analysis of 208,727 passenger vehicle records involved in fatal crashes in the United States between 2017 and 2021 revealed that 17.6% of the cases were related to drowsy drivers [4]. In Spain, the Dirección General de Tráfico (DGT) stated that between 15% and 30% of vehicle accidents are directly or indirectly caused by drowsiness [5]. This problem is not only evident in developed countries, but also generates concern in Latin American countries such as Peru, which according to the National Institute of Statistics and Informatics (INEI) of Peru, 116,659 traffic accidents were recorded in 2016, of which 0.97% (1,131 accidents) were attributed to driver tiredness or fatigue, with Lima being the most affected region with 813 accidents, followed by Puno with 38 and Arequipa with 34 [6]. These data highlight the need to implement preventive measures, especially in areas with high accident rates due to fatigue.

To address this problem, recent advances in artificial intelligence have enabled the development of automated driver monitoring systems, Convolutional Neural Networks (CNNs) are advanced artificial intelligence models widely used in tasks such as image classification, segmentation, object detection and video processing. Such models are composed of several layers, such as the convolution layer, the clustering layer and the fully connected layer, which allow extracting essential features from the input data [7], these deep learning networks have gained recognition for their ability to process images with high efficiency and detect complex patterns [8], which makes them particularly suitable for drowsiness detection applications.

The objective of this research project was to evaluate and compare different CNN architectures for drowsiness detection in drivers. In particular, we sought to determine which of these architectures offered the best performance in the identification of drowsiness cues, such as the state of the eyes (open, closed or undetected) and the presence of yawning. For the yawning state, images of drivers in different states were used according to the dataset in the Kaggle database [9], and a new dataset specific to the state of the eyes was constructed, which included images in a wide variety of situations. The training was performed with 4 convolutional neural network architectures: LeNet [10], DenseNet121 [11], InceptionV3 [12], MobileNet [13]. The architectures were selected for their diversity, ranging from simple and efficient models to more advanced and specialized ones.

All research has its limits, and this study is no exception. Although the models evaluated have demonstrated high performance in drowsiness detection in stable conditions, their effectiveness can be affected by external factors, such as abrupt changes in lighting and extreme variations in the driver's posture, which can reduce their accuracy in more challenging scenarios. In addition, the availability of computational resources was a determining factor in the training process, as the use of a Google Colab Pro account was practically indispensable for reasonable processing times.

The main contributions of this study consist of the comparative evaluation of four CNN architectures to determine their strengths and limitations in drowsiness detection. In addition, two datasets have been worked with the purpose of improving generalization and robustness in various driving conditions. Finally, real-time tests were carried out to analyze the practical applicability of the models in real-world scenarios.

The remainder of this paper is organized as follows: Section II discusses related work on drowsiness detection using CNN and other artificial intelligence techniques. Section III describes the methodology, including details of the datasets, preprocessing steps, and training setup. Section IV presents the experimental results and performance comparisons. Section V discusses the findings and their implications. Finally, Section VI concludes the study and raises possible future directions for improving drowsiness detection systems.

II. Related Works

Recent studies have extensively explored the use of artificial intelligence for traffic accident prevention. For example, Ma, Chau, and Yap [14] focused on developing a fatigue detection system for nighttime driving conditions using in-depth video sequences, overcoming the limitations of RGB video-based systems in low-light environments. Their goal was to leverage Kinect sensor data to detect signs of fatigue such as yawning and posture changes. They used a Two-stream CNN architecture that includes a spatial stream to capture static features, such as driver posture, and a temporal stream to analyze changes between frames, using motion vectors instead of dense optical flow. The results of both flows were combined using an SVM classifier. The system achieved an accuracy of 91.57%, significantly outperforming systems using only RGB video in daylight environments.

Zhao et al. [15] designed a CNN-based algorithm to detect driver fatigue by analyzing the state of the eyes and mouth using the Eye and Mouth CNN (EM-CNN) network. The methodology included face and facial point detection using Multitask Cascaded Convolutional Network (MTCNN) to extract the regions of interest (ROI) of eyes and mouth. Subsequently, EM-CNN classified whether the eyes and mouth were open or closed, and the indicators percent eye closure time (PERCLOS) and percent mouth opening (POM) were calculated to assess fatigue. The results showed an accuracy of 93.623%, with a sensitivity of 93.643% and a specificity of 60.882%, demonstrating high effectiveness in detecting fatigue in a real driving environment.

Li, Gao and Suganthan [16] developed an advanced system for driver fatigue recognition using electroencephalography (EEG) signals, which reflect brain activity and exhibit high inter-subject variability, complicating cross-recognition tasks. Their goal was to improve the ability of the models to extract more distinguishable features from the decomposed EEG signals. The proposed methodology included decomposing the signals into components of different frequency bands using techniques such as discrete wavelet transform (DWT), empirical wavelet (EWT), empirical mode decomposition (EMD) and variational mode decomposition (VMD). These components were processed by independent CNNs with a Component-Specific Batch Normalization (CSBN) layer for each component to reduce inter-individual variability. The model, a hybrid ensemble of CNNs, was evaluated on cross-recognition tasks, achieving an average accuracy of 83.48%, with the DWT-based model being the most effective, outperforming existing approaches by more than 5%.

Chirra, Uyyala y Kolli [17] designed a system based on Deep CNN with the aim of detecting drowsiness in drivers based on the analysis of the state of the eyes. To achieve this, they used the Viola-Jones algorithm to detect the face and extract the eye region as the ROI. Subsequently, they applied a stacked CNN architecture including four convolution layers, each followed by normalization, ReLU activation and MaxPooling, which allowed them to extract relevant features from the images. Finally, a SoftMax layer classified the driver as drowsy or non-drowsy. The model was trained with 1200 images and evaluated with 1150, reaching an accuracy of 96.42%. This approach overcame the limitations of traditional CNN methods, which presented difficulties in pose accuracy during regression, thus demonstrating high effectiveness in accurately detecting drowsiness in drivers.

Flórez [18] designed a real-time drowsiness detection system using computer vision, using CNN to analyze visual features such as eye closure and yawning. The goal of his research was to develop an efficient model for drowsiness detection in drivers. He evaluated several CNN architectures, such as InceptionV3, VGG16 and ResNet50V2, as well as two custom models, DD-AI and DD-AI-G, adapted for implementation on an NVIDIA Jetson Nano device. The results obtained in simulated and real environments showed an accuracy of 91.48% in simulations and 86.28% in real driving conditions, with the DD-AI-G model standing out for its superior performance.

III. Methodology

A. Type and Level

This research was of an applied type, as it focused on implementing and evaluating CNN architectures for drowsiness detection in drivers. It was descriptive, since it evaluated and compared the performance of different CNN architectures without developing new theories, using quantitative metrics for its evaluation Study design.

B. Study Design

This study employed a non-experimental design, since the independent variables were not manipulated, but observed in their performance under controlled conditions. Likewise, a cross-sectional design was used, collecting and analyzing the data at a specific moment in time. A quantitative approach was used, measuring the performance of the models through metrics such as accuracy, recall and F1-score.

C. Sample

Two datasets were used for the research. The first was a public dataset from the Kaggle platform, which contains images classified in the categories of yawning ('yawning') and non-yawning ('no_yawning') drivers. The second dataset was created specifically for this research, with images classified into three main categories: eyes open ('open'), eyes closed ('closed') and images where the eyes were not detected ('no_detected').

The images collected for the second dataset include a variety of features, such as eyes with makeup, eyes of different ethnicities (Caucasian, Asian, Afro-descendant and Latin American), with lenses (cool, warm or neutral colors) and without lenses, as well as different eye colors (very dark, medium dark, warm dark, warm light, and cool light). In addition, variations such as irritated eyes, aged eyes and eyes looking away from the eye were included.

On the other hand, images classified as 'non-detected' comprise cases in which the eyes were not detected due to factors such as dark or reflective lenses, obstructions by hair, hats, caps or hands, inadequate lighting (very bright or dark), blurring (total or slight), or artistic make-up that confuses the processing.

The inclusion criteria for both datasets were images of acceptable quality for neural network processing.

The sample size includes two datasets: a public dataset from Kaggle, consisting of 2892 images distributed in 2083 for training, 519 for validation and 290 for testing; and a proprietary dataset, with a total of 4593 images, of which 3307 are for training, 825 for validation and 461 for the test set.

D. Procedure

This study started with data preprocessing, where the original dataset of the yawning state, represented in Fig. 1, was modified by duplicating the images by flipping them vertically and converting them to black and white. The conversion to black and white helped to improve performance on nighttime images by removing the color ranges present during the day and allowing CNNs to perform better predictions of both daytime and nighttime images. In addition, a green facial mesh was applied to the drivers' faces using the MediaPipe library and its FaceMesh function, which allowed the image to be cropped and focused exclusively on the facial region. These modifications were implemented after evaluating that this format offered better results in previous tests.

Fig. 1. Images of the original yawning state dataset: yawn and no_yawn.

General preprocessing was performed for the eye status dataset and the yawning status dataset as shown in Fig. 2 and Fig. 3. This process included normalizing the pixel values, scaling them from 0 to 1, and resizing the images to 224x224 pixels. In addition, the data were organized into batches of 64 and sorted. For eye status ("open", "closed" and "no_detected"), categorical classification was used, while for yawning status ("yawn" and "no_yawn"), binary classification was employed.

For eye and yawning status, approximately 70.0% of the images were separated for training, 20.0% for validation, and 10.0% of the images for testing.

Fig. 2. Images of the preprocessed dataset of the eye status: open, closed and no-detected.

Fig. 3. Images of the preprocessed yawning state dataset: yawn and no_yawn.

For the configuration of architectures, we used the optimizer Adam, using a learning rate of 0.001, and the loss function categorical_crossentropy for the eye state and binary_crossentropy for the yawning state. For evaluation, the accuracy metric was used to monitor performance in each epoch. In addition, the L2 regularization technique (kernel_regularizer=l2(0.01)) was applied to the fully connected layers and a 50% dropout layer was added to prevent over-fitting.

The modifications implemented in each architecture are illustrated in Fig. 4, Fig. 5, Fig. 6 and Fig. 7, showing the structure of LeNet, DenseNet121, InceptionV3 and MobileNet with the optimization techniques applied.

In order to optimize the training process, the Early Stopping technique was applied, which ended the training if the loss in the validation set did not improve after five consecutive epochs. And in case the training was terminated, the weights corresponding to the best performance were restored.

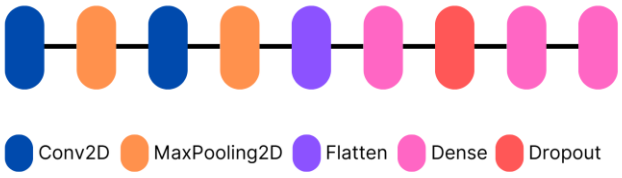


Fig. 4. Structure of the LeNet architecture with optimization techniques.

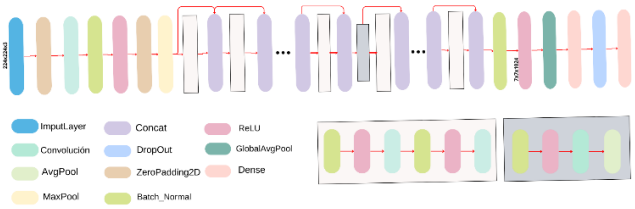


Fig. 5. Structure of the DenseNet121 architecture with optimization techniques.

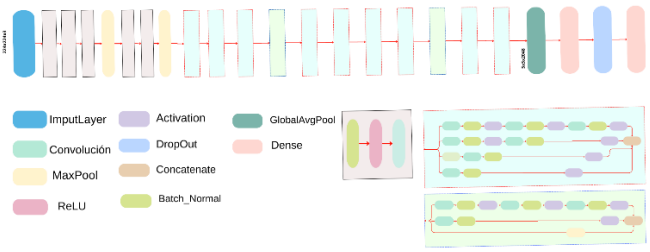


Fig. 6. Structure of the InceptionV3 architecture using optimization techniques.

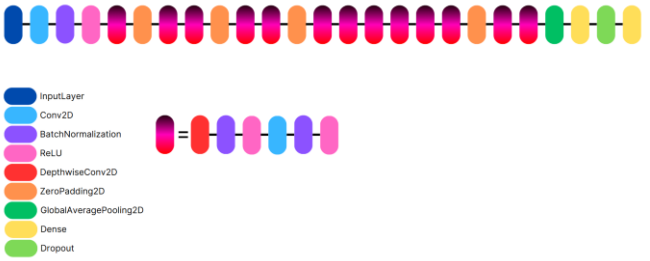


Fig. 7. Structure of the MobileNet architecture with optimization techniques.

Next, each convolutional neural network architecture was trained: LeNet, DenseNet121, InceptionV3 and MobileNet. In the case of DenseNet121, InceptionV3 and MobileNet, pre-trained models were used with the weights of ImageNet.

The training was carried out on Google Colab Pro using an A100 GPU, which allowed processing time to be optimized. Without this configuration, training the heavier architectures would have taken 4 to 5 days, requiring continuous connection. The trained models were stored in .h5 and tflite formats for future evaluation.

Subsequently, the trained models were stored in .h5 and.tflite formats to facilitate their evaluation and further use. A prototype was developed in Google Colab to perform the corresponding evaluations, which activated the camera and processed the video in real time. Each captured frame was preprocessed and sent to each trained model, allowing the driver's status to be displayed immediately.

IV. Results

This section presents the results obtained from the evaluation of the LeNet, DenseNet121, InceptionV3 and MobileNet architectures in drowsiness detection. The results are structured in the following analyses:

A. Training and Validation Curves

Fig. 8 shows the accuracy and loss curves for yawning detection. MobileNet and InceptionV3 achieved fast convergence and maintained stability during training. DenseNet121 completed its training in fewer epochs with competitive performance. In contrast, LeNet exhibited greater variability in accuracy and difficulties in generalization.

Fig. 9 illustrates the performance in eye state detection. MobileNet and DenseNet121 demonstrated rapid convergence, with a steep reduction in loss from the earliest epochs. InceptionV3, although progressively improving, showed a less pronounced decrease in loss. LeNet exhibited fluctuations in both accuracy and loss, evidencing instability in training.

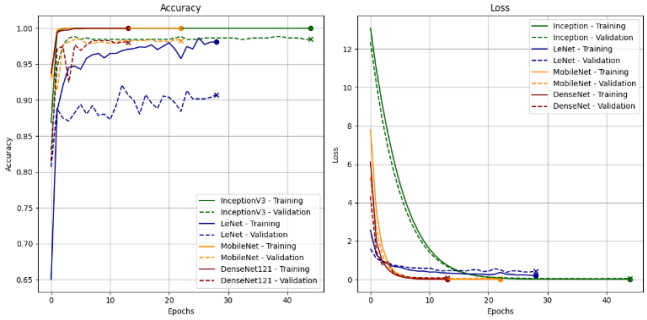


Fig. 8. Accuracy and loss curves during training for yawning state detection using LeNet, DenseNet121, InceptionV3 and MobileNet architectures.

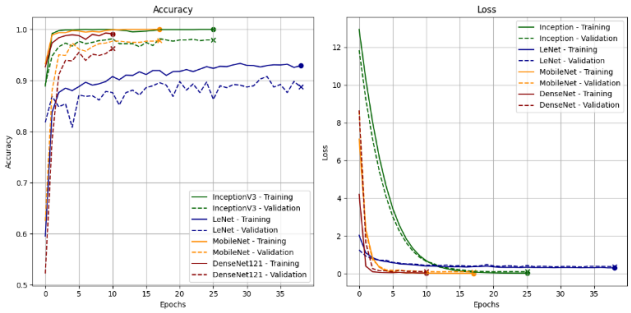


Fig. 9. Accuracy and loss curves during MobileNet, InceptionV3, DenseNet121 and LeNet training for the eye state.

B. Accuracy in Validation, Training and Testing

Fig. 10 presented the performance of the models in each phase of the training, validation and testing process. DenseNet121 and MobileNet achieved accuracies of 99.66% and 99.31%, respectively, in the testing phase for yawning detection, with minimal variations with respect to validation. In eye state detection, DenseNet121 achieved an accuracy of 96.53%, while MobileNet recorded 93.28% in the test phase.

InceptionV3 showed an accuracy of 100% in the training phase for eye state detection, with a reduction to 59.87% in the test phase. LeNet obtained an accuracy of 97.24% in yawning detection in the test phase, with lower values than those obtained by DenseNet121 and MobileNet.

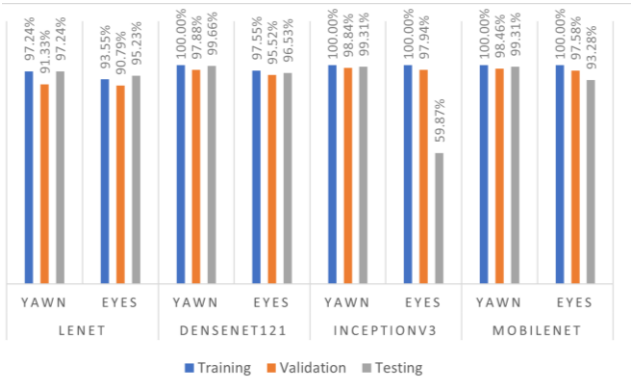


Fig. 10. Comparative performance of CNN architectures in the training, validation and testing phases for yawning and eye states.

C. Confusion Matrix

Fig. 11 shows the confusion matrices of LeNet, DenseNet121, InceptionV3 and MobileNet in the classification of 'Yawn' and 'No Yawn' states. It was observed that DenseNet121 performed the best, with high accuracy and minimal error incidence. InceptionV3 and MobileNet showed balanced performance, with similar error rates. In contrast, LeNet presented greater difficulty in detecting 'Yawn', registering more false negatives compared to the other architectures.

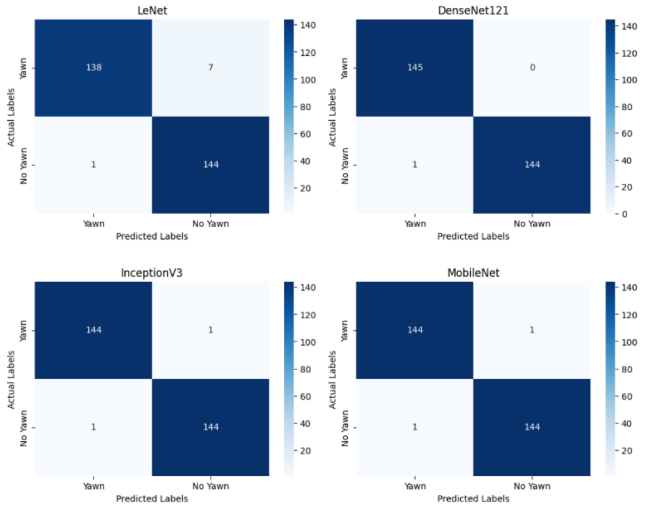


Fig. 11. Confusion matrices of LeNet, DenseNet121, InceptionV3 and MobileNet architectures in yawning state classification (Yawn and No-Yawn).

The confusion matrices of the LeNet, DenseNet121, InceptionV3 and MobileNet architectures for the classification of eye states ('Open', 'Closed' and 'No-Detected') are presented in Fig. 12.

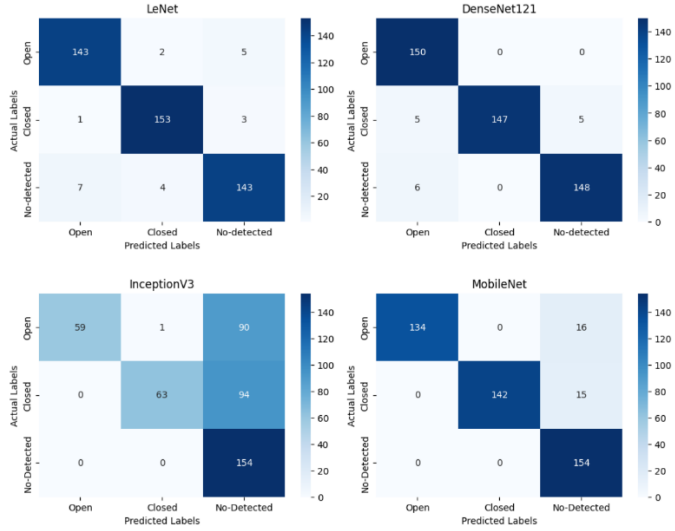


Fig. 12. Confusion matrices for the LeNet, DenseNet121, InceptionV3 and MobileNet architectures in the classification of eye states (Open, Closed and No-Detected).

DenseNet121 performed best, with high true positive values and minimal errors, with no false positives in the 'No-Detected' class. LeNet showed adequate performance, although it recorded false negatives in 'Open' and 'Closed'. MobileNet presented a balance between accuracy and sensitivity, with moderate confusions between 'Open' and 'No-Detected'. In contrast, InceptionV3 had the highest number of false negatives in 'Open' and 'Closed', reflecting a lower ability to correctly classify these categories.

D. Ranking Metrics

Table I presents the classification metrics obtained for each architecture in yawning state detection. DenseNet121 achieved an F1-Score of 1.00 in both classes, with precision of 0.99 and recall of 1.00 in "Yawn". InceptionV3 and MobileNet recorded an F1-Score of 0.99, with balanced accuracy and recall in both categories. LeNet obtained an F1-Score of 0.97, with an accuracy of 0.99 in "Yawn" and a recall of 0.95 in the same class. In the yawning state.

TABLE I. YAWNING STATE CLASSIFICATION REPORTS

Architecture	Class	Accuracy	Recall	F1-Score
LeNet	Yawn	0.99	0.95	0.97
	No Yawn	0.95	0.99	0.97
	Macro AVG	0.97	0.97	0.97
DenseNet121	Yawn	0.99	1	1
	No Yawn	1	0.99	1
	Macro AVG	1	1	1
InceptionV3	Yawn	0.99	0.99	0.99
	No Yawn	0.99	0.99	0.99
	Macro AVG	0.99	0.99	0.99
MobileNet	Yawn	0.99	0.99	0.99
	No Yawn	0.99	0.99	0.99
	Macro AVG	0.99	0.99	0.99

TABLE II. EYE CONDITION CLASSIFICATION REPORTS

Architecture	Class	Accuracy	Recall	F1-Score
LeNet	Closed	0.95	0.95	0.95
	Open	0.96	0.97	0.97
	No-detected	0.95	0.93	0.94
	Macro AVG	0.95	0.95	0.95
DenseNet121	Closed	0.93	1	0.96
	Open	1	0.94	0.97
	No-detected	0.97	0.96	0.96
	Macro AVG	0.97	0.97	0.97
InceptionV3	Closed	0.98	0.4	0.57
	Open	1	0.39	0.56
	No-detected	0.46	1	0.63
	Macro AVG	0.81	0.6	0.59
MobileNet	Closed	1	0.9	0.95
	Open	1	0.89	0.94
	No-detected	0.83	1	0.91
	Macro AVG	0.94	0.93	0.93

Table II presents the ranking metrics for each architecture. DenseNet121 obtained the best performance with an average F1-Score of 0.97. MobileNet showed a balanced performance with an F1-Score of 0.93. LeNet showed consistent values across all classes with an F1-Score of 0.95. InceptionV3 recorded the lowest performance, with a noticeable reduction in recall for "Closed" and "Open".

E. Real-Time Testing

The prototype created at Google Colab prepared each frame to meet the input requirements of the architectures, such as image size and format. Subsequently, each model generated its prediction and the level of certainty expressed as a percentage. This level of certainty indicates the model's confidence in its prediction:

- Close to 100%: High prediction confidence.
- Near 50% or less: Low confidence, suggesting doubt or ambiguity in the prediction.

The tests were carried out under real conditions, as shown in Fig. 13, where the models faced different situations such as subject movements, light variations and facial expressions.

1) Yawning state predictions:

a) LeNet: Its level of certainty varies between 56% and 100%, remaining at 100% in stable conditions (with little light variation and minimal user movements). However, in extreme scenarios, such as constant user movements or environments with low illumination, its transparency decreases significantly, leading to incorrect predictions on several occasions.

b) DenseNet121: Its accuracy level ranges between 90% and 100%, remaining at 99% in stable conditions. In extreme scenarios it maintains excellent accuracy, showing great robustness to environmental variations.

c) InceptionV3: Its accuracy fluctuates between 60% and 100%, stabilizing around 91% under normal conditions. It is the model with the greatest variability in its level of certainty, but even so it adapts well to changes in illumination and extreme conditions, offering reliable predictions.

d) MobileNet: With a range of certainty between 95% and 100%, it remains practically 99% in optimal conditions. In extreme scenarios, it shows an outstanding performance, standing out for its accuracy and high reliability.

2) Eye condition predictions:

a) LeNet: Although its performance in yawn detection was the lowest, it surprises with decent results in eye status, even in low light environments. Its performance is almost comparable to MobileNet and even becomes better in some cases, showing a very solid behavior in stable conditions. It could be considered as the second best model for this condition.

b) DenseNet121: In challenging environments, it shows greater instability, with notable fluctuations in its predictions. However, in standard conditions, it achieves a satisfactory performance, although it lags behind other more consistent models.

c) InceptionV3: It shares similar characteristics with DenseNet121 in terms of instability in difficult scenarios. Although it achieves very good results in controlled environments, its variability also places it among the least reliable models in this test.

d) MobileNet: It stands out as the most robust and consistent model for the eye condition, maintaining outstanding performance even in difficult or low-light environments. Its ability to adapt to adverse conditions clearly positions it as the best model for this condition.

Fig. 13. Images of the prototype executed in real time.

V. DISCUSSIONS

The results obtained in this study showed that the DenseNet121 and MobileNet architectures significantly outperformed the models used in previous research, achieving accuracies of over 99% in yawn detection and 93% in eye detection.

Unlike the model of Zhao et al [15], which used EM-CNN and MTC-CNN for face detection, this evaluation identified MobileNet and DenseNet as the best performing architectures. MobileNet achieved 99.31% test accuracy for yawn detection and 93.28% for eye classification, while DenseNet obtained 99.66% and 96.53%, respectively. These results exceed the 93.62% reported in their study, evidencing the stability of MobileNet and DenseNet in diverse scenarios. Furthermore, unlike PERCLOS and POM-based approaches, the landmark-based classification enabled a more accurate segmentation of facial regions, optimizing the detection of fatigue-relevant facial states.

The study by Chirra, Uyyala, and Kolli [17] used a stacked CNN together with the Viola-Jones algorithm, obtaining an accuracy of 96.42% in eye state detection. In the present work,

DenseNet121 achieved an accuracy of 96.53% without the need for additional face detection algorithms, demonstrating its ability to extract relevant features efficiently. The improved performance can be attributed to the preprocessing techniques applied, such as normalization, grayscale conversion and face segmentation with MediaPipe, in addition to the implementation of L2 regularization and dropout, which contributed to reduce overfitting and improve the generalization capability of the model.

Also, unlike the study by Ma, Chau, and Yap [14], which employed a Two-stream CNN with Kinect sensors to improve fatigue detection in nighttime conditions, this study evaluated real-time CNN architectures without the need for additional sensors. While the video stream-based approach achieved 91.57% accuracy, MobileNet and DenseNet121 achieved up to 99% in yawning and eye state detection, while maintaining stability in the face of illumination and motion variations. These results suggest that optimized models can be an efficient alternative for drowsiness detection without requiring specialized hardware.

On the other hand, in the study by Li, Gao and Suganthan [16] they achieved an accuracy of 83.48% when combining CNN with EEG signals, a methodology that, although useful, is more complex and less accurate than the visual feature analysis performed in this study. The results obtained with LeNet, DenseNet121 and MobileNet in both states evaluated reflect that facial image-based techniques are more accurate and practical for vehicular implementations.

Finally, the study by Florez [18] used CNN in a real-time system with InceptionV3, VGG16, ResNet50V2, obtaining an accuracy of 91.48% in simulations and 86.28% in real driving. In this study, InceptionV3 showed 99.31% in test, but its real-time certainty ranged from 60% to 100%, with a drop to 59.87% in eye-state detection. MobileNet and DenseNet121 were more stable, with 99% real-time certainty. This confirms that, although the models achieve high accuracy in controlled tests, their performance in real environments can be affected, with MobileNet standing out as the most robust.

VI. Conclusions and Recommendations

In this study, different convolutional neural network architectures were evaluated and compared for drowsiness detection in drivers, focusing on two key aspects: eye state and yawning state. For yawning state, the architectures that topped the list in accuracy upon training were InceptionV3 with a validation accuracy of 98.84%, followed by MobileNet with 98.46%, DenseNet121 with 97.88% and finally LeNet with 91.33%. For eye status, InceptionV3 stood out with a validation accuracy of 97.94%, followed by MobileNet with 97.58%, then InceptionV3 with 95.52%, and finally LeNet with 90.79%.

The results of the confusion matrices and the classification report clearly reflect this superior performance. LeNet, DenseNet121, InceptionV3 and MobileNet achieved outstanding classifications in the yawning and eyes state having very few errors with the exception of InceptionV3 which showed less consistent performance in the eyes state, especially in the 'Open' class, where a significant number of errors were observed.

Finally, in real-time testing with the prototype, MobileNet proved to be the most robust and reliable architecture for both yawning and eye detection. Its outstanding accuracy remained consistent even under challenging conditions. LeNet, although it came in last place in yawning state detection, surprised by showing solid performance in eye state detection, obtaining comparable or even better results to MobileNet in some scenarios. DenseNet121 showed solid and consistent performance in yawning state detection, positioning itself as a reliable alternative to MobileNet. However, its performance in eye state detection was more unstable in challenging environments. For its part, InceptionV3, while achieving acceptable results in stable conditions, presented the greatest variability between architectures in both tasks. Its performance was less consistent in challenging environments, especially in eye-state detection, where it showed a higher number of errors compared to the other models.

In conclusion, when comparing the architectures, MobileNet stood out as the best choice for its consistency and accuracy, even in challenging conditions such as low illumination or constant user movements. Although DenseNet121 and InceptionV3 also offered good performance in stable environments, MobileNet stood out for its adaptability and comprehensive performance. On the other hand, LeNet, although the simplest architecture in this study, showed surprisingly good performance in eye detection, despite its lack of optimization compared to more modern architectures. It is important to note that the architectures were not modified, as the goal was to evaluate them as they are, respecting their internal structure. Although it is possible to improve the architectures by adding layers, in this study we chose to add the same layers at the beginning and at the end of all the architectures, maintaining their original shape and respecting the design with which they were built.

For future research, the integration of other types of data, such as heart rate or electromyographic activity, could be explored to improve accuracy and robustness in the detection of physiological states. It would be interesting to develop a prototype that combines these new instruments with the data obtained in this study, offering a more complete analysis. This approach could open new possibilities for more personalized and effective health monitoring systems.

Finally, based on the evaluations, an application called 'Drowse Alert' was developed using the MobileNet model, which obtained the best performance. The app is currently in closed testing on Google Play, available only to selected users. The link to access the application is as follows: https://play.google.com/store/apps/details?id=com.invoryan.drowse, as shown in Fig. 14.

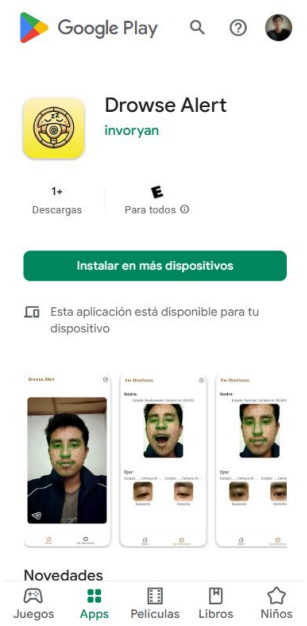


Fig. 14. Screenshot of the "Drowse Alert" application in Google Play.

REFERENCES

[1] Y. Albadawi, M. Takruri, and M. Awad, "A review of recent developments in driver drowsiness detection systems," Sensors, vol. 22, no. 5, p. 2069, 2022.

[2] K. Peña Prado, "Somnolencia en conductores de transporte público regular de pasajeros de Lima Metropolitana – Perú. 2016," Master's thesis, Universidad Peruana Cayetano Heredia, Lima, Peru, 2017.

[3] World Health Organization, "Road traffic injuries," Who.int, Dec. 13, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. [Accessed: Sep. 23, 2024].

[4] B. C. Tefft, "Drowsy driving in fatal crashes, United States, 2017–2021 (Research Brief)," AAA Foundation for Traffic Safety, Washington, D.C., 2024.

[5] Dirección General de Tráfico, "Conducir con sueño o cansancio," Www.dgt.es, 2024. [Online]. Available: https://www.dgt.es/muevete-con-seguridad/evita-conductas-de-riesgo/Conducir-con-sueno-o-cansancio. [Accessed: Sep. 24, 2024].

[6] Instituto Nacional de Estadística e Informática, "Análisis de los accidentes de tránsito ocurridos en el año 2016," Plataforma del Estado Peruano, 2017. [Online]. Available: https://www.inei.gob.pe/Est/Lib1528/cap03. [Accessed: Sep. 24, 2024].

[7] P. Purwono et al., "Understanding of convolutional neural network (CNN): A review," Int. J. Robotics Control Syst., vol. 2, no. 4, pp. 739–748, 2022.

[8] A. Ghosh, A. Sufian, F. Sultana, A. Chakrabarti, and D. De, "Fundamental concepts of convolutional neural network," in Recent Trends and Advances in Artificial Intelligence and Internet of Things, V. Balas, R. Kumar, and R. Srivastava, Eds., Intelligent Systems Reference Library. Springer, 2020. [Online]. Available: https://doi.org/10.1007/978-3-030-32644-9_36.

[9] D. Perumandla, "Drowsiness dataset," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/dheerajperumandla/drowsiness-dataset. [Accessed: Sep. 20, 2024].

[10] M. V. Vijaya Saradhi, P. Venkateswara Rao, V. Gokula Krishnan, K. Sathyamoorthy, and V. Vijayaraja, "Prediction of Alzheimer's disease using LeNet-CNN model with optimal adaptive bilateral filtering," Int. J. Commun. Netw. Inf. Secur., vol. 15, no. 1, pp. 52–58, 2023.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, 2017, pp. 4700–4708.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 2818–2826.

[13] A. G. Howard, "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017. [Online]. Available: https://arxiv.org/abs/1704.04861.

[14] X. Ma, L.-P. Chau, and K.-H. Yap, "Depth video-based two-stream convolutional neural networks for driver fatigue detection," in 2017 International Conference on Orange Technologies (ICOT), Singapore, pp. 155–158, 2017. [Online]. Available: https://doi.org/10.1109/ICOT.2017.8336111.

[15] Z. Zhao, N. Zhou, L. Zhang, H. Yan, Y. Xu, and Z. Zhang, "Driver fatigue detection based on convolutional neural networks using EM-CNN," Computational Intelligence and Neuroscience, vol. 2020, Art. no. 7251280, 11 pages, 2020. [Online]. Available: https://doi.org/10.1155/2020/7251280.

[16] R. Li, R. Gao, and P. N. Suganthan, "A decomposition-based hybrid ensemble CNN framework for driver fatigue recognition," Information Sciences, vol. 624, pp. 833–848, 2023. [Online]. Available: https://doi.org/10.1016/j.ins.2022.12.088.

[17] V. R. R. Chirra, S. R. Uyyala, and V. K. K. Kolli, "Deep CNN: A machine learning approach for driver drowsiness detection based on eye state," Revue d'Intelligence Artificielle, vol. 33, no. 6, pp. 461–466, Dec. 2019. [Online]. Available: https://doi.org/10.18280/ria.330609.

[18] R. D. Florez Zela, "Diseño e implementación de un sistema detector de somnolencia en tiempo real mediante visión computacional usando redes neuronales convolucionales aplicado a conductores," Undergraduate Thesis, Universidad Nacional de San Antonio Abad del Cusco, 2024. [Online]. Available: http://hdl.handle.net/20.500.12918/8298.

Integrating Deep Learning in Art and Design: Computational Techniques for Enhancing Creative Expression

Yanjie Deng[1], Qibing Zhai[2]*
School of Art, Sichuan Top IT Vocational Institute, ChengDu, SiChuan, 611743, China[1]
School of Art and Design, XiHua University, ChendDu, SiChuan, 610039, China[2]

***Abstract*—Deep learning and art design are being integrated, which is an innovative process that has the potential to reframe the way the human imagination is defined. This paper is an exploration of a broad field that showcases how AI enhances the experience of artist practice, especially content deep learning. This study comprises an exhaustive analysis of the cutting-edge models including generative adversarial networks (GANs), neural style transfer, and multimodal AI that assist in the creation, modification, and optimization of the artistic experience. This research points to implementations of those in the visual arts, graphic design, and interactive media while providing contemporary examples where deep learning has been an addition to traditional media and created new forms of art. Besides, the paper points to the challenges and ethical considerations concerning algorithmic art, including issues of authorship, biases, and intellectual property. The integration of computational methods in the realm of artistic expression is made in the paper and the paper provides insights into the change that deep learning can affect for artists, designers, and technologists.**

***Keywords*—Deep learning; art; design; creative expression; computational techniques**

I. INTRODUCTION

Technology and creativity have always been allied with the help of innovation, which enables artists and designers to explore new avenues of expression. The application of deep learning, a subfield of artificial intelligence (AI), is undoubtedly one of the most tremendous innovations of recent years that has invaded the arts and design [1]. Through its ability to let machines learn patterns, copy styles, and provide original content, the deep learning process has been a game-changer that has altered the approach to the conceptualization, production, and experience of creative works. This fusion of computing and artistry is not only a means but rather a new branch that challenges the conventional ideas of creativity and craftsmanship [2-3].

The functioning of deep learning algorithms is based on the processes of artificial neural networks mimicking the operations of the human brain, which processes vast amounts of data to recognize patterns and make decisions. This ability has found new depth in areas such as computer vision, natural language processing, and generative modeling. When used for art and design, these systems can do a wide variety of things, including generating realistic images, changing artistic styles, or speeding up the intricate design process. Today, artists and designers have tools that can supplement their creative efforts, giving them both efficiency and entirely new creative avenues [4-5].

One of the first breakthroughs in the field of AI in art was the neural style transfer, a technology that could bring together two separate images through the style of the first one. The demonstration of the power of AI to form an artistic composition through the achievements that followed its introduction marked the beginning of a series of advancements [6-7]. The introduction of generative adversarial networks (GANs) by Ian Goodfellow in 2014, set the stage for the development of the field that was characterized by the ability of computers to generate photorealistic images, merge those with new artworks, or even adopt the styles of famous artists. AI-driven tools exist beyond merely the generation of content; they also facilitate the examination of questions relating to abstraction, surrealism, and multimedia interactivity which are unusual as yet uncharted territories [8-9].

These new technologies create significant changes. For long-time painters, deep learning is an example of the merging of traditional and computational techniques which finally leads to hybrid outcomes that neither of the methods alone can produce [10]. Those in the graphic design world can tap into the upper hand of AI to make machines complete mundane tasks, perfect formatting patterns, and find design options in record time. Interactive media artists and game developers harness AI's potent force of synthesizing rich environments, non-linear plots, and personalized user experiences. The array of opportunities is just astonishing, only limited by creators' inventiveness and the prowess of foundational algorithms [11-12].

Notwithstanding the upsides, making a connection between deep learning and art & design in contemporary times is not a walk in the park. Modifications are being made to the definition of authorship and ownership as machines are used for a tremendous part of the creative process [13-14]. What is the rightful owner of an AI-created artwork? Can a machine be acknowledged as an artist? They are philosophical questions at the root level that lead the inquiry into the very essence of creativity and intellectual property in the time of AI. Moreover, AI systems being biased from the data supplied to them might lead to one-sided representation and a real threat to the principle of inclusiveness. The usage of AI in art and design

could give rise to a certain fear regarding human creativity, as entrenched automated systems often catch more credit for art than actual human inspiration and experience [15-17].

However, even with these caveats, the combined use of deep learning and creativity is undoubtedly on the rise. Major corporations, research institutions, and even individual artists are testing the limits of AI to prove that creativity can work in partnership with technology. The very nature of platforms such as RunwayML, Google's Magenta, and NVIDIA's GauGAN is that they give creators at all skill levels the technology they need to be able to use the adventure of the AI art world. These tools serve as a bridge between computational complexity and creative simplicity, allowing artists and designers to channel their energies into the creation of art and design while harnessing the capabilities of the machine [18-19].

Deep learning, in contrast to mere aesthetics, has far-reaching implications in the area of art and design. By providing visual and interactive instruments that allow students and professionals to understand intricate ideas, it plays a significant role in education. In the case of architecture, for example, AI aids the design of energy-efficient buildings by analyzing data from the environment and optimizing space usage [20]. Forecasting fashion trends, adapting creative designs to individual preferences, and improving production operations are some of the ways deep learning shapes the fashion industry. The multifaceted usage of social media using deep learning signifies its vigor as a catalyst for creativity in a variety of industries [21].

A. Objectives

- To study how deep learning has changed the way creativity and artistic workflow interact: On the way, one sees how the use of tools like GAN and neural style transfer naturally leads to more innovation and faster work in the areas of art and design.
- To analyze the fundamental principles and ethical aspects: The issues of combined identity, ideas creation, and unfairness in AI art are the topics to be considered here.
- To point out new prospects and trends in the industry: It means to share ideas about the already existing and future technologies that may affect the integration of AI in creative professions.

To sum up, deep learning is not just a choice in technology, but a life-changing force in the creation of art and design that brings human and machine interaction in a new way. A creative revolution is taking place now, with deep learning leading the way through the provision of tools that are creative, tackle problems, and stimulate new forms of expression [22]. This research will uncover the many ways in which deep learning fosters, confronts, and rearranges the artistic scene thus providing an insight into the creative process in a digital future [23-24]. With this analysis, it aims at promoting a better comprehension of the opportunities, and barriers of AI as a helper in artistic and design processes.

The rest of this paper is arranged as follows: Section II (Literature Review) recaps previous research on AI in art and design, pointing out major approaches, contributions, and drawbacks. Section III (Methodology) elaborates on the implemented deep learning techniques, the data preprocessing steps, the design of the model, and the evaluation metrics. Section IV (Results) reports the study's empirical outcomes, and Section V (Discussion) then weighs the significance of these results against traditional artistic practices and also compares AI-driven approaches with them. Section VI (Conclusion and Future Work) majorly encapsulates the study's significant findings, furthers the implications of said findings for future research, highlights limitations, and provides directions for future research.

II. Literature Review

The intersection of deep learning with art and design has attracted significant attention in the last years resulting in the emergence of a growing body of literature that emphasizes the transformative potential of these technologies. Diverse aspects of this field have been explored by researchers from varying technological foundations of deep learning algorithms to their practical applications in creative workflows [25-26]. The gist of this section is given via a deep insight into the significant studies that have influenced our awareness of the part of deep learning in art and design thereby the methodologies they have adopted and their findings. The findings of each study point to the various aspects of this domain, demonstrating how computational techniques facilitate, challenge, and change conditional creativity.

The literature review has been carefully restructured into three main subsections. The first subsection titled AI Model Development elaborately covers the fundamental research of neural networks applied in art. The model architectures and the training methodologies are described in detail. The second subsection titled Applications in Art and Design surveys recent studies on the implications of deep learning for creative workflows, including its deployment in digital painting, graphic design, and fashion items. The last subsection entitled Ethical Considerations studies the challenges of authorship, bias, and the impact of AI on human artistic agency. Table I shows significant contributions by the related research. The detailed systematic comparison of processes, data, measures, and issues of previous studies' results in this table helps clarify the point that the AI-based art is contrasted with the conventional ones. The inclusion of this comparative analysis makes the literature review an all-inclusive review covering the latest research in AI creativity.

McLain [27] examines the concept of signature pedagogies in design and technology education. They do this by using Shulman's framework. They use the concept of distinctive teaching methods which in this case are called ideators, realizes, and critics. The terms designer, maker, and evaluator are synonymous with them. Their study underscores the role of project-based learning, collaboration, creativity, and problem-solving in shaping these pedagogies, and proposes the framework which highlights the unique contributions of design education to the curriculum.

TABLE I. LITERATURE COMPARISON

Author(s)	Focus Area	Key Methods/Approach	Findings/Contributions	Implications
McLain et al.	Signature pedagogies in design and technology education	Literature review on Shulman's framework; project-based learning emphasizing ideating, realizing, and critiquing	Identified unique contributions of design education to curriculum through collaboration, creativity, and problem-solving	Offers a framework for enhancing teaching practices in design education.
Jin et al.	Metaverse in art design education	Virtual learning through Xirang games; qualitative and quantitative analyses	Demonstrated increased creativity and critical thinking via virtual presence and collaborative learning strategies	Provides a model for integrating metaverse technologies into future educational practices.
Noble et al.	Art education and teacher training	Practitioner-led action research; constructivist pedagogy	Enhanced visual literacy and creativity through community practice involving teachers, artists, and museum professionals	Suggests collaborative approaches for improving art education outcomes.
Saleeb et al.	Virtual learning in practical education	2D/3D virtual environments; learning analytics	Validated the effectiveness of virtual tools for achieving hands-on learning outcomes	Supports the shift to online delivery for traditionally face-to-face programs.
Li et al.	AI in conceptual product design	Systematic review of cross-modal tasks (DLCMT)	Identified challenges like multimodal data and creativity demands; proposed solutions and future research directions	Highlights AI's potential to redefine the conceptual design process.
Zhao et al.	Emotional aspects of AI in art and design	Review of AI technologies and human collaboration in art	Emphasized the collaborative potential between AI and human designers for enhancing aesthetic resonance and creativity	Proposes a future where AI complements rather than replaces human creativity.
Qiu et al.	AI in art design and souvenir creation	Application of AI and deep learning for design automation	Demonstrated AI's ability to enhance creativity and meet market trends in visual arts	Encourages using AI for practical applications in art and design.
Cai et al.	Art education system with CAD and deep learning	CAD-integrated instructional system; evaluation mechanisms	Improved learning outcomes, skill mastery, and creativity in students	Advocates for using AI-powered systems in teaching to expand creative potential.
German et al.	AI as a co-creative partner in design	AI algorithms trained on small datasets; evaluation of design variations	Showed how AI can inspire creativity and novel designs; addressed ethical concerns	Supports the integration of AI as a tool for augmenting human creativity while recognizing ethical considerations.

Jin et al. [28] investigate the use of the metaverse in art design education during the pandemic focusing on the activities of Xirang games through which innovative teaching strategies were introduced. The limitations of traditional and the gap between the goals that were proposed to be achieved using these new technologies were the central challenges that they were addressing. They proposed the solutions for them as virtual presence and collaborative learning groups. With qualitative and quantitative methods, they tested and illustrated the metaverse's role in the empowerment of creativity, critical thinking, and engagement demonstrated in a new way for the future educational models.

Noble [29] investigates art education through practitioner-led action research and museum and artist-led CPD programs. The results of the study deliver a message about how with constructivist pedagogy one can enable participatory, and experimental approaches such as thus being part of art education not only will the participants achieve an increase in visual literacy and creativity, but also will teachers help students to be more creative. The essential part of this research is the construction of communities of practice which are formed by teachers, artists, and museum professionals. The research revealed the great possibility of the transformation of art and design teaching.

Saleeb [30] engages in questioning the necessity of face-to-face learning in practical fields like engineering and design, by virtual learning environments to achieve learning outcomes and foster creative thinking using two. They establish the working of virtual learning environments utilizing such 2D and 3D media and assuring the effective creation of application and creativity. Learning analytics are used as validating tools to show the ability of the aforementioned to drive learning performance, the acquisition of skills and an argument for education such as that of practical artware in a remote way.

Li et al. [31] support by considering text on product conceptual design requiring balancing of products in the cross-modal. Their systematic review included a variety of methodologies including text-to-3D and sketch-to-3D transformations, it particularly drew attention to the fact that current knowledge is lacking on the subject matter and set directions for further research. They stress how AI will transform the world of design by tackling these challenges.

Zhao et al. [32] study the potential of AI in art and design, notably its ability to mimic emotional articulation and produce twofold effects such as conjuring up emotions and presenting beauty through some. In addition, they recognize AI's limitations, but they are more concerned about the cooperative aspect between human beings and AI technology. The authors predict a scenario in the future where AI will play a part in human creative processes, thus a creative cooperation between man and machine in the sphere of art can be expected.

Qiu et al. [33] draw attention to the necessity of AI and deep learning in creative arts and custom souvenir making; they automate those tasks such as 'image recognition' and 'market analysis' totally dependent upon human beings. Art is facilitated by AI that aligns itself closely with trends about consumers, their psychology. Thus, the research puts emphasis on the role of AI in both sides - creativity, and practical issues purposely to provide evidence of its increased influence on visual arts.

Cai et al. [34] have gone ahead and proposed an art educational system that includes the use of CAD and deep-learning models. The user's system is the creative stylist who takes the creative direction and integrates various products from the teaching materials. The results show an improvement in the learning outcomes and creativity that the students can exhibit, and this exemplifies how artificial intelligence and technology can completely change art education and can shape the creative talents of the young.

The part that Germany et al. [35] dealt with is the one where they took actors to examine AIs as co-creative partners in design where algorithms help come up with various types of new schemes. Their study is a clear example of how AI could improve creativity without the need for human input, and as a result, it leads to unexpected solutions. The research also deals with the ethical concerns that AI would have on the creative industries and provides a more holistic view of its benefits and drawbacks.

III. METHODOLOGY

This study adopts a comprehensive methodology to explore the integration of deep learning in art and design, focusing on its applications, challenges, and transformative potential. The approach is interdisciplinary, which combines application specific computation experimentation, creative data analysis, and evaluation of aesthetic results. It includes stages such as extracting essential data, preprocessing the data, creating a model, applying ability and testing the model. The methodology of this research is aimed to verify to what extent deep learning technologies can be leveraged by digital artists whilst also answering the concerns of accuracy, authorship, and the ability to scale.

A. Data Collection and Preprocessing

The basis for any successful deep learning project is the data sourced from various places and available in different styles. This study utilized online-source images by which the researchers intended to gather both diverse kinds of data and numerous examples of the same types of art. Such a collection would, therefore, include pictures of publicly owned artworks, user-designed projects of various types, and special manufacturing facilities as well as color-variant designs to increase diversity. The data compared different forms such as painting, digital illustration, photography, and graphic design to ensure that all kinds of art were analyzed.

Data preprocessing included various steps such as image normalization, augmentation (including random cropping, rotation, and color jittering), and edge detection for improved feature extraction. The dataset was divided into training (70%), validation (15%), and test (15%) sets. Each model was fine-tuned iteratively, dropout rates of between 0.3 and 0.5 were used to prevent overfitting. Evaluation was done using the conventional deep learning benchmarks, like mean square error (MSE) for GAN outputs and precision-recall curves for classification tasks.

The critical preprocessing step was the one that set the stage for the direct use of the data by deep learning models and involved the augmentation of data by a series of cropping, scaling, rotation, and color adjustments among others to increase its diversity. Furthermore, the quality of the data was improved by using techniques such as image enhancement and feature extraction-the first one highlights defects, and transforms the image into a standard pattern, while the second one searches for key features in each image and forms the basis for training the models. The feature extraction procedures helped to identify the desired attributes in all the images, such as shapes, textures, and color palettes, which were then used for model training.

B. Model Development

The following stage involved the careful selection and configuration of deep learning architectures that were best suited to specific creative endeavors. The study also took advantage of more advanced designs such as Generative Adversarial Networks (GANs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs). Models were chosen based on their appropriateness for specific tasks. For example, high-quality visual art was created using GANs, while CNNs were employed for style transfer and object recognition. RNNs were used to reveal the temporal component relevant to the particular creative activity, thus making diverse media such as animations and interactive systems.

The models were trained using the pre-processed datasets in which the hyperparameters were optimally set for the best possible performance. In this respect, high-tech techniques such as transfer learning and fine-tuning were used to boost the re-training of the pre-trained models plus for improving their precision. The training cycle was completed by evaluation steps that resulted in the requests for the modification of the models so that they were not only logically precise but also relevant to the contextual ones.

C. Creative Tool Implementation

Once the models were trained, they were packaged with user-friendly design tools intended for artists and designers. These applications included features such as style transfer, form and composition recommendations, and color palette optimizations. For example, the style transfer module enabled users to make use of the characteristics of one artwork to another to make a picture with characteristic components of two artworks or one artwork and other subject matter. Similarly, the composition tool gave recommendations on how to distribute elements visually in a balanced and harmonious way, based on the design principles that have been mastered.

In order to examine both usability and effectiveness, the tools were submitted to user testing. Artists and designers gave feedback on how these tools contributed to their creative workflows. This feedback approach led to continual modification of the tools in order to match the users' wants and needs.

D. Performance Evaluation

To assess the models and tools' performance, a set of metrics was defined, including their visual appeal, artistic creativity scores, and user engagement. Visual allure was determined through both automated evaluation metrics and the subjective feedback of evaluators. The factor of originality and novelty of the outputs was considered for measuring artistic

creativity. Whereas, the user engagement was captured through the time and frequency of tool usage. These evaluations have given valuable insights into the effectiveness of the solutions and their potential for being broadly adopted.

The provisioned system unifies deep learning algorithms in a systematic channel. The structure starts from data gathering where the data is taken from artistic styles, design concepts, and external influences processing. The raw data is upgraded through enhancements, augmentations, and feature extractions by preprocessing. The use of a multi-model deep learning module is noted: CNNs for deep learning of feature-based classification, GAIs-based generation of artistic content, AIs with temporal artistic applications (RNNs), and creativity's abstraction. As performance standards like creativity and user engagement will be taken into account while evaluating the final outputs. The whole structure is depicted in Fig. 1, thereby showing the interactive feedback loop for continuous model optimization.

E. Working of the Proposed Model

The new model as depicted in "Fig. 1" consists of multiple interacting components taking the shape of a smooth workflow for integrating deep learning into art and design. The project kicks off with the input ready, the data, including design concepts, art style data, artist preferences, and factors outside the artist's control, that will be used. In the data preprocessing stage, the input data is formed, together with, and the following techniques are engaged, such as image enhancement, feature extraction, and data augmentation, all these are part of the data preparation phase which will be then analyzed further.

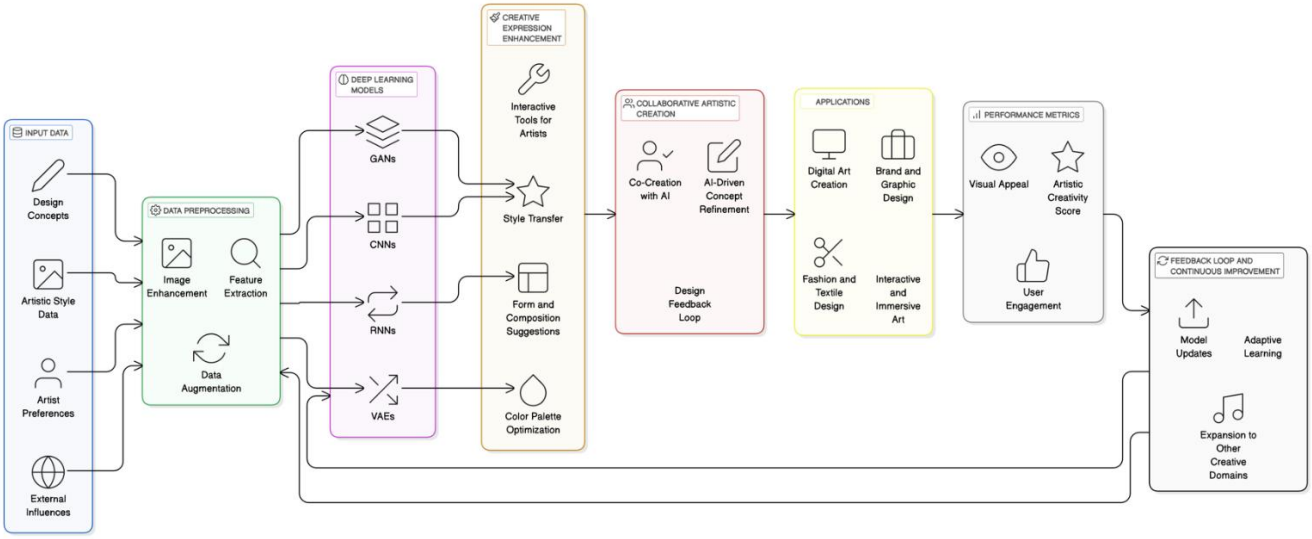


Fig. 1. Proposed model diagram.

The preprocessed data is then passed to the deep learning model made up of various advanced architectures, such as GANs, CNNs, RNNs, and VAEs. Each of these models does a different task, such as creating new artistic content, changing styles, and giving form and composition tips. These tools generate some of the model outputs that can help creative people express themselves better by providing them with interactive tools, suggesting the best color combinations, and going through the idea drafting with AI and the artist together, the outcome being a creation that normally could not have been done.

The proposed model's applications are different including digital art production, brand and graphic design, fashion and textile design, and immersive art. The model's performance is evaluated based on metrics such as visual appeal, artistic creativity scores, and user engagement. A feedback loop guarantees continuous improvement, with model updates and adaptive learning extending the system's capabilities to other creative fields.

The primary architectures under deep learning used in this study are the Generative Adversarial Networks (GANs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Variational Autoencoders (VAEs). GAN has been constructed by having a generator and discriminator with convolutional layers, batch normalization, and the LeakyReLU activation. The CNN model was the ResNet-based architecture that was pre-trained on ImageNet for transfer learning, with fine-tuning to enhance and classify the artwork's style, using additional layers. The neural architecture for the RNN was made up of Long Short-Term Memory (LSTM) units so that the dynamic and artistic sequences could be mapped in a chronological sequence. For the purpose of training VAEs, a framework of inference in terms of variables was adapted by the latent space's dimension being set to 128. GANs and VAEs were obtained by means of the Adam optimizer with the learning rate set at 0.0002, and the rate for CAT and RNN was equal to 0.001 as well. In addition, the models were trained on the NVIDIA RTX 3090 GPU for 50 epochs with a batch size of 32.

In conclusion, the methodology being proposed is an integration of deep learning, which is mechanical in nature, with soft aspects of art evaluation. The fusion of both creative and scientific disciplines results in a robust grounding for AI's role in art and design. The model presented in Fig. 1 exemplifies this interdisciplinary cooperation and poses a guideline for future research in this intriguing field.

IV. RESULTS

The proposed deep learning models were evaluated based on their ability to perform artistic tasks. The Art Images Dataset, which is a dataset comprising various art forms including paintings, drawings, engravings, and sculptures sourced from Kaggle, was used. This diversity of datasets was a solid foundation for the model performance evaluation in various domains of creativity. The results are summarized in terms of accuracy metrics and creativity scores, and visual representations demonstrating the competitive performance of different models are also included.

A. Model Accuracy

The models were analyzed through measuring the precision of their created, transformed, or improved artistic works, and also, their precision of stylistic and compositional faithful. In the graphical representation "Fig. 2", it was evident that CNNs achieved superior performance in accuracy across all categories compared to other models, with an average accuracy of 90% and 92% for paintings and drawings respectively. This can be attributed to the ability of the CNN in the automatic feature extraction of high-resolution artistic images.

GANs were the next best-performing models, having averaged an accuracy of 85%. The advantage lies in the creation of original and astonishing artistic work that they can attain in tasks exceedingly well. The VAEs are also capable of performing great, especially in quite more abstract and surreal compositions; more so in painting tasks with an accuracy of 88%. RNNs might not be quite as accurate they did perform quite well in the paradigms where dynamic sequences were demanded such as animations and sequential art. "Table II shows model accuracy metrics.

B. Creativity Scores

In this study, the measure of creativity used a grading method in which a score of 1 manifest the lowest level of creativity and a score of 5 manifests the highest level of creativity in the output. Furthermore, the type of art input (drawings and paintings) is also a significant factor that pushed the CNNs to achieve the highest scores (the results are shown in "Fig. 3". As a result, the model's abilities in picking up the fine details and the flexibility in its style options could be highlighted.

TABLE II. MODEL ACCURACY METRICS

Art Category	GAN Accuracy (%)	CNN Accuracy (%)	RNN Accuracy (%)	VAE Accuracy (%)
Painting	85	90	75	88
Sculpture	80	84	70	82
Engraving	78	83	68	80
Drawing	88	92	73	85

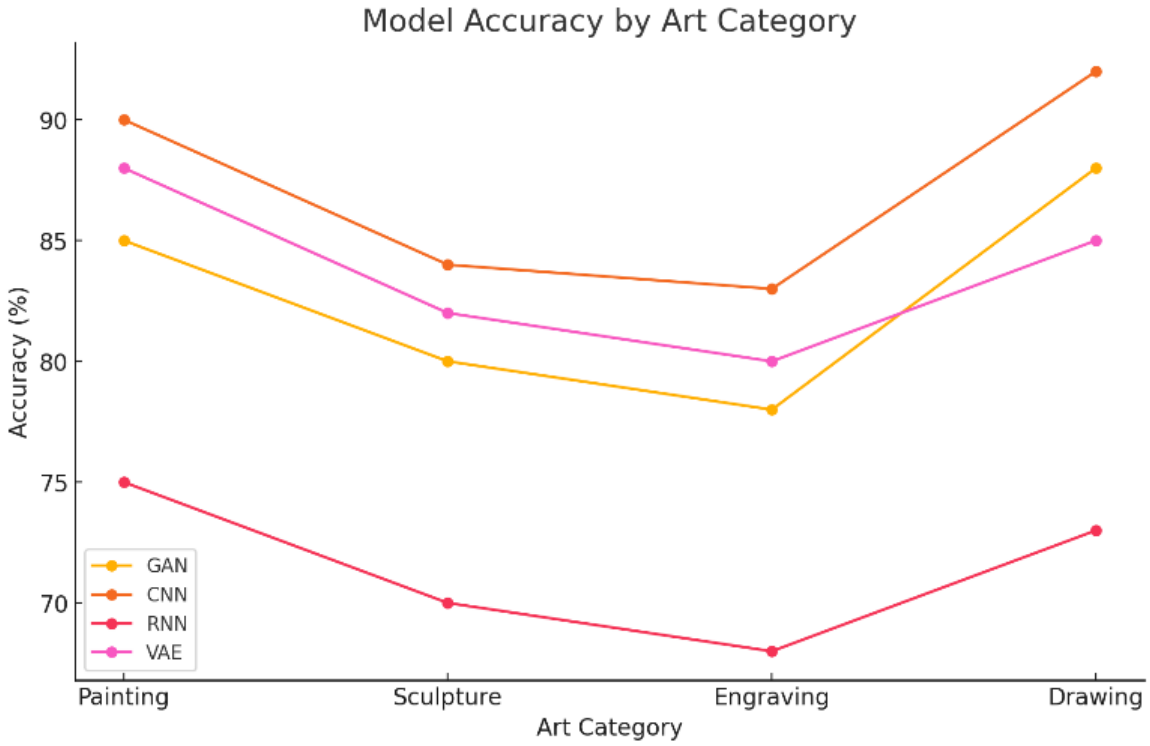


Fig. 2. Model accuracy by art category.

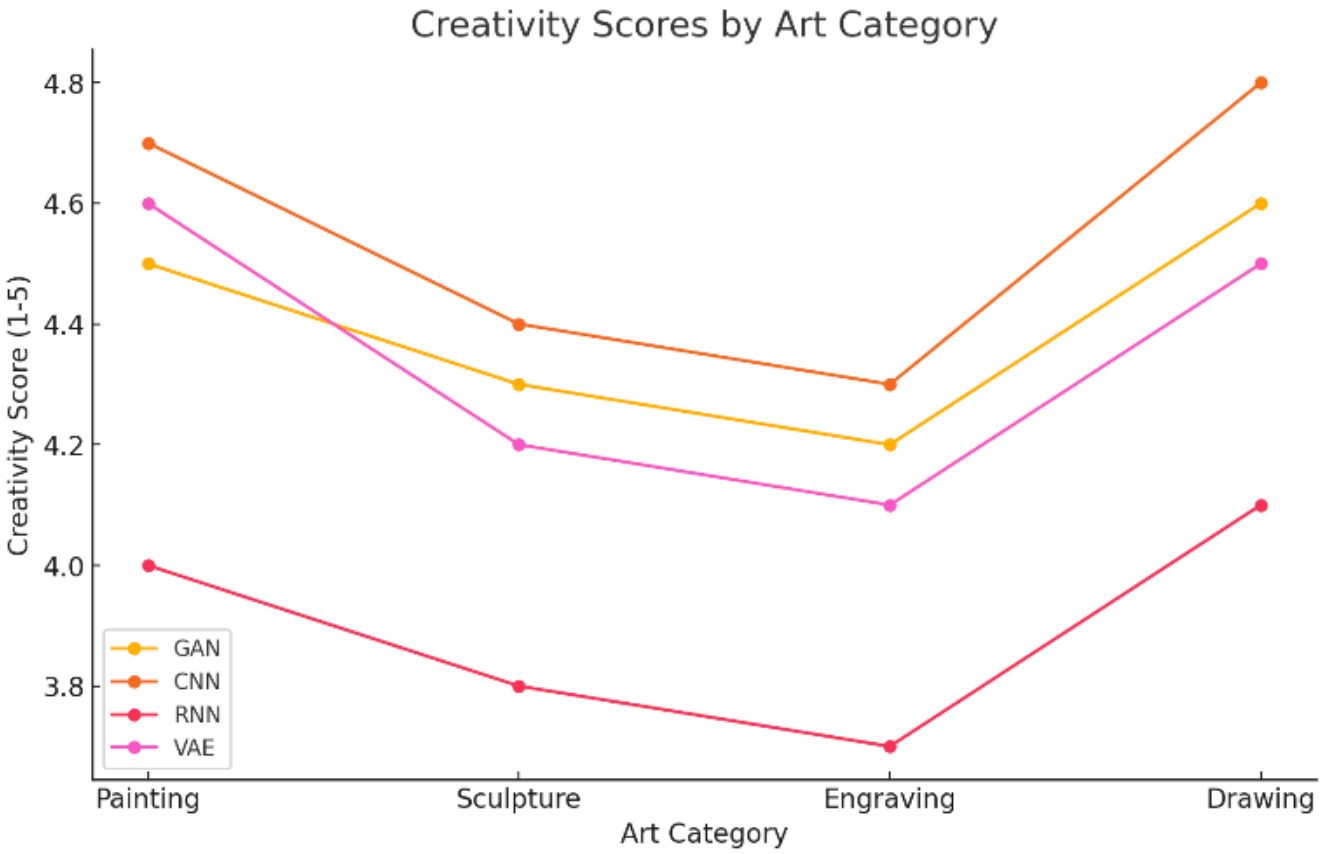


Fig. 3. Creativity scores by art category.

The provided examples also produced quite creatively the drawn and painted characters. They scored 4.5 and 4.6 in this part of the competition, respectively. The utilization of the option that is entirely new in art creation was a major part of their success. VAEs were particularly for tasks that required some level of abstraction, for example, SCULPTURE, and the CHOCOLATE area earned points of 4.2 and 4.1, respectively. On the other hand, RNNs were the technology that was good at most elements of art but were not very good at easy art forms such as static pieces. Instead, the program had much more success in the dynamic tasks and the tasks that required a user to be mobile. For instance, the level of user input in the problem was pretty high, and therefore it proved that the RNNs were extremely capable of doing these types of tasks. "Table III shows creativity scores metrics.

TABLE III. CREATIVITY SCORES METRICS

Art Category	GAN Creativity Score	CNN Creativity Score	RNN Creativity Score	VAE Creativity Score
Painting	4.5	4.7	4.0	4.6
Sculpture	4.3	4.4	3.8	4.2
Engraving	4.2	4.3	3.7	4.1
Drawing	4.6	4.8	4.1	4.5

C. Comparative Insights

The observations that have been made show that the specific type of deep learning model chosen has an important part in determining the number of outputs that will be produced as well as the quality of the outputs in outputs in the entire variety of art categories. CNNs are the ones that come out on top in the accuracy of work and creativity for static visuals. GANs are the leaders in content creation, even more so in daring and experimental art styles, which are uncontrollable. VAEs are not very good in terms of precision but still have the unique advantage of surrealism and conceptual art. RNNs are also at the experienced art level. However, they could be seen as they will help future interactive and narrative-driven pieces of art develop further.

D. Proposed Model Performance

The proposed model, as illustrated in Fig. 1, incorporates various deep learning techniques to provide a complete solution for art and design jobs. The pipeline of the model—from the data preparation stage to the application of GANs, CNNs, RNNs, and VAEs—takes a comprehensive view of the creative improvement process. The performance metrics obtained from the model and the visual and creativity scores obtained from the outputs prove the efficacy of the proposed model in combining computational methods with artistic expression.

The AI systems have undergone assessment on different parameters like accuracy, creativity factor, and user engagement level. For calculating accuracy, the percentage of the correctly classified or generated art styles compared to the manually annotated database is used. Creativity scores were rated by human judges on a scale ranging from 1 to 5, which considered the factors such as originality, aesthetic value, and conformity with artistic intent. The frequency of the use of tools and the duration of their usage were the features that indicated user engagement. Besides achieving the accuracy (92%) in drawing classification the CNNs system also obtained outstanding creativity scores (4.6) for style generation.

V. Discussion

Although AI systems are extremely productive and versatile, they cannot replace traditional art methods, which are indispensable for creativity that is sensitive to context and human nuance. When it comes to AI-generated content, human emotions remain absent and subjective intent is lacking. But then again, speed, relevance and the fusion of media forms are areas in which AI brings innovation. For example, traditional oil painting that requires time-consuming layering and drying is replaced by the AI-based digital painting tools that the artist can use to apply colors and textures in real-time. A direct comparison of accuracy and efficiency between AI-generated and traditionally-created designs revealed that AI-assisted workflows reduced production time by approximately 40% while maintaining stylistic fidelity.

With the emergence of AI-generated art comes the need to consider the ethics of authorship, originality, and intellectual property. AI models can be trained on publicly available artwork and may inadvertently replicate copyrighted styles which causes a fear of ownership in the art world. At the same time, training data biases can lead to the reinforcement of stereotypes and the limitation of diversity in generated content. To fix these problems, it is important to have transparency in AI development, dataset curation that represents all groups, and a clear definition of human-AI collaboration on creative projects.

In order to reduce the risk of biases in the AI-generated artistic outputs, we made sure to curate a dataset that is made of many different cultural, gender, and thematic representations. We made use of data augmentation techniques, which helped us in balancing styles that were not adequately represented through other means. We also employed fairness-aware learning algorithms such as adversarial debiasing to make the model less reliant on the dominant artistic movements. The last step we adopted involved humans evaluating the designs to make sure they were ethically aligned.

In conclusion, the findings indicate that deep learning can significantly change art and design. The proposed method provides an adaptable framework for increasing creativity and innovation by adjusting the models to particular artistic domains and utilizing their distinctive benefits. The employment of deep learning is transformative for conventional creative practices and opens new avenues for novel expressions, and interdisciplinary collaboration in the expressive arts sector.

VI. Conclusion

The integration of deep learning into art and design has been a groundbreaking process that has enabled the release of creativity, efficiency, and innovation at levels which have never been seen before the technologies were involved. The research conducted in this paper shows the abilities of models like CNNs, GANs, RNNs, and VAEs to improve the creative workflows of artists and designers in different disciplines like painting, sculpture, engraving, and drawing. The results revealed that CNNs achieved the best accuracy (92%) and creativity scores (4.8) in static visual art forms, whereas GANs were the most capable in generating new and experimental content and scored 4.6 on the creativity scale for drawings. VAEs seem to have some application in art based on a conceptual and abstract basis, while RNNs pointed out their usefulness in dynamic and narrative-driven tasks. These findings confirm that different models have their own unique advantages in the creative process, which can be utilized to fit art and design needs. The pipeline model which incorporates these various technologies into a unified structure was thus demonstrated to be very much adaptable to fit many purposes, efficiently with respect to the application of other forms of technology, a point which was justified both by quantitative metrics and visual evaluations.

Even though this study has shown encouraging outcomes, it still has some limitations. On the one hand, the general use of the Art Images Dataset, which is rather inclusive, could give the wrong impression that it is the only relevant source of information; it is highly unlikely however to know the breadth of artistic styles and cultural backgrounds without using some additional sources. On the other hand, deep learning models, which are groundbreaking technologies for creativity, come at a cost: they need to rely on the biases inherent in the training data which, in turn, can generate imprecise results. The requirements for computation when it comes to training these systems and fine-tuning them are a barrier to entry in terms of affordability. The next phase of research could be based on these limitations by using a broader variety of datasets, considering smaller versions of the models that usually are more cost-effective for making a larger number of people benefit from them and analyzing the ethical aspects of AI-generated art. All these measures would ensure that not only are the deep learning technologies used in creative industries but also the whole process is more accountable and widely applicable. Growth of future research should check real-time AI-supported artistic collaboration systems, incorporating customer input in the reactive fashion of the generator. Developing AI's use in 3D design, simulated immersive environments, and bespoke story-telling could open more doors. Moreover, art created using AI that can be made more interpretable will give clarity and make AI remain an add-on instead of a substitute for human creativity.

Acknowledgment

This work was supported by Key Project of the Research Center for Beautiful Countryside Construction and Development of Chengdu Key Research Base of Social Sciences: Imaging Research of Daoming Bamboo Weaving National Non-Genetic Inheritors, Project No. CCRC2020-2.

REFERENCES

[1] M.-J. Lee and E. Choi, "A Study on Creative Nail Art Design Generation Based on Text Prompt: Focused on Image-Generating Artificial Intelligence Models, DALL-E 2 and Bing Image Creator," Journal of the Korean Society of Cosmetology, vol. 29, no. 4, 2023, doi: 10.52660/jksc.2023.29.4.1058.

[2] L. Zhao, "International Art Design Talents-oriented New Training Mode Using Human-Computer Interaction based on Artificial Intelligence," International Journal of Humanoid Robotics, vol. 20, no. 4, 2023, doi: 10.1142/S0219843622500128.

[3] R. Risandhy and S. Q. Nada, "Perancangan Video Motion Graphic mengenai Dampak Artificial Intelligence dalam Art & Design," Journal of Visual Communication Design, vol. 3, no. 2, 2023.

[4] J. Lively, J. Hutson, and E. Melick, "Integrating AI-Generative Tools in Web Design Education: Enhancing Student Aesthetic and Creative Copy Capabilities Using Image and Text-Based AI Generators," DS Journal of Artificial Intelligence and Robotics, vol. 1, no. 1, 2023, doi: 10.59232/air-v1i1p103.

[5] A. A. Alegaonkar and M. A. Avachat-Shirke, "IS ARTIFICIAL INTELLIGENCE KILLING ARTISTIC SKILLS IN DESIGNERS?," ShodhKosh: Journal of Visual and Performing Arts, vol. 4, no. 2SE, 2023, doi: 10.29121/shodhkosh.v4.i2se.2023.484.

[6] Y. Zhang and Z. Jin, "Optimization Strategy of Cultural Creativity Product Art Design Based on Artificial Intelligence and CAD," Comput Aided Des Appl, vol. 21, no. S14, 2024, doi: 10.14733/cadaps.2024.S14.299-314.

[7] K. Terzidis, F. Fabrocini, and H. Lee, "Unintentional intentionality: art and design in the age of artificial intelligence," AI Soc, vol. 38, no. 4, 2023, doi: 10.1007/s00146-021-01378-8.

[8] L. Yin and Y. Zhang, "Artistic Style Transformation Based on Generative Confrontation Network," Comput Aided Des Appl, vol. 21, no. S13, 2024, doi: 10.14733/cadaps.2024.S13.48-61.

[9] Y. Ran, "GAN and art: Facilitation of artistic production and expression based on artificial intelligence," Applied and Computational Engineering, vol. 16, no. 1, 2023, doi: 10.54254/2755-2721/16/20230887.

[10] K. German, M. Limm, M. Wölfel, and S. Helmerdig, "Co-designing object shapes with artificial intelligence," in Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST, 2020. doi: 10.1007/978-3-030-53294-9_21.

[11] J. Hutson and M. Lang, "Content creation or interpolation: AI generative digital art in the classroom," Metaverse, vol. 4, no. 1, 2023, doi: 10.54517/m.v4i1.2158.

[12] R. West and A. Burbano, "Ai, arts & design: Questioning learning machines," Artnodes, vol. 2020, no. 26, 2020, doi: 10.7238/a.v0i26.3368.

[13] I. Mahy and D. Bubna-Litic, "Breaking Through: A journey towards deep learning for the 21st century," 2015.

[14] N. Meng, J. Yang, and H. Wang, "Icon Art Design with Generative Adversarial Network under Deep Learning," Wirel Commun Mob Comput, vol. 2022, 2022, doi: 10.1155/2022/3499570.

[15] Y. S. Chang, Y. Y. Wang, and Y. Te Ku, "Influence of online STEAM hands-on learning on AI learning, creativity, and creative emotions," Interactive Learning Environments, 2023, doi: 10.1080/10494820.2023.2205898.

[16] S. A. Chauncey and H. P. McKenna, "A framework and exemplars for ethical and responsible use of AI Chatbot technology to support teaching and learning," Computers and Education: Artificial Intelligence, vol. 5, 2023, doi: 10.1016/j.caeai.2023.100182.

[17] S. K. Choi, S. DiPaola, and L. Gabora, "Art and the artificial," Journal of Creativity, vol. 33, no. 3, 2023, doi: 10.1016/j.yjoc.2023.100069.

[18] D. Satrinia, R. R. Firman, and T. N. Fitriati, "Potensi Artificial Intelligence dalam Dunia Kreativitas Desain," Journal of Informatics and Communication Technology (JICT), vol. 5, no. 1, 2023, doi: 10.52661/j_ict.v5i1.164.

[19] J. Ploennigs and M. Berger, "AI art in architecture," AI in Civil Engineering, vol. 2, no. 1, 2023, doi: 10.1007/s43503-023-00018-y.

[20] M. A. Ali Elfa and M. E. T. Dawood, "Using Artificial Intelligence for enhancing Human Creativity," Journal of Art, Design and Music, vol. 2, no. 2, 2023, doi: 10.55554/2785-9649.1017.

[21] A. F. C. A. Fathoni, "Leveraging Generative AI Solutions in Art and Design Education: Bridging Sustainable Creativity and Fostering Academic Integrity for Innovative Society," in E3S Web of Conferences, 2023. doi: 10.1051/e3sconf/202342601102.

[22] B. Hokanson, "The Technology of the Question: Structure and Use of Questions in Educational Technology," Educational Technology, vol. 55, no. 6, 2015.

[23] Y. Li, "Research on Modern Book Packaging Design under Aesthetic Evaluation Based on Deep Learning Model," 2021. doi: 10.1155/2021/8214200.

[24] D. C. Elton, Z. Boukouvalas, M. D. Fuge, and P. W. Chung, "Deep learning for molecular generation and optimization - a review of the state of the art," ArXiv, no. March, 2019.

[25] S. Deshpande and A. Purwar, "Computational creativity via assisted variational synthesis of mechanisms using deep generative models," Journal of Mechanical Design, vol. 141, no. 12, 2019, doi: 10.1115/1.4044396.

[26] S. José Venancio Júnior, "Extentio: Artistic Human-Machine Collaboration," in LINK 2021 Conference Proceedings, 2022. doi: 10.24135/link2021.v2i1.143.

[27] M. McLain, "Towards a signature pedagogy for design and technology education: a literature review," Int J Technol Des Educ, vol. 32, no. 3, 2022, doi: 10.1007/s10798-021-09667-5.

[28] Y. Jin and Z. Tiejun, "The application of Metaverse XiRang game in the mixed teaching of art and Design in Colleges and Universities," Educ Inf Technol (Dordr), vol. 28, no. 12, 2023, doi: 10.1007/s10639-023-11844-z.

[29] K. Noble, "'Getting Hands On with Other Creative Minds': Establishing a Community of Practice around Primary Art and Design at the Art Museum," International Journal of Art and Design Education, vol. 40, no. 3, 2021, doi: 10.1111/jade.12371.

[30] N. Saleeb, "Closing the chasm between virtual and physical delivery for innovative learning spaces using learning analytics," International Journal of Information and Learning Technology, vol. 38, no. 2, 2021, doi: 10.1108/IJILT-05-2020-0086.

[31] X. Li, Y. Wang, and Z. Sha, "Deep Learning Methods of Cross-Modal Tasks for Conceptual Design of Product Shapes: A Review," 2023. doi: 10.1115/1.4056436.

[32] W. Zhao and Y. Sun, "The Exploration of Emotional Aspects of Artificial Intelligence (AI) in Artistic Design," International Journal of Interdisciplinary Studies in Social Science, vol. 1, no. 1, 2024, doi: 10.62309/bk757m16.

[33] L. Qiu, A. R. A. Rahman, M. S. Bin Dolah, and S. Ge, "The Research on the Application of Artificial Intelligence in Visual Artbased on Souvenir Design," WSEAS Transactions on Information Science and Applications, vol. 21, 2024, doi: 10.37394/23209.2024.21.6.

[34] Q. Cai, X. Zhang, and W. Xie, "Art Teaching Innovation Based on Computer Aided Design and Deep Learning Model," Comput Aided Des Appl, vol. 21, no. S14, 2024, doi: 10.14733/cadaps.2024.S14.124-139.

[35] K. German, M. Limm, M. Wölfel, and S. Helmerdig, "Towards Artificial Intelligence Serving as an Inspiring Co-Creation Partner," EAI Endorsed Transactions on Creative Technologies, vol. 6, no. 19, 2019, doi: 10.4108/eai.26-4-2019.162609.

Scallop Segmentation Using Aquatic Images with Deep Learning Applied to Aquaculture

Wilder Nina[1], Nadia L. Quispe[2], Liz S. Bernedo-Flores[3], Marx S. García[4], Cesar Valdivia[5], Eber Huanca[6]
VEOX Inc., Arequipa, Peru[1,2]
Instituto Tecnológico de la Producción - CITE pesquero Piura[4]
Departamento de Ingeniería Eléctrica y Electrónica, Universidad Católica San Pablo, Arequipa, Peru[3,5,6]

***Abstract*—This study evaluates the performance of deep learning-based segmentation models applied to underwater images for scallop aquaculture in Sechura Bay, Peru. Four models were analyzed: SUIM-Net, YOLOv8, DETECTRON2, and CenterMask2. These models were trained and tested using two custom datasets: SEG_SDS_GOPRO and SEG_SDS_SF, which represent diverse underwater scenarios, including clear and turbid waters, varying current intensities, and sandy substrates. The primary aim was to automate scallop identification and segmentation to improve the efficiency and safety of aquaculture monitoring. The evaluation showed that SUIM-Net achieved the highest accuracy of 93% and 94% on the SEG_SDS_GOPRO and SEG_SDS_SF datasets, respectively. CenterMask2 performed best on the SEG_SDS_SF dataset, with an accuracy of 96.5%. Additionally, a combined dataset was used, where YOLOv8 achieved an accuracy of 88%, demonstrating its robustness across varied conditions. Beyond scallop segmentation, the models were extended to detect six additional marine classes, achieving a maximum accuracy of 39.90% with YOLOv8. This research underscores the potential of deep learning techniques to revolutionize aquaculture by reducing operational risks, minimizing costs, and enhancing monitoring accuracy. The findings contribute valuable insights into the challenges and opportunities of applying artificial intelligence in underwater environments.**

Keywords*—Image segmentation; object detection; deep learning; computer vision; aquaculture; scallop segmentation; aquatic images

I. Introduction

Scallop cultivation has been developed in Peru since the 1970s, initially on the central coast. Due to its rapid growth, it expanded to other regions along the Peruvian coastline. By the early 1990s, scallop farming became an important economic activity, generating significant employment opportunities in aquaculture [1]. Sechura Bay, located on the northern coast of Peru (latitude: -5.742118, longitude: -80.867822) (see Fig. 1), is one of the largest semi-enclosed bays in the Peruvian sea, covering approximately 89 kilometers. It is bordered by Punta Gobernador to the north and Punta Aguja to the south, with a cultivation area of 6,752.48 hectares (Ha). The bay has depths of less than 30 meters within 10 kilometers of the coastline. Numerous scallop hatcheries have been established on the seabed in this region. During the early stages of production, scallop seeds are classified based on their valve length, which typically ranges from 2.5 cm to 4.0 cm. They are cultivated at a density of 1.5 bunches per square meter, where each bunch contains 96 seeds [2]. During hatchery cultivation, seeds that are confined or have irregular sizes are extracted and replanted in better-suited areas. At this stage, growth and mortality rates are also monitored to estimate harvest levels [3]. Scallop stock assessments in hatcheries are conducted monthly or at least once before harvesting. These assessments involve divers manually collecting population samples using a 1m² squared frame tool. Divers carefully place the frame on the seabed and harvest the scallops within its boundaries. On the support boat, the collected scallops are counted and measured using a vernier caliper and a malacometer, and the data is recorded. This process is repeated multiple times to estimate the total scallop stock in units and bunches [2]. However, these manual operations pose significant risks to divers. The average hatchery depth is 10 meters, requiring divers to spend extended periods underwater. While safety guidelines recommend a maximum working time of four hours per day, divers often exceed this limit, working up to seven or eight hours. Additionally, divers rely on inadequate compressed air systems, which can lead to nitrogen narcosis, posing a severe health risk, including fatal consequences [4]. To address these challenges, the integration of advanced methodologies and technologies for scallop production assessment is essential. Automation in data collection and analysis can significantly reduce operational risks, costs, and time associated with aquaculture monitoring.

This study explores the application of deep learning and computer vision to automate scallop segmentation, aiming to enhance accuracy, improve efficiency, and minimize risks. Two custom datasets (SEG_SDS_GOPRO and SEG_SDS_SF) were developed to evaluate the performance of four segmentation models: SUIM-Net, YOLOv8, DETECTRON2, and CenterMask2. Additionally, these models were extended to detect six marine classes, expanding their applicability in aquaculture. This research provides valuable insights into leveraging artificial intelligence for sustainable and efficient scallop farming.

This research was carried out as part of a project (Monitoring platform for non-extractive sampling of hydrobiological resources through the development of a customized underwater robot with advanced deep learning and computer vision techniques. Application case: sea floor hatchery of scallops (argopecten purpuratus) in Sechura bay) financed by the Peruvian research council FONDECYT.

This work is organized into six sections. Section II introduces related works, including the methodology for performing the datasets and SUIM-Net, YOLOv8, DETECTRON2, and CenterMask2 models; Section III describes the datasets and methods; Section IV describes experiments and analysis; Section V describes discussion of Results; finally, Section VI summarizes the work.

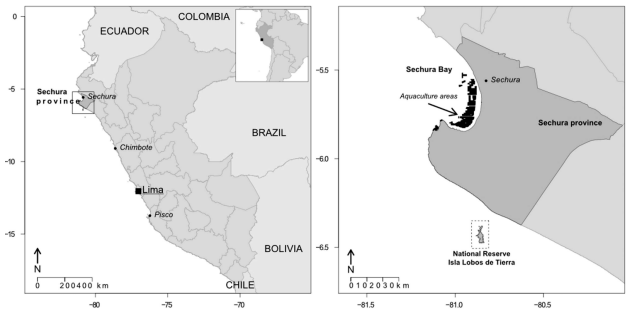


Fig. 1. Left panel: Situation the study setting Sechura Bay in Secura province in Peru. Right panel: Sechura Bay in the province of Sechura, indicating the aquaculture concession area [3].

II. Related Works

One of the most popular underwater datasets for marine species detection and classification is F4K [5]. This dataset was performed using 10 cameras between 2010 and 2013 in Taiwan and has been used for multiple detection and classification algorithms. The F4K dataset is large and contains videos and images with complex scenes and diversity of marine species. Another large data set is the Electronic Library of Deep Sea Images (J-EDI) [6] which are consistuded of videos and images of deep-sea organisms captured by remotely operated underwater vehicles (ROVs) [7]. This images are labeled at the image level and have been used to interconnect convolutional neural networks - CNN for the detection of deep-sea hydrobiological organisms. The author in [8] considers three objectives for collecting underwater images: (1) Broad diversity of underwater scenes, having different quality degradation characteristics and a wide range of image content (2) Big amount of data and (3) High quality of images. The author in [9] introduces USR-248 dataset, which is a large-scale data set of three sets of images, which were rigorously collected during ocean explorations, field experiments and some resources are publicly available online. The UFO-120 dataset is depicted in [10] which contains more than 1500 trining sample. These images were captured in oceanic explorations in many different locations and kinds of water. [9] presents the USR-248 dataset, which main characteristic is its capability for supervised training. The author in [11] presents a survey of deep learning techniques for performing the underwater image classification (see Table I).

TABLE I. Undersea Dataset [11]

Name	Object	Class	Images
Sipper	Zooplankton	81	>750K
WHOI	Plankton	70	>3.4M
ZooScan	Zooplankton	20	3,771
ZOOVIS	Zooplankton	6	>685K
ISIIS	Plankton	108	>42K
PlanktonSet	Plankton	121	>60K
ZooLake	Plankton	35	17,943
F4K	Pez tropical	23	27,370
LifeCLEF14	Pez tropical	10	19,868
LifeCLEF15	Pez tropical	15	>20K
Temperate fish	Temperate Fish	4	619
Fish-gres	Pez	8	3248
MLC	Arrecife Coral	9	2055
CADDY	Gestos de buzo	16	10,322

The underwater environment is one of the most challenging conditions for object detection using sensors. For this, sonars and cameras are mainly used. Sonars are sensitive to geometric structure and provide information in an environment of very low or no visibility, but their drawback is that they can only measure the difference in distance between objects [12]. On the other hand, underwater images allow colors, textures and edges to be easily detected as long as visibility conditions exist. However, in a real environment these conditions are altered, resulting in degradation of image quality due to wavelength-dependent absorption and dispersion, including forward and backward scattering. In addition, marine snow introduces noise and increases dispersion effects, reducing visibility, contrast and even distorting colors. Despite this, it has more potential to detect characteristics of objects, compared to sonar sensing (applied for scallop production assesment). But for this it is necessary to pre-process the images to improve their quality and extract information [6], [8]. Object detection in computer vision can be acquired with a single camera or multiple cameras (stereo vision). In [13] was developed algorithms for image detection using a stereo vision camera system for detecting objects of sea floor without previous information. The region of interest - ROI was determined by pixel similarity concentrations in *area*, *color* and *shape*. These criteria were applied to segment and validate the image recognition algorithms. For the experimental tests [13] used three underwater datasets: Garda, Portofino and Soller; each one has its own characteristics and challenges, which allowed the author to evaluate it in different underwater situations, demonstrating that the proposed algorithm is robust to changes in lighting and turbidity of the water, however it does not consider the challenges of superimposing objects intermingled on top of each other.

The UFO-120 dataset [10], it addresses the problem of simultaneous enhancement and super-resolution (SESR) for underwater robotic vision, providing an efficient solution for real-time applications. It also introduces the Deep SESR model, which is a generative model based on a residual network that can learn to restore the perceptual qualities of the image. The proposed Deep SESR model offers perceptually improved FC imaging and saliency prediction through a single efficient inference. Enhanced images restore color, contrast and sharpness at higher scales to facilitate better visual perception. For semantic segmentation [14] proposed to improve the encoder and decoder structures of the DeepLabv3+ network, in order to improve the appearance of the segmenting object and prevent its pixels from mixing with the pixels of other classes, improving the accuracy of the segmentation of the object boundaries and preserving feature information. To do this, it added the unsupervised color correction - UCM method in the encoder structure to improve the image quality, then added two upsampling layers to the decoder structure to obtain greater feature information using the (*backbone*) Xception_65. Regarding the data set, this was self-made, some images were obtained from public resources on the Internet and another part was obtained from video images taken by an underwater laboratory robot (HUBOS-2K, Hokkaido University).

Due to the absorption of light and the deeper the water, underwater images usually acquire a greenish-blue color. To counteract this effect [15] applies a combination of maximum RGB method and gray tone method to achieve underwater vision improvement. Then it proposes a method based on

CNN for the detection and classification of objects in real time, using the third version of You Only Look Once network (YOLOv3), according to the characteristics of underwater vision, two improved schemes are applied to modify the CNN structure. The proposed YOLOv3 framework is divided into three parts: feature extraction, object detection, and prediction of bounding box coordinates and object confidence. Feature extraction is performed using the Darknet-53 network. Object detection is performed by detecting objects at multiple scales, and predicting bounding box coordinates and object confidence is performed by a fully connected neural network. The database used for the tests was obtained from a video recorded by an underwater robot, with approximately 18000 images. In [16] was proposed an underwater scallop recognition algorithm using an improved version of the YOLOv5 [17] neural network. The authors first designed a new lightweight backbone model to replace the original one of YOLOv5, using group convolution and inverse residual block, which helps improve the accuracy and accelerate the detection speed. The proposal used the k-means algorithm for clustering analysis of the data set to reduce the initial prediction layer of the model and the enhancement module was used by the adaptive dark channel algorithm to improve the clarity of the blurred image. The data set used was self-constructed in a laboratory environment, with a swimming pool and a GoPro 5 camera, 2200 labeled images were acquired. The data was increased to 4400 using dark channel image enhancement. The experimental results indicate that the precision rate, recall rate, F1 and mPA of the proposed algorithm reached 90.8%. In [18] was proposed state-of-the-art real-time object detection algorithms are trained and inferred on underwater images of a hypothetical inshore aquaculture operation to investigate model selection and hyper-parameters for object detection in underwater images. The simulation results show that 54.2% mean average precision is achieved by YOLOX-m and 97.1 frames per second inference processing confirmed by YOLOX-tiny.

In [19] was proposed a detection and Classification of Subsea Objects in Forward-Looking Sonar and Electro-Optical Sensors for ROV Autonomy The objects of interest are MLOs(Mine Like Objects) that need to be located, identified, and inspected by an autonomous submersible robot. SeeByte used Deep Learning Neural Network (DNN) on both of these sensor feeds yielding a very robust detection and classification system.

state-of-the-art real-time object detection algorithms are trained and inferred on underwater images of a hypothetical inshore aquaculture operation to investigate model selection and hyper-parameters for object detection in underwater images. The simulation results show that 54.2% mean average precision is achieved by YOLOX-m and 97.1 frames per second inference processing confirmed by YOLOX-tiny.

Related work is crucial to contextualize our results within the advancement of underwater image segmentation. Previous studies have addressed segmentation in marine environments with different approaches, from CNN-based models to more advanced architectures such as Transformer-based vision models. In particular, works such as [10] and [11] have demonstrated the importance of improving image quality to optimize detection. Our study complements these efforts by evaluating models in real-world conditions of fan shell farming in Sechura Bay, an environment that presents unique challenges of visibility and environmental variability.

III. Datasets and Methods

In this study, a customized dataset of scallop images was collected from a seabed nursery in the bay of Sechura (Peru), under various underwater environmental conditions. The Fig. 2 illustrates the methodology, which is structured in three main stages: Image Acquisition, Image Preprocessing, and Scallop Training and Detection.

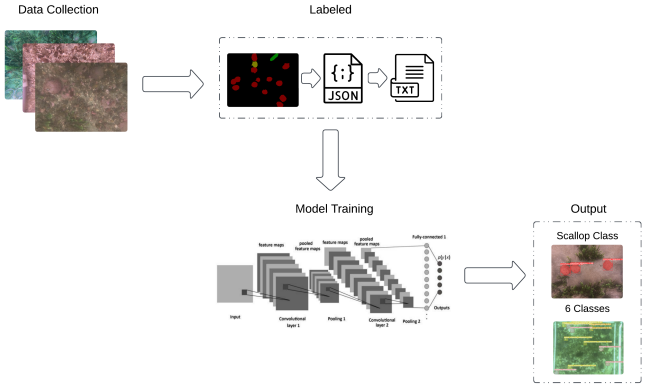


Fig. 2. Structure of the methodology used.

A. Datasets

Unlike reference datasets in underwater image segmentation, such as F4K and UFO-120, which were collected in previous decades, our study is based on recent images captured in Sechura Bay. Data collection was conducted in 2023 and 2024 using high-resolution cameras, which ensures that the dataset accurately reflects current ecosystem conditions. This update is fundamental to improve the applicability of the models in real aquaculture monitoring scenarios, allowing for more accurate and adaptive detection of environmental changes (Table II).

TABLE II. Descriptions of Datasets

DataSet name	Type	Description
SEG_SDS_GOPRO (SEGMENTATION SCALLOP DATASET WITH GOPRO)	Training and Validation	It uses a PVC sampling structure with a GoPro H9 camera for image segmentation.
SEG_SDS_SF (SEGMENTATION SCALLOP DATASET WITH SMARTFRAME)	Training and Validation	It uses a steel tube structure called smartframe v1.0 adapted Raspberry Pi V2.1 camera include enclosure.

Given the limited availability of databases featuring marine species in the reviewed state-of-the-art literature, a custom dataset was created to guarantee the presence of scallops in the images. This approach also facilitates pretraining the neural network with relevant data. Table III lists the visited locations within Sechura Bay (see Fig. 1) where scallop images were collected. These locations span the northern,

central, and southern regions, capturing a variety of underwater topographies, including sandy, rocky and mixed terrains.

TABLE III. HATCHERIES WHERE SCALLOP IMAGES WERE ACQUIRED

Zone	Scallop hatchery name in Sechura bay
North	Chulliyache - Caballero de los Mares, boarding at the Matacaballo DPA
Center	Las Delicias, boarding at the Parachique Zonal Fishery Terminal
South	Barrancos - Amigos Unidos, boarding at the Parachique Zonal Fishery Terminal
South	Sea Corp Camp, Vichayo Aquaculture Production Center (CPAV), Vichayo Zone

Scallop images were captured manually with the support of a professional diver and a support boat using two types of cameras: (1) GoPro Hero 9 and (2) Raspberry Pi V2.1. The cameras were mounted on a square frame with a 1-meter edge, elevating them 40 cm above the seabed. Fig. 3 illustrates the image acquisition process.

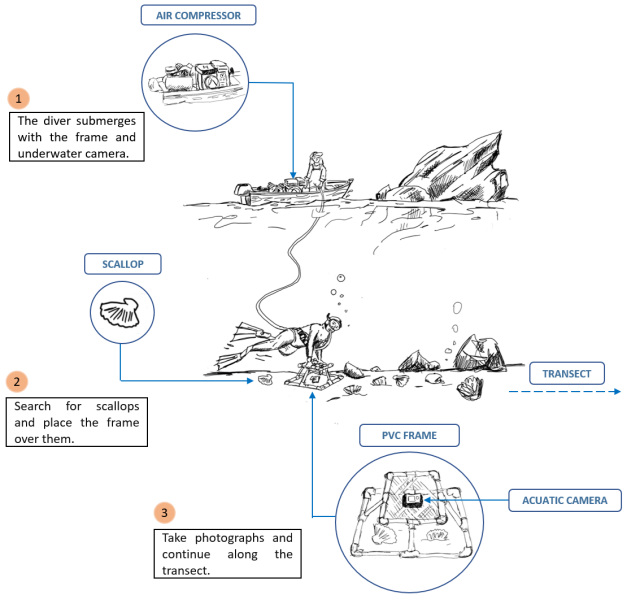


Fig. 3. Image acquisition process.

To ensure accurate scallop identification, the collected images were manually labeled with the assistance of a hydrobiology specialist from the Center for Productive Innovation and Technology Transfer (CITE) in Piura, Peru. This labeling process was critical for creating a reliable training dataset.

Fig. 4 and 5 present the underwater environment of Sechura Bay, revealing the composition of the seafloor and the diversity of its hydrobiological resources. These images provide a detailed visual representation of the natural habitat inhabited by scallops, along with other marine species, highlighting the biological richness captured in the dataset. Understanding the characteristics of this ecosystem is fundamental for the correct annotation and classification of the images, ensuring the accuracy and reliability of the dataset used in the training of deep learning models.

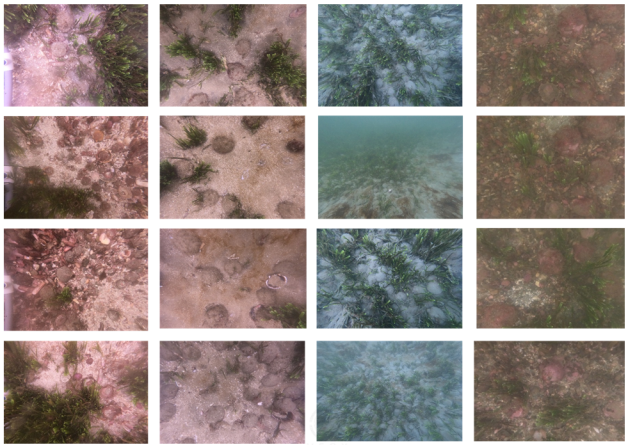


Fig. 4. Dataset SEG_SDS_GOPRO acquired in Sechura Bay

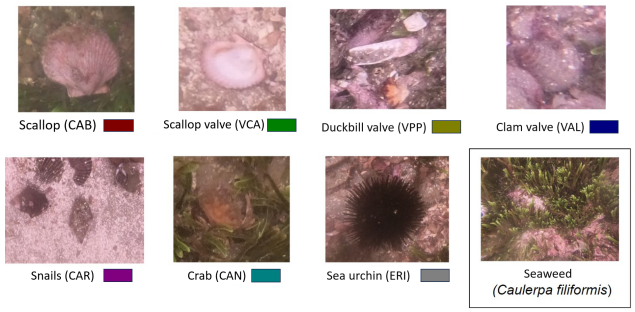


Fig. 5. Marine species considered in the labeling.

Additionally, Fig. 6 shows the SEG_SDS_SF dataset acquired in Sechura Bay.

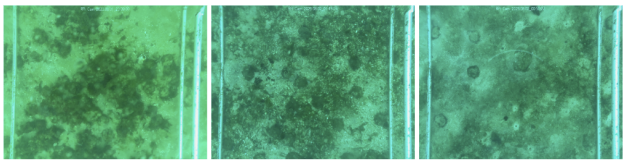


Fig. 6. Dataset SEG_SDS_SF acquired in Sechura Bay.

In Table IV, the classes with which the training is carried out and the labels of each of them from the SEG_SDS_GOPRO and SEG_SDS_SF dataset are observed.

TABLE IV. DATASET CLASSES AND LABELS

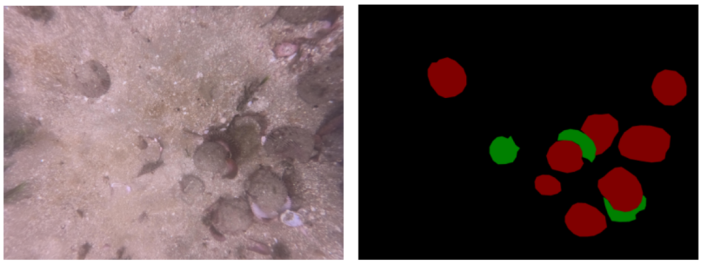
DataSet name	Classes	Labels
SEG_SDS_GOPRO	Scallop	2418
	Scallop Shell	165
	Duck Beak Shell	163
	Snail	57
	Crab	9
	Clam Shell	4
SEG_SDS_SF	Scallop	715
	Scallop Shell	42
	Duck Beak Shell	683

B. Labeling Image Process

The acquired images were labeled and reviewed by the CITE expert for validation or correction of the labeling. For

labeling images, seven species of scallops where considered: (1) Scallop - CAB, (2) Scallop valve - VCA, (3) Duckbill valve - VPP, (4) Clam valve - VAL, (5) Snails - CAR, (6) Crab - CAN, and (7) Sea urchin - ERI, Fig 7 shows each scallop specie image. For each one a different color was assigned. The seaweed (*Caulerpa filiformis*) in labeling process is considered as part of seabeed. The images were labeled with the "labelme" tool.

A validation table was created to document image characteristics, including image ID, collection day, filtering status, species present, quantity per species, and seabed type. This table facilitated coordination with the specialist for label verification and served as an inventory.



a) Original image b) Labeled image

Fig. 7. Results of labeling image process.

C. Models

The models selected for this study were SUIM-Net, YOLOv8, DETECTRON2 and CenterMask2, due to their balance between accuracy and computational efficiency in underwater image segmentation. Also, at the time of selection, these models represented some of the most recent and advanced versions of their respective architectures, which guaranteed better performance in detection and segmentation tasks.

1) SUIM-Net: The SUIM-Net is a fully convolutional semantic segmentation model (FCN) proposed by [20], which is based on an encoder-decoder architecture with skip connections between composite layers. The network employs residual learning and optional residual skip blocks to enhance performance. The encoder extracts features from input images, and the decoder generates binary pixel labels per channel for each object category. It utilizes a proprietary dataset called SUIM (Semi-supervised Underwater Image Manipulation) which is presented in Fig. 8. For training, it applies various image transformations for data augmentation. The results obtained show that SUIM-Net exhibits improved execution time compared to FCN models, SegNet, UNet, VGG-based encoder-decoders, DeepLab, and PSPNet

2) YOLOv8: YOLOv8, developed by Ultralytics, builds upon YOLOv5 with key improvements, the model is shown in Fig. 9 where it uses a similar backbone with modifications in the CSPLayer, now called the C2f module, to combine high-level features and contextual information, enhancing detection accuracy. YOLOv8 adopts an anchor-free design with a decoupled head, allowing separate processing of objectivity, classification, and regression tasks for improved accuracy.

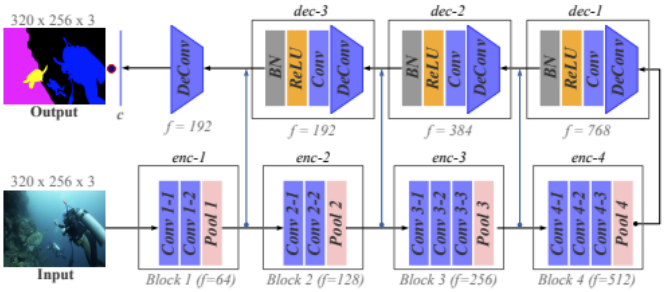


Fig. 8. SUIM-Net model [20].The end-to-end architecture of SUIM-Net$_{VGG}$: first four blocks of a pre-trained VGG-16 model are used for encoding, followed by three mirrored decoder blocks and a deconv layer.

The output layer applies the sigmoid function for objectivity scores and softmax for class probabilities [21]. YOLOv8 also includes a semantic segmentation variant, YOLOv8-Seg, which uses CSPDarknet53 as the backbone and C2f for feature extraction. It features two segmentation heads for mask prediction and detection heads with five modules for bounding box prediction [22].

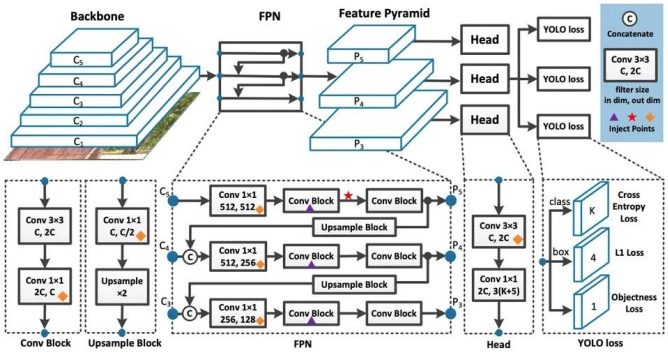


Fig. 9. YOLOv8 model [23].

3) DETECTRON2: DETECTRON2 [24] allows us to implement models of different architectures. These pretrained models in the DETECTRON2 Zoo Model have a structure that follows the meta-architecture provided by the base code. It is a framework that includes high-quality implementation of state-of-the-art instance segmentation algorithms, such as Faster R-CNN and Mask R-CNN. The architecture of the DETECTRON2 model in Fig. 10, consists of three main parts: the backbone network, region proposal network (RPN), and ROI (Region of Interest) head. Given an input image, the backbone network extracts feature maps at various scales with different receptive fields. The RPN detects object regions from feature maps and various scales by default. Finally, the box head crops the feature maps into different sizes and finds the location of the boxes and classification labels in addition to fully connected layers. In our tests, we used ResNet with transformers. ResNets are convolutional neural networks that utilize skip connections enabling a deep architecture with many layers.

4) CenterMask2: The CenterMask2 model is an instance segmentation approach that decomposes this task into two main parallel subtasks [26]. The model is shown in Fig. 11,

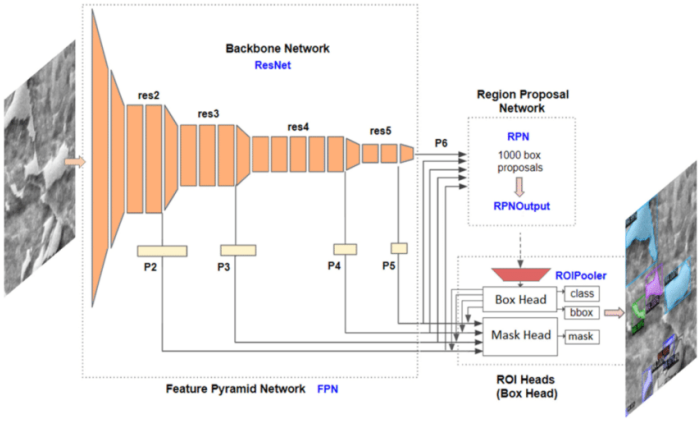


Fig. 10. DETECTRON2 model [25].

where first it performs a local prediction to separate instances even under overlapping conditions. Secondly, it carries out global prominence generation to segment instances in the complete image, pixel by pixel. Subsequently, the outputs of these two branches are assembled to form the final instance mask. It utilizes four types of backbone networks such as MobileNetV2, VoVNetV2-19, VoVNetV2-39, and ResNet-50. It is noteworthy that the accuracy of this model surpasses that of other instance segmentation methods.

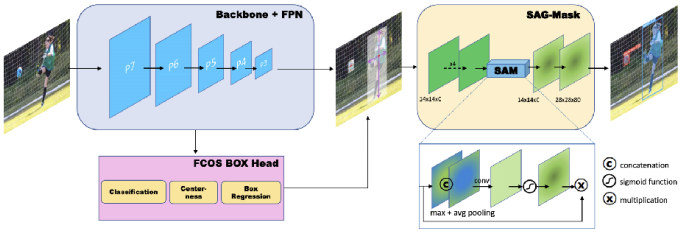


Fig. 11. CenterMask2 model [26].

IV. Experiments and Analysis

The experiments aimed to evaluate the performance of four deep learning models—SUIM-Net, YOLOv8, DETECTRON2, and CenterMask2—in scallop segmentation and multi-class detection using two custom datasets: SEG_SDS_GOPRO and SEG_SDS_SF. Each dataset represents different underwater conditions to assess model robustness. These models were selected based on their capabilities in segmentation and detection tasks in challenging environments. SUIM-Net was chosen for its specialization in underwater image segmentation, particularly in low-visibility conditions. YOLOv8 was included due to its high-speed detection and strong generalization in real-time applications. DETECTRON2, provides advanced segmentation architectures optimized for complex object detection, while CenterMask2 enhances instance segmentation with high precision in intricate visual scenes.

A. Training and Validation Scallop

The four models demonstrated varying segmentation capabilities on the two datasets. SUIM-Net performed well on SEG_SDS_GOPRO, accurately identifying scallops in challenging conditions. CenterMask2 excelled on SEG_SDS_SF, showing strong segmentation results in clearer underwater settings. YOLOv8 and DETECTRON2 provided balanced performance across both datasets, effectively handling scallop detection under diverse environmental conditions.

To provide a clear understanding of their performance, Fig. 12 presents a visual comparison of segmentation results for a representative sample across all four models. This visualization effectively highlights each model's strengths and limitations in detecting scallops, offering valuable insight into their accuracy and adaptability.

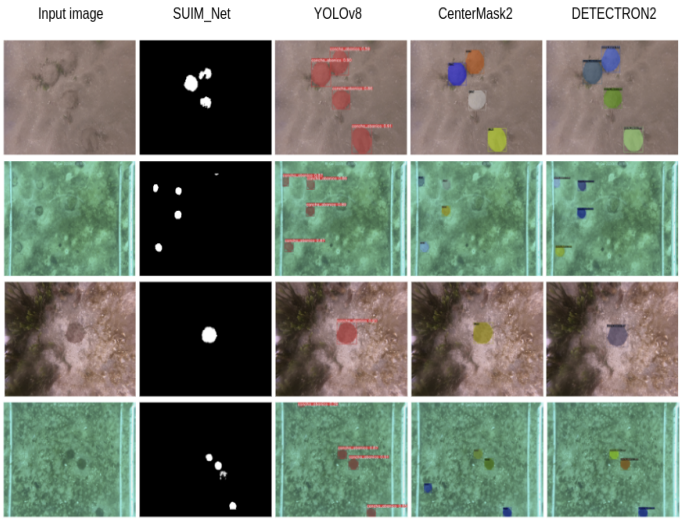


Fig. 12. Comparison of the 4 trained models.

B. Training and Validation of Six Classes

Three different models, SUIM_Net, YOLOv8 and DETECTRON2, are used for the detection of six specific marine-related classes in images. Each model employs different segmentation methodologies and detection strategies, making them suitable for different contexts and scenarios. These models have been trained to identify six marine-related classes: scallop, scallop shell, duckbill shell, snail, crab, and clam shell. Each model independently processes a set of images, predicting and delineating instances of these classes based on their unique capabilities and detection approach. Leveraging their individual strengths, the models aim to provide comprehensive and accurate identification of target objects within the images. The results obtained from each model are shown in the following images, allowing a direct comparison of their predictions. Fig. 13, 14, and 15 show the detections performed by SUIM_Net, YOLOv8, and DETECTRON2, respectively.

Fig. 13. Prediction of 6 classes using the SUIM_Net model.

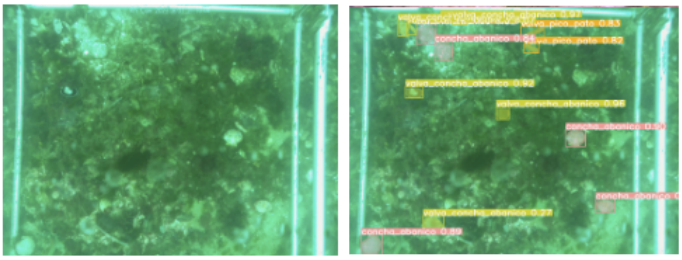


Fig. 14. Prediction of 6 classes using the YOLOv8 model.

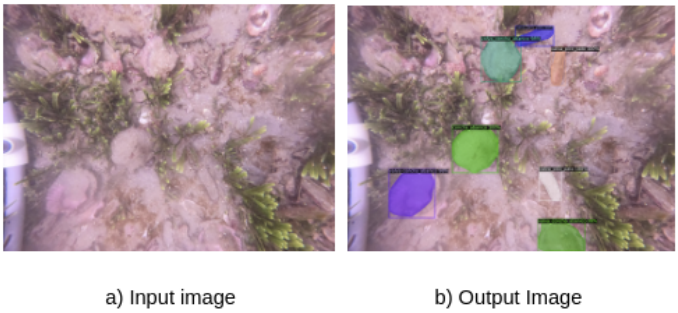


Fig. 15. Prediction of 6 classes using the DETECTRON2 model.

V. Discussion of Results

The results obtained in the SEG_SDS_GOPRO and SEG_SDS_SF datasets present significant variations due to differences in environmental conditions and characteristics of each dataset. SEG_SDS_GOPRO was captured with a GoPro Hero 9 mounted on a PVC structure, which allowed obtaining high-resolution images, although with possible optical distortions due to water refraction. In contrast, SEG_SDS_SF used a Raspberry Pi V2.1 camera integrated into a smartframe, resulting in lower-resolution images and greater variability in lighting. These differences in capture conditions directly impact the performance of the models, as they influence the accuracy of object segmentation and detection.

1) Results of the 2 datasets with the Scallop class: The SUIM-Net model was applied to the SEG_SDS_GOPRO and SEG_SDS_SF datasets for scallop segmentation, achieving accuracies of 93% and 94%, respectively. The loss plots shown in Fig. 16 illustrate the model's performance during training. In Graph (a), corresponding to the SEG_SDS_GOPRO dataset, the loss steadily decreases across epochs, with minor fluctuations, indicating consistent learning and optimization during training. Similarly, Graph (b), for the SEG_SDS_SF dataset, the loss drops sharply at first and then stabilizes, demonstrating effective learning and optimization in both cases.

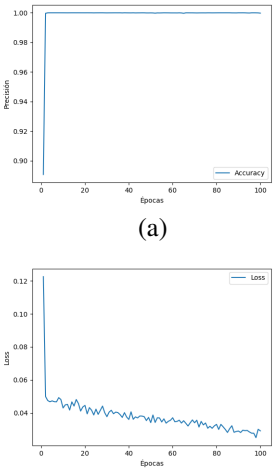


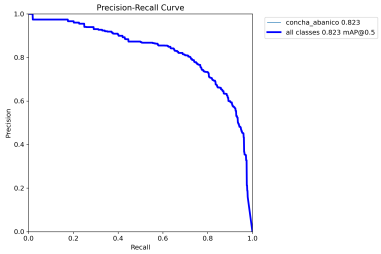
Fig. 16. Loss graphs SUIM_Net SEG_SDS_GOPRO (a) and SEG_SDS_SF (b) datasets.

While the YOLOv8 model was applied for the scallop images collected in Sechura bay. The accuracy of the a) SEG_SDS_GOPRO data set was 82%, and for the b) SEG_SDS_SF data set, its The accuracy was 95% Likewise, training was carried out with c) SEG_SDS_GOPRO + SEG_SDS_SF (Fig. 17) which obtained an accuracy of 88%. The following images refer to the experiments with each of the data sets.

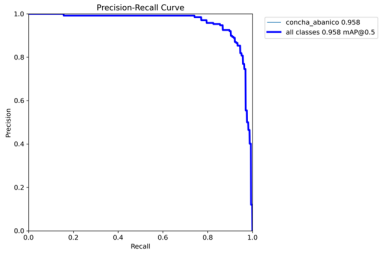
In our research, we evaluated the performance of DETECTRON2. We use different data sets: SEG_SDS_GOPRO and SEG_SDS_SF. Our goal is to understand how this model performs in different environments and imaging conditions. We start by evaluating the model trained on the (a) SEG_SDS_GOPRO dataset. He The results revealed an accuracy of 82% using DETECTRON2. This figure represents the model ability to correctly identify objects present in test images. On the other hand, when evaluating the model trained with the (b) SEG_SDS_SF data set, We see a noticeable improvement in accuracy. We achieved 96% accuracy using DETECTRON2. Likewise, the evaluation was carried out with the (c) SEG_SDS_GOPRO + SEG_SDS_SF data set, which obtained an accuracy of 80%. It is important to note that these results provide an initial view of the performance of our models with DETECTRON2.

In the following Fig. 18 you can see the loss function graphs of the datasets, where we can see that our function is decreasing which indicates that our model is efficient.

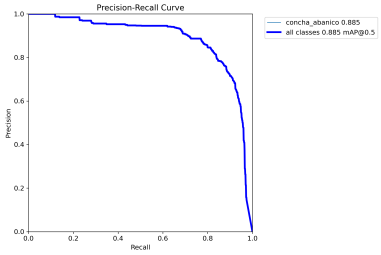
In Fig. 19, the performance of the CenterMask2 instance segmentation model was evaluated on different different data sets: SEG_SDS_GOPRO and SEG_SDS_SF. Results were visualized using average precision (mAP50) graphs, which represent the highest detection precision at 50%. The mAP50



(a)



(b)



(c)

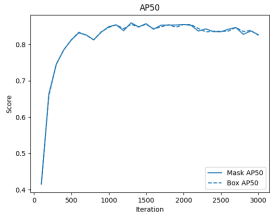
Fig. 17. Segmentation results with YOLOv8 model.

plots for the (a) SEG_SDS_GOPRO set showed an accuracy score of 81% for the CenterMask2 model. On the other hand, in the set (b) SEG_SDS_SF, CenterMask2. The model achieved a noticeably higher accuracy of 96% based on mAP50 charts. Finally, the data set of both datasets gave an accuracy of 83%.

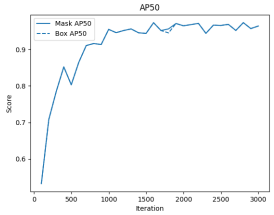
TABLE V. DATASET AND PRECISION RESULTS FOR EXPERIMENTS SCALLOP

Modelo	DataSet	F-Score	mAP
SUIM-Net	SEG_SDS_GOPRO + SEG_SDS_SF	52.45%	82.04%
	SEG_SDS_GOPRO	59.47%	93.06%
	SEG_SDS_SF	60.64%	94.05%
YOLOv8	SEG_SDS_GOPRO + SEG_SDS_SF	83.00%	88.50%
	SEG_SDS_GOPRO	76.00%	82.30%
	SEG_SDS_SF	91.00%	95.80%
DETECTRON2	SEG_SDS_GOPRO + SEG_SDS_SF	71.44%	80.84%
	SEG_SDS_GOPRO	75.62%	82.51%
	SEG_SDS_SF	85.60%	96.40%
CenterMask2	SEG_SDS_GOPRO + SEG_SDS_SF	74.46%	83.35%
	SEG_SDS_GOPRO	72.15%	81.23%
	SEG_SDS_SF	81.69 %	96.56%

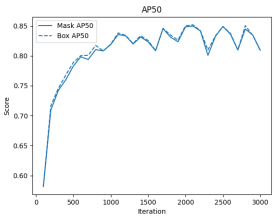
2) Results of the 2 datasets with the 6 classes: The Fig. 20 shows the precision and loss metrics of the SUIM_Net model with the SEG_SDS_GOPRO + SEG_SDS_SF dataset, which was trained with the 6 classes in the image. You can see how the precision increases, giving greater reliability to the model and the loss decreases. which indicates that the model makes more accurate predictions.



(a)

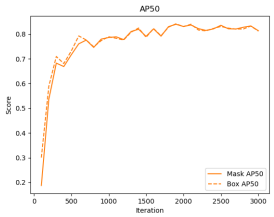


(b)

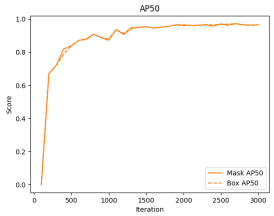


(c)

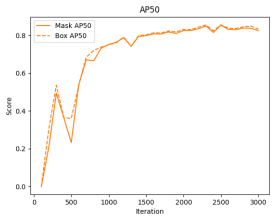
Fig. 18. Segmentation result with DETECTRON2 model.



(a)

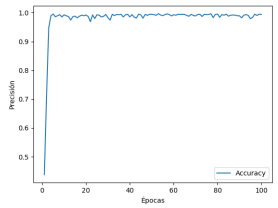


(b)

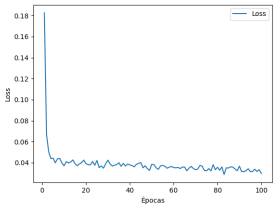


(c)

Fig. 19. Segmentation result with CenterMask2 model.



(a)



(b)

Fig. 20. Segmentation result with SUIM_Net model SEG_SDS_GOPRO + SEG_SDS_SF data set.

In Fig. 21, the performance of the model was evaluated using the Average Precision (AP) metric for six different classes. The precision results for each class are as follows: For the Crab class, the AP was 0.0, while for the Snail class it was 3.06%. The Scallop class demonstrated significantly higher accuracy, with an AP of 60.22%. Likewise, the Clam Shell class obtained an AP of 0.0, while for the Scallop Shell class it was 26.26% and for the Duck Beak Shell class it was 52.45%. These AP values provide a measure of the model's accuracy in detecting objects for each specific class, with higher values indicative of better detection.

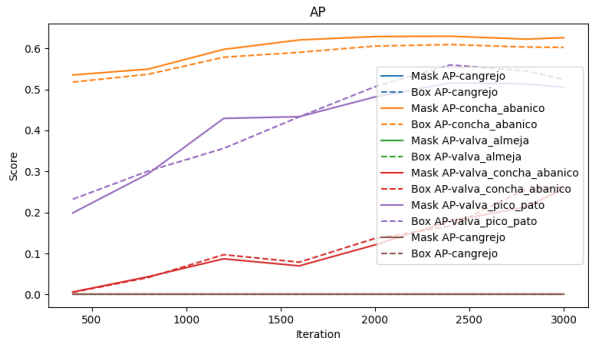


Fig. 21. Segmentation result with DETECTRON2 model SEG_SDS_GOPRO + SEG_SDS_SF data set.

In Fig. 22, the performance of the model was evaluated using the mean precision (mAP) metric for six different classes. The precision results for each class are as follows: For the Crab class, the mAP was 0.0, while for the Snail class it was 1.93%. The Scallop class demonstrated significantly higher accuracy, with a mAP of 88.50%. Likewise, the Clam Shell class obtained a mAP of 0.0, while for the Scallop Shell class it was 45.50% and for the Duck Beak Shell class it was 86.40%. These mAP values provide a measure of the model's accuracy in instance segmentation for each specific class.

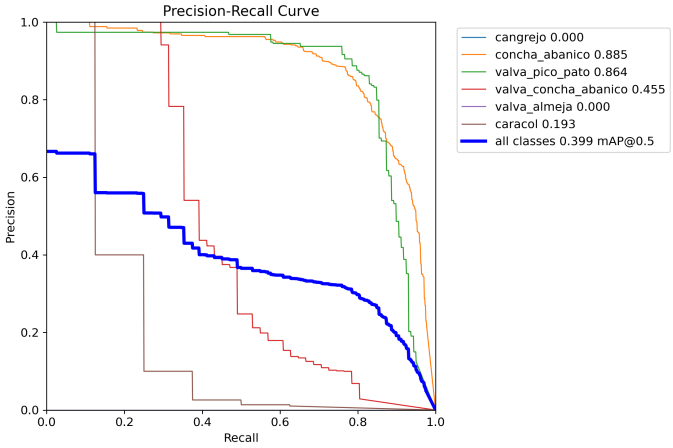


Fig. 22. Segmentation result with YOLOv8 model SEG_SDS_GOPRO + SEG_SDS_SF data set.

TABLE VI. DATASET AND PRECISION RESULTS FOR EXPERIMENTS 6 CLASSES

Modelo	DataSet	F-Score	mAP
SUIM-Net	SEG_SDS_GOPRO + SEG_SDS_SF	38.02%	35.23%
YOLOv8	SEG_SDS_GOPRO + SEG_SDS_SF	40.00%	39.90%
DETECTRON2	SEG_SDS_GOPRO + SEG_SDS_SF	37.01%	36.58%

The results obtained in our experiments reflect a high level of accuracy in the segmentation of scallops and other elements of the underwater environment. Rather than just referencing previous studies, we highlight that SUIM-Net achieved 94% accuracy on SEG_SDS_SF, while CenterMask2 reached 96.5%, which validates the effectiveness of our proposal. Compared to other works in underwater segmentation, our models demonstrated better adaptation to low visibility conditions and lighting variations, highlighting the importance of the collected dataset and the architecture of the selected models. Detailed metrics can be found in Tables V and VI.

VI. CONCLUSIONS

The SEG_SDS_GOPRO and SEG_SDS_SF datasets were designed to address the challenges of scallop segmentation in diverse underwater environments. SEG_SDS_GOPRO represents more complex conditions, including turbid waters and heterogeneous seabeds, while SEG_SDS_SF captures clearer and more structured environments. The combination of both datasets enhances model robustness by providing a balance between variability and specificity. However, the use of different camera devices (GoPro Hero 9 vs. Raspberry Pi V2.1) introduces variability in image quality, lighting conditions, and resolution, which may have impacted the model's performance.

The application of deep learning-based segmentation models demonstrated strong performance, highlighting their potential for automating aquaculture monitoring. Among the tested models, SUIM-Net achieved the highest accuracy 93% in the challenging SEG_SDS_GOPRO dataset, while CenterMask2 excelled 96.5% in the structured SEG_SDS_SF dataset, showcasing their adaptability to different underwater conditions. These results emphasize the importance of dataset-specific model selection for underwater segmentation tasks.

In addition to scallop segmentation, the models were extended to detect six additional marine classes, expanding their applicability. YOLOv8 achieved the highest accuracy 39.90%, followed closely by DETECTRON2 36.58%, while SUIM-Net had the lowest performance 35.23%. These results suggest that YOLO-based architectures, originally designed for object detection, can be effective alternatives for underwater segmentation tasks.

A key limitation was the small dataset size, constrained by logistical challenges in marine data collection and manual annotation times. Future research should integrate semi-automated labeling, data augmentation, and synthetic data generation to improve scalability.

Additionally, while this study focuses on dataset collection and segmentation performance, a more in-depth analysis of underwater image characteristics is necessary. Factors such as image noise, light distortions, and water turbidity can have a significant impact on model robustness and segmentation accuracy.

This research demonstrates the potential of deep learning models for real-time underwater aquaculture monitoring, emphasizing the importance of dataset diversity, environmental adaptability, and model selection. Further improvements in dataset size, validation methods, and robustness analysis in challenging underwater conditions will be crucial for developing more reliable AI-driven solutions for sustainable aquaculture.

ACKNOWLEDGMENT

This work was subsidized by CONCYTEC through the PROCIENCIA program within the framework of the "Special projects for the reactivation E067-2021-02" competition, according to contract or agreement 042-2021.

REFERENCES

[1] A. A. Sanchez Fernandez Baca, "The mariculture of fan shell (argopecten purpuratus) in peru and its relationship with biotrade."

[2] C. Benites Rodriguez, "The development of mariculture in peru with emphasis on fan shell (argopecten purpuratus) and shrimp (penaeus vannamei)," 1988.

[3] L. G. Yenque Moran, "Comparative analysis of the growth of the fan shell, argopecten purpuratus with respect to the suspended culture system and the bottom culture system in the company association de pescadores artesanales acuicultores chulliyachi–sechura," 2021.

[4] C. M. Salazar, R. Bandin, F. Castagnino, and B. Monteferri, "Peruvian fishing gear and methods: illustrative series," 2020.

[5] R. B. Fisher, Y.-H. Chen-Burger, D. Giordano, L. Hardman, F.-P. Lin *et al.*, *Fish4Knowledge: collecting and analyzing massive coral reef fish video data*. Springer, 2016, vol. 104.

[6] J. A. for Marine-Earth Science and Technology, "Jamstec e-library of deep-sea images," *Sensors*, 2016.

[7] J. Neira, C. Sequeiros, R. Huamani, E. Machaca, P. Fonseca, and W. Nina, "Review on unmanned underwater robotics, structure designs, materials, sensors, actuators, and navigation control," *Journal of Robotics*, vol. 2021, pp. 1–26, 2021.

[8] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE Transactions on Image Processing*, vol. 29, pp. 4376–4389, 2019.

[9] M. J. Islam, S. S. Enan, P. Luo, and J. Sattar, "Underwater image super-resolution using deep residual multipliers," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 900–906.

[10] M. J. Islam, P. Luo, and J. Sattar, "Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception," *arXiv preprint arXiv:2002.01155*, 2020.

[11] S. Mittal, S. Srivastava, and J. P. Jayanth, "A survey of deep learning techniques for underwater image classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[12] Z. Chen, Z. Zhang, F. Dai, Y. Bu, and H. Wang, "Monocular vision-based underwater object detection," *Sensors*, vol. 17, no. 8, p. 1784, 2017.

[13] D. L. Rizzini, F. Kallasi, F. Oleari, and S. Caselli, "Investigation of vision-based underwater object detection with multiple datasets," *International Journal of Advanced Robotic Systems*, vol. 12, no. 6, p. 77, 2015.

[14] F. Liu and M. Fang, "Semantic segmentation of underwater images based on improved deeplab," *Journal of Marine Science and Engineering*, vol. 8, no. 3, p. 188, 2020.

[15] F. Han, J. Yao, H. Zhu, and C. Wang, "Underwater image processing and object detection based on deep cnn method," *Journal of Sensors*, vol. 2020, 2020.

[16] S. Li, C. Li, Y. Yang, Q. Zhang, Y. Wang, and Z. Guo, "Underwater scallop recognition algorithm using improved yolov5," *Aquacultural Engineering*, vol. 98, p. 102273, 2022.

[17] H. Wang, S. Sun, X. Wu, L. Li, H. Zhang, M. Li, and P. Ren, "A yolov5 baseline for underwater object detection," in *OCEANS 2021: San Diego – Porto*, 2021, pp. 1–4.

[18] A. Imada, T. Katayama, T. Song, and T. Shimamoto, "Yolox based underwater object detection for inshore aquaculture," in *OCEANS 2022, Hampton Roads*, 2022, pp. 1–5.

[19] V. Nguyen, L. Bezanson, and B. Kinnamman, "Detection and classification of subsea objects in forward-looking sonar and electro-optical sensors for rov autonomy," in *OCEANS 2022, Hampton Roads*, 2022, pp. 1–8.

[20] M. J. Islam, C. Edge, Y. Xiao, P. Luo, M. Mehtaz, C. Morse, S. S. Enan, and J. Sattar, "Semantic segmentation of underwater imagery: Dataset and benchmark," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 1769–1776.

[21] J. Terven and D. Cordova-Esparza, "A comprehensive review of yolo: From yolov1 to yolov8 and beyond," *arXiv preprint arXiv:2304.00501*, 2023.

[22] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[23] Y. b. M. M. Contributors, "https://github.com/open-mmlab/mmyolo/tree/main/," Mayo 2023.

[24] G. M. Merz, Y. Liu, C. J. Burke, P. D. Aleo, X. Liu, M. C. Kind, V. Kindratenko, and Y. Liu, "Detection, instance segmentation, and classification for astronomical surveys with deep learning (deepdisc): Detectron2 implementation and demonstration with hyper suprime-cam data," 7 2023. [Online]. Available: http://arxiv.org/abs/2307.05826

[25] M. Ackermann, D. İren, S. Wesselmecking, D. Shetty, and U. Krupp, "Automated segmentation of martensite-austenite islands in bainitic steel," *Materials Characterization*, vol. 191, p. 112091, 07 2022.

[26] Y. Lee and J. Park, "Centermask : Real-time anchor-free instance segmentation," 2020.

Performance Optimization with Span<T> and Memory<T> in C# When Handling HTTP Requests: Real-World Examples and Approaches

Daniel Damyanov, Ivaylo Donchev
Department of Information Technologies, Veliko Tarnovo University, Bulgaria

***Abstract*—Optimization of application performance is a critical aspect of software development, especially when dealing with high-throughput operations such as handling HTTP requests. In modern C#, structures Span<T> and Memory<T> provide powerful tools for working with memory more efficiently, reducing heap allocations, and improving overall performance. This paper explores the practical applications of Span<T> and Memory<T> in the context of optimizing HTTP request processing. Real-world examples and approaches that demonstrate how these types can minimize memory fragmentation are presented, avoid unnecessary data copying, and enable high-performance parsing and transformation of HTTP request data. By leveraging these advanced memory structures, developers can significantly enhance the throughput and responsiveness of their applications, particularly in resource-constrained environments or systems handling many concurrent requests. This paper aims to provide developers with actionable insights and strategies for integrating these techniques into their .NET applications for improved performance.**

***Keywords*—NET; C#; optimization; memory optimization; span; memory; development; HTTP requests; data structures**

I. Introduction

As software applications evolve and network communications become increasingly complex, code efficiency and performance are becoming key. In .NET, types such as Span<T> [1] and Memory<T> [2] provide powerful memory management and data processing tools that can play an essential role in optimizing performance when handling HTTP requests [3]. Efficient memory management is a critical concern in modern high-performance applications, especially when dealing with high volumes of HTTP requests and responses. In scenarios such as web servers, API gateways, and microservices, where throughput and responsiveness are paramount, excessive memory allocations and garbage collection (GC) [4] overhead can severely impact performance and scalability. Traditional approaches to handling HTTP data, like using StreamReader or working directly with strings and arrays, often lead to unnecessary memory copying and fragmentation, which increases the load on the garbage collector (GC) and reduces the overall efficiency of the system. To address these issues, modern versions of C# introduced Span<T> and ReadOnlySpan<T> – lightweight, stack-allocated types designed to operate directly on memory without allocations. These types allow developers to access and manipulate slices of memory, buffers or arrays without creating new objects, thereby minimizing memory allocations and avoiding costly GC operations. Span<T> enables mutable operations on memory, while ReadOnlySpan<T> provides safe, immutable access, making them ideal for handling both mutable and immutable HTTP request and response data. This paper explores how these advanced memory constructs can be applied in real-world HTTP request handling scenarios to optimize performance. We demonstrate practical examples of using Span<T> and ReadOnlySpan<T> to efficiently parse, modify, and process HTTP request and response data, highlighting the performance improvements gained from reduced memory usage and minimized garbage collection. By leveraging these tools, developers can significantly improve the throughput and scalability of their .NET applications, particularly in resource-constrained or high-load environments.

The following sections are organized as follows:

Section II comments on typical use cases of the Span and Memory structures. It points out their advantages, proven by test results. Emphasis is placed on the optimization of HTTP requests, sorting algorithms and parallel data processing.

Section III summarizes the comparison of the different memory management techniques in C#.

Section IV discusses the results of similar studies; focuses on the advantages of new memory management structures, including reducing memory allocation and optimizing GC operations. Attention is paid to trade-offs and limitations related to the use of Span and Memory, practical implications for HTTP request handling and similar features in other programming languages.

The conclusion motivates once again developers to use the new memory management structures to achieve higher performance of their applications.

II. Common Use Cases of Span and Memory in Applications and Real Tests

When working with large text data, like reading content from files or HTTP responses, the traditional approach involves loading all the content into memory as a string. This can lead to a significant memory load, especially when the data is large. One of the common problems when dealing with HTTP requests is the incorrect handling of large JSON responses [5]. When receiving a large JSON response from a server, trying to desearilize the entire response at once can lead to memory and performance problems.

The following example shows one such problematic JSON response that is handled incorrectly, and this leads to a large memory load.

One common problem when dealing with data from HTTP requests is the inefficient handling of large text data, which can lead to unnecessary data copying and increased memory usage.

Span<T> can help optimize data processing by allowing us to work directly on arrays of bytes without copying. Span<T> comprises just two fields, a pointer and a length. For this reason, it can represent only contiguous blocks of memory [6]. Span by nature is mutable and allows for modifications to the underlying data. This example shows how a large text response is handled inefficiently, resulting in high memory overhead and latency.

```
HttpResponseMessage response = await _httpClient.GetAsync(url);
response.EnsureSuccessStatusCode();
string responseText = await response.Content.ReadAsStringAsync();
string firstWord = responseText.Split(' ')[0];
Console.WriteLine($"First word: {firstWord}");
```

Fig. 1. Getting a response and taking the result from the null index.

The problem with the code above (Fig. 1) is the large and inefficient use of memory: the entire textual response is held in memory as a string. Therefore, unnecessary data copying is also performed. The Split() method creates new arrays of strings. To avoid these problems, Span<T> can be used to work directly on the byte arrays and to extract the necessary information efficiently. Using Span is only synchronous, and it should be noted that implementation in asynchronous methods requires additional code writing, since using synchronous methods in asynchronous ones would lead to unpredictable results or thread blocking.

Avoiding asynchronous disadvantages, another fast data processing can be used and the use of asynchronous operations without accompanying difficulties that can be achieved using Memory<T>, which allows working with subsets of data without additional copying.

Processing large text data from HTTP responses can be inefficient if the data is behaved like strings. Memory<T> can help optimize this process by working directly with the byte arrays. In .NET, Memory<T> and ReadOnlyMemory<T> provide powerful memory tools that enable efficient data processing without unnecessary copying and reduce memory load. These structures can be used in both synchronous and asynchronous methods (Fig. 2), making them suitable for a wide range of applications.

```
1 reference
public async Task ProcessHttpResponseAsync(string url)
{
    try
    {
        HttpResponseMessage response = await _httpClient.GetAsync(url);
        response.EnsureSuccessStatusCode();
        byte[] responseBytes = await response.Content.ReadAsByteArrayAsync();
        Memory<byte> memory = new Memory<byte>(responseBytes);
        ProcessResponseData(memory);
    }
    catch (HttpRequestException e)
    {
        Console.WriteLine($"Request error: {e.Message}");
    }
}

1 reference
private void ProcessResponseData(Memory<byte> memory)
{
    ReadOnlySpan<byte> span = memory.Span;
    int spaceIndex = span.IndexOf((byte)' ');
    ReadOnlySpan<byte> firstWordSpan = spaceIndex == -1 ? span : span.Slice(0, spaceIndex);
    string firstWord = Encoding.UTF8.GetString(firstWordSpan);
    Console.WriteLine($"First word: {firstWord}");
}
```

Fig. 2. Use of memory in http requests.

A. *Benefits of Memory<T> and ReadOnlyMemory<T>*

- Efficient memory usage: Memory<T> and ReadOnlyMemory<T> allows processing of parts of arrays without creating new objects, which reduces memory load.
- Flexibility: They can be used in both synchronous and asynchronous methods, providing a safe way to work with data between await points.
- Improved data processing code: By using Memory<T>, efficiency of applications can be improved when dealing with large amounts of data, such as text responses from HTTP requests or binaries.

B. *Test Performance When Reading Query Data*

When working with large files, it is often necessary to read only part of the content to avoid unnecessary memory load. Big data can fill up a system's RAM, leading to delays or crashes (OutOfMemoryException) [7]. Also, big data leads to slow processing speed [8] – a significant amount of time to reduce the efficiency of applications. Big data can contain unstructured or poorly formatted parts, which can complicate its processing.

Using Memory<T> allows working with parts of the data without loading the entire file into memory. Reading a specific block of the file and using Memory<T> to work with that block provides efficiency and flexibility (Fig. 3).

```
HttpResponseMessage response = await _httpClient.GetAsync(url);
response.EnsureSuccessStatusCode();

byte[] responseBytes = await response.Content.ReadAsByteArrayAsync();
Memory<byte> memory = new Memory<byte>(responseBytes);
```

Fig. 3. Reading data with Memory<T>

The following is a comparative analysis between the use of Memory<T> and the standard implementation with StreamReader. Two methods will be implemented to read the response from the query (Fig. 4, Fig. 5).

```
async Task ReadLargeHttpResponseWithStream(string url)
{
    HttpResponseMessage response = await _httpClient.GetAsync(url);
    response.EnsureSuccessStatusCode();

    using (Stream responseStream = await response.Content.ReadAsStreamAsync())
    {
        using (StreamReader reader = new StreamReader(responseStream))
        {
            char[] buffer = new char[8192];
            int bytesRead;
            while ((bytesRead = await reader.ReadAsync(buffer, 0, buffer.Length)) > 0)
            {

                ProcessLargeData(buffer.AsMemory(0, bytesRead));
            }
        }
    }
}
1 reference
void ProcessLargeData(ReadOnlyMemory<char> data)
{
    string content = new string(data.Span);
    Console.WriteLine(content);
}
```

Fig. 4. Reading data with StreamReader.

```
public async Task ReadLargeHttpResponseWithMemory(string url)
{
    HttpResponseMessage response = await _httpClient.GetAsync(url);
    response.EnsureSuccessStatusCode();

    byte[] responseBytes = await response.Content.ReadAsByteArrayAsync();
    Memory<byte> memory = new Memory<byte>(responseBytes);

    ProcessLargeData(memory);
}
1 reference
private void ProcessLargeData(Memory<byte> data)
{
    string content = System.Text.Encoding.UTF8.GetString(data.Span);
}
```

Fig. 5. Reading data with Memory<T>.

HttpClient and URL were used for the tests. The URL points to a large file (100 MB) which can be downloaded for the tests. Running the benchmark leads to the following results (Fig. 6):

Method	Mean	Error	StdDev
ReadLargeHttpResponseWithMemory	520.4 ms	8.48 ms	7.94 ms
ReadLargeHttpResponseWithStream	680.3 ms	12.6 ms	11.8 ms

Fig. 6. Test results.

- ReadLargeHttpResponseWithMemory(): This method uses Memory<byte> and shows an average execution time of 520.4 ms. Using Memory<T> provides advantages in terms of efficient memory management and performance.
- ReadLargeHttpResponseWithStream(): This method uses a traditional approach with Stream and shows an average execution time of 680.3 ms. The traditional method is slower due to the additional time it takes to work with StreamReader and convert data.

Several of the most important advantages of using relevant optimizations can be noted:

- Efficient Memory Management: Working with Memory<byte> reduces the need to create redundant copies of data, which saves memory and resources.
- Flexibility: Memory<T> allows for easy retrieval and handling of partial blocks of data, which is useful when working with large files or data streams.
- Better performance: Reading only the required portion of the file and working with Memory<T> can improve application performance by reducing processing time and memory usage.

C. Other applications

1) Search algorithm optimization: Algorithms for searching large arrays or text data often require high performance and low memory usage. Traditional approaches may involve creating new copies of the data, which increases memory costs and slows down processing. Span<T> allows working with pieces of data directly, without creating new copies, which is useful in search algorithm optimization. The following test will be conducted. A version of the familiar binary search algorithm [9] is implemented but using the Span structure. It must be compared to a method representing the basic implementation. Finally, the test is run again.

Fig. 8 shows the only difference is that the lookup data is sent with a Span. In the benchmark made, in Fig. 7, 1 million elements were feeded.

Method	Mean	Error	StdDev
StandardBinarySearch	0.120 ms	0.001 ms	0.002 ms
SpanBinarySearch	0.115 ms	0.002 ms	0.003 ms

Fig. 7. Binary search and span.

```
static int BinarySearchWithSpan(Span<int> data, int target)
{
    int low = 0;
    int high = data.Length - 1;

    while (low <= high)
    {
        int mid = low + (high - low) / 2;
        if (data[mid] == target)
            return mid;
        if (data[mid] < target)
            low = mid + 1;
        else
            high = mid - 1;
    }
    return -1;
}
```

Fig. 8. Binary search with a span.

It is noticeable that the difference is small, but it will be tested whether Span<T> offers any advantages in the context of specific use cases. One of the advantages is that ready-made methods of the structure itself can be used, and additional functionality can be added without wasting time on it.

2) Parallel data processing: Parallel data processing can lead to high memory costs if each thread creates its own copies of the data. This can reduce the efficiency of the application and increase processing time. Parallel processing involves dividing a task into smaller subtasks that can be run simultaneously [10]. C# has various tools for parallelization., such as Task, Parallel.For, and async/await. The following test will be performed. 10 million characters simulating text will be sent to be processed by the program. Two variants have been developed: with string and with Memory (Fig. 9 and Fig. 11).

In the test done, better performance is again visible, albeit with a small lead in time (Fig. 10).

```
public async Task ParallelProcessingWithMemory()
{
    var tasks = new Task[_blockCount];

    for (int blockIndex = 0; blockIndex < _blockCount; blockIndex++)
    {
        int start = blockIndex * _blockSize;
        int end = start + _blockSize;
        var blockMemory = _textMemory.Slice(start, _blockSize);

        tasks[blockIndex] = Task.Run(() =>
        {
            ProcessTextBlock(blockMemory.Span);
        });
    }

    await Task.WhenAll(tasks);
}
```

Fig. 9. Parallel processing with memory

Method	Mean	Error	StdDev
ParallelProcessingWithMemory	100.000 ms	2.000 ms	3.000 ms
ParallelProcessingWithString	150.000 ms	3.000 ms	4.000 ms

Fig. 10. Task and memory test result

The use of Memory<T> allows for the safe sharing of data between different threads, without creating redundant copies. Processing with Memory<T> will be faster because Memory<T> reduces the cost of creating new copies of the data and provides more efficient access to memory.

```
0 references
public async Task ParallelProcessingWithString()
{
    var tasks = new Task[_blockCount];

    for (int blockIndex = 0; blockIndex < _blockCount; blockIndex++)
    {
        int start = blockIndex * _blockSize;
        int end = start + _blockSize;
        string blockString = _largeText.Substring(start, _blockSize);

        tasks[blockIndex] = Task.Run(() =>
        {
            ProcessTextBlock(blockString);
        });
    }

    await Task.WhenAll(tasks);
}
```

Fig. 11. Parallel processing with string

III. Results

The results of comparing different memory management techniques in C# - ReadOnlyMemory<T>, ReadOnlySpan<T>, Memory<T>, and Span<T> - demonstrates clear advantages in performance optimization for HTTP request handling. Key findings are as follows:

- ReadOnlyMemory<T>: Offers efficient read-only data access with reduced memory allocations and garbage collection pressure. It is well-suited for handling immutable HTTP request data like headers and bodies, providing safety with minimal performance overhead.
- ReadOnlySpan<T>: Provides the most efficient, zero-allocation solution by leveraging stack-based memory management. It excels in short-lived, high-performance scenarios, such as parsing HTTP requests without persisting data. However, it is limited by its inability to be used across asynchronous boundaries.
- Memory<T>: Allows mutable access to memory, making it versatile for scenarios where data needs to be modified. Its ability to reuse memory buffers reduces garbage collection events and enhances performance in systems with high-throughput HTTP requests.
- Span<T>: Similar to Memory<T>, but focused on stack-based memory, making it ideal for fast, non-persistent data processing. It shares the limitations of ReadOnlySpan<T> in async contexts but provides excellent performance when dealing with mutable, transient data.

Overall, these advanced memory manipulation types significantly reduce memory allocations, improve execution time, and lower GC pressure compared to traditional approaches like StreamReader. The most substantial performance gains are observed when using Span<T> and ReadOnlySpan<T>, which minimize overhead through zero-copy, stack-based memory operations.

IV. Discussion

The special language features in C# and other modern languages that allow more efficient management of sequentially located data structures in memory are relatively new, and there is not much scientific research related to them.

The study [11] is focused on how Span is designed to optimize memory usage and enhance processing speed in .NET applications, with an emphasis on the characteristic collections, which generally store the data in the heap memory, which consumes more RAM and increases the workload of the Garbage Collector. One of the strongest features of Span is to keep data on the stack which enables better performance.

The study [12] presents an in-depth comparison of the use of different algorithms on the traditional List, Array, etc. collections. Experimentation and analysis reveal the different performance of these algorithms using C#.

A comparison of the effectiveness of different strategies related to the optimization of memory management procedures and in particular the release of resources implemented in the C#, Java and C++ languages is discussed in [13]. The surprising conclusion is made that C#'s garbage collection system consistently outperformed the others due to its optimized procedures of asynchronously deallocating memory.

Notorious key techniques in Memory Optimization include avoiding unnecessary object allocations; use value types for small, immutable data; pool reusable objects with ObjectPool<T>; optimize collections (e.g. prefer array or Span<T> over List<T> when possible); avoid large object heap fragmentation by reusing buffers; utilize asynchronous programming effectively to reduce memory pressure.

The test results clearly demonstrate that leveraging modern memory types in C#, such as ReadOnlyMemory<T>, ReadOnlySpan<T>, Memory<T>, and Span<T>, provides significant performance improvements in scenarios involving HTTP request handling. The implications of these findings are particularly relevant for applications that deal with high

volumes of HTTP traffic, where memory allocations and GC overhead can become major bottlenecks.

D. Memory Efficiency and Allocation Reduction

One of the key benefits of using these memory types is the reduction in memory allocations. Traditional methods, such as reading data with StreamReader, create new strings and objects on the heap, which leads to frequent memory allocations. This not only consumes more memory but also increases the frequency of GC cycles, leading to performance degradation.

By contrast, Memory<T> and Span<T> significantly reduce the need for heap allocations by reusing memory buffers or leveraging stack-based memory management. This is particularly important in real-time systems or services that handle numerous HTTP requests, as it helps maintain consistent performance under heavy load. The zero-allocation nature of ReadOnlySpan<T> and Span<T>, in particular, shows tremendous potential for short-lived operations where both memory and speed are critical.

E. Garbage Collection Optimization

Reducing GC pressure is a critical factor in achieving high-performance applications, especially in scenarios involving concurrent HTTP requests. The frequent creation and destruction of objects in heap-based memory can result in excessive GC activity, leading to increased latency and jitter. The findings highlight that Memory<T> and Span<T>, by reducing object creation, lead to fewer GC interruptions and smoother application performance.

In particular, ReadOnlyMemory<T> and ReadOnlySpan<T> proved highly efficient for handling immutable data like HTTP headers or request bodies, where copying or modifying data is unnecessary. These types allow direct access to memory without the need for costly allocations, which lowers GC frequency and minimizes its impact on performance.

F. Trade-offs and Limitations

While these memory types offer clear performance gains, there are trade-offs that developers need to consider. For instance, Span<T> and ReadOnlySpan<T> are stack-allocated and, therefore, cannot be used across asynchronous method calls. This limits their applicability in scenarios where asynchronous programming is heavily used, such as modern HTTP request pipelines built on async/await patterns. In these cases, developers must rely on heap-based Memory<T> or ReadOnlyMemory<T>, which still provide performance benefits but with slightly higher overhead compared to their stack-based counterparts.

Additionally, while these types reduce memory allocations and improve performance, they introduce added complexity in managing memory manually. Developers must be more mindful of buffer management and ensure that memory is handled properly to avoid issues such as memory leaks or unsafe access. This represents a trade-off between performance and ease of use, particularly for teams or projects where rapid development is prioritized over fine-tuned optimization.

G. Practical Implications for HTTP Request Handling

For real-world applications that handle HTTP requests, especially high-throughput services such as web servers, API gateways, or microservices, the use of Memory<T> and Span<T> can lead to substantial performance improvements. The ability to avoid data copying, minimize memory fragmentation, and reduce GC pressure can help these systems scale more effectively, handling larger workloads with lower resource consumption.

However, these performance benefits come with the requirement for a deeper understanding of memory management in .NET. Developers will need to weigh the trade-offs between the additional complexity and the performance gains, particularly when deciding between the simplicity of StreamReader and the efficiency of the newer memory types.

H. Future Considerations

As C# continues to evolve, further optimizations and tools that will make memory management both easier and more efficient can be expected. The results suggest that adopting these newer memory types now can provide immediate benefits in terms of performance, but future improvements in language features, runtime optimizations, and libraries may help bridge the gap between ease of use and high performance. Additionally, as more developers adopt these patterns, best practices will likely emerge, helping to mitigate the challenges associated with the manual memory management required by Span<T> and Memory<T>.

I. Similar Features in Other Programming Languages

Languages like C++, Rust, D, and Swift have similar features. C++ has std::span which is a very lightweight abstraction (a class template), but powerful tool for working with contiguous sequence of data. A typical implementation holds a pointer to the data, if the extent is dynamic, the implementation also holds a size. The main advantage is that it is a non-owning type (a reference-type rather than a value type). It never allocates nor deallocates anything and does not keep smart pointers alive.

```
void show(std::span<int> data)
{
    for (const auto& x : data)
        std::cout << x << ' ';
    std::cout << '\n';
}

int main()
{
    int myArray[] = { 1,2,3,4,5 };
    show(myArray);
}
```

Fig. 12. Using std::span as function parameter.

In C++, std::span can also be used to simplify syntax. For example, the implementation of arrays in C++ is such that they "don't know" their size. When an array is passed as an

argument to a function, actually a constant pointer to the first element of the array is passed and the function "does not know" the number of elements in the array. However, if this array is passed to the function as span, there is no need to pass an additional parameter with the size of the array (Fig. 12).

Instead of Span, Rust has a slice, which is a view into a collection of elements, like std::span. D language offers slices as well, providing a safe way to handle arrays. Swift has ArraySlice, which allows sub-ranges of an array without copying. Each of these languages emphasizes safety and efficiency in handling contiguous data sequences.

Java doesn't have a direct equivalent of Span, but List.subList() allows to work with a sublist view of a list. Although it doesn't offer the same level of low-level control, it does provide a way to handle segments of collections efficiently.

V. Conclusion

Both Span<T> and Memory<T> provide powerful ways to work with memory in C#, especially when performance and efficiency are critical. Overall, our findings confirm that using ReadOnlyMemory<T>, ReadOnlySpan<T>, Memory<T>, and Span<T> provide significant improvements in performance, particularly for HTTP request processing. While there are trade-offs in terms of complexity and applicability in asynchronous programming, these memory types offer a powerful way to optimize memory usage, reduce GC overhead, and improve the scalability of modern .NET applications.

Developers aiming for high-performance HTTP request handling should seriously consider integrating these memory types into their systems to achieve better throughput and responsiveness.

Other practical applications of Span<T> include image processing or numerical computations – situations when working with performance-critical code. In this case, Span<T> can help avoid allocations and reduce the overhead of garbage collection. Similar is the situation when interaction with native code via P/Invoke or other interop mechanisms is needed. Here Span<T> can be used to represent contiguous memory regions efficiently.

Wherever working with asynchronous I/O operations, data buffers or lazy initialization are needed, Memory<T> is a great solution.

References

[1] Microsoft .NET documentation, Microsoft Learn, Span<T> Struct, online: https://learn.microsoft.com/en-us/dotnet/api/system.span-1?view=net-9.0

[2] Microsoft .NET documentation, Microsoft Learn, Memory<T> Struct, online: https://learn.microsoft.com/en-us/dotnet/api/system.memory-1?view=net-9.0

[3] Smith, J., Build your own web server from scratch in Node.JS: Learn network programming, HTTP, and WebSocket by coding a web server (Build Your Own X From Scratch), Independently published, 2024

[4] Jones, R., Hosking, A., Moss, E., The garbage collection handbook: The art of automatic memory management, Chapman and Hall/CRC; 2nd edition, 2023

[5] Price, M., C# 9 and .NET 5 – Modern cross-platform development: Build intelligent apps, websites, and services with Blazor, ASP.NET Core, and Entity Framework Core using Visual Studio Code, Packt Publishing; 5th edition, 2020

[6] Albahari, J., C# 12 in a Nutshell: The Definitive Reference, O'Reilly Media; 1st edition, 2023

[7] Kokosa, K., Nasarre, Chr., Gosse, K., Pro .NET memory management: For better code, performance, and scalability, Apress; Second edition, 2024

[8] Rasmussen, B. (2014), High-performance Windows Store apps (developer reference), Microsoft Press; 1st edition, 2014

[9] Cormen, T., Leiserson, Ch., Rivest, R., Stein, C., Introduction to Algorithms, The MIT Press; 4th edition, 2022

[10] Sarcar, V., Parallel programming with C# and .NET: Fundamentals of concurrency and asynchrony behind fast-paced applications, Apress, 2024

[11] Akdoğan, H., Duymaz, H., Kocakır, N., Karademir, Ö., Performance analysis of Span data type in C# programming language. Turkish Journal of Nature and Science. October 2024; Issue 1, pp. 29-36. doi:10.46810/tdfd.1425662

[12] Shastri, S., Singh, A., Mohan, B., Mansotra, V., Run-time analysis of searching and hashing algorithms with C#, 2016, Available from: https://www.researchgate.net/publication/326331475_Run-Time_Analysis_of_Searching_and_Hashing_Algorithms_with_C

[13] Henriques, L., Bernardino, J., Performance of memory deallocation in C++, C# and Java, CAPSI 2018 Proceedings. 10., https://aisel.aisnet.org/capsi2018/10

A Systematic Review of the Benefits and Challenges of Data Analytics in Organizational Decision Making

Juan Carlos Morales-Arevalo, Ciro Rodríguez
Faculty of Systems Engineering and Computer Science, Universidad Nacional Mayor de San Marcos, Lima - Perú

***Abstract*—Data analytics has been relied heavily in organizational decision-making, which allows accuracy, timeliness, and data-driven processes in a wide range of industries. These factors are influential as the figure and complexity of data are on the rise, along with problems like authentication, integration, and organizational resistance. The current study seeks to systematically review the benefits and challenges of data analytics on decision-making in different sectors using the PRISMA guidelines. A total of 32 articles published from 2020 until 2024 were identified through this review from reputable databases, including Scopus, Web of Science, IEEE Xplore, ProQuest, and Emerald Insight. These insights underscore the power of data analytics in driving change, enabling more accurate, faster, and aligned decision-making with organizational objectives. Challenges remain though, including the availability of broken data systems, hindrance due to a non-standardized norm across the whole sector, and resistance in places where data literacy is low or cultures resist data-driven practices. To mitigate the challenges, this review offers organizations practical recommendations for management. Companies that successfully incorporate analytics into their overall business strategies and create an organization-wide value for data and insights will be able to leverage analytics more effectively to enhance efficiency, encourage innovative growth, and navigate future disruptions. However, tackling these challenges is more than just optimizing performance—it is about future-proofing organizations in a world increasingly defined by data.**

***Keywords*—Data analytics; decision-making; data-driven processes; big data analytics; systematic review**

I. Introduction

Implementation of this method helps organizations in smart and accurate decision making which is an integral process of the global world today. This declaration comes atop a growing volume of data available today (and growing complexity as well), which presents both opportunity and challenges. Enabling this have been advances in technology such as machine learning and cloud computing that can analyze large amounts of data in real-time [1]. In particular, the use of data analytics in areas such as supply chain management, medical, and marketing is changing the way organizations can improve their actions and forecast future procedures in their working environment, based on [2], the analysis of large-scale data is the basis of customizing approaches, enabling organizations to adapt their services and products to the characteristics of each customer positively impacting the productivity and retention of customers. In addition, [3] noted that data analytics provides advantageous sustainable competitive advantages leading to innovation in emerging markets, thus directly enhancing strategic decision-making and all the different strategies that can be made. Moreover, [4] pointed out that big data analytics considerably enhance real-time decision-making, specifically in healthcare supply chains, by enabling efficient operations management and overcoming essential implementation hurdles.

In this sense, the identification of the main challenges in the field of data analytics for organizational decision-making highlights significant challenges related to data management and quality, as well as to the effective implementation of analytical methodologies. [5] Among these challenges, the integration of advanced techniques such as machine learning into traditional systems is especially problematic, due to the lack of clarity in objectives and limitations in organizational maturity for ad hoc projects. Instead [6], current approaches are focused on technology rather than socio-technical aspects, creating a disconnect between these technologies and their strategic and operational implementation. To cope with these challenges, methodologies focusing on end-to-end team, project, and data management have been introduced, a more holistic approach to data science projects. In organizational contexts [7], the successful integration of business analytics is contingent on a blend of organizational, technological, and environmental variables, which serve as driving forces behind its success.

This review provides strategic insights into leveraging data analytics to overcome barriers and optimize decision-making processes; by addressing these challenges with holistic approaches, organizations can better align technical tools with operational strategies, ultimately enhancing performance and driving meaningful innovation.

This study analyzes the effects of being data-driven on organizations, listing the benefits and challenges it presents as well as its strategic importance in varied industries. Furthermore, it offers actionable insights to assist organizations in effectively adopting analytics and deriving maximum value from it in their decision-making processes [8].

The rest of the paper is structured as follows: Section II discusses the related work. Section III explains the methods of carrying out this systematic review, including inclusion and exclusion criteria and sources of data. Results are described in Section IV, organized around major themes. An extended discussion of the implications, challenges, and opportunities arising from data analytics in organizations is covered in Section V. Finally, Section VI provides the conclusions and Section VII directions for future work.

II. Related Work

Data analytics is increasingly critical for organizations, as it helps organizations make more informed decisions based on data. Prior research addressed the adoption of data analytics, benefits of data analytics, and challenges of data analytics per industry.

Multiple scholars have studied how big data analytics improves decision making and business intelligence. Patricio-Peralta et al. discussed the role of big data in shaping marketing approaches to enhance customer targeting and engagement using predictive analytics techniques [2]. Likewise, Al Nuaimi & Awofeso emphasized the vital significance of big data for healthcare supply chain management while also adding that it greatly supplements healthcare optimization leading to operational efficiency and be the followed standard [4]. Their studies are consistent with Rahman's systematic review on empowering business intelligence in healthcare capabilities, which show that data-driven decision-making improves resource allocation as well as patients' outcomes [9].

The quality and maturity of data is another important area of analytics adoption research. Galetsi et al. highlighted the challenge of data fragmentation and lack of interoperability as a key challenge that many organizations face to this day, limiting the reliability of insights derived from analytics [1]. Likewise, Al-Sai et al. presented a Big Data Maturity Model, emphasizing that the readiness of the organization for Big Data needs to be evaluated before integrating analytics into the decision-making process of the organization completely [5]. Hence, organizations focused on utilizing analytics must follow a stringent strategy for managing the data.

There have been several studies that also covered the obstacles to not adopting data analytics in organizations. Horani et al. highlights cultural and leadership challenges as major barriers, also mentioning that the lack of executive driving force can slow down analytics adoption [7]. Gonzales & Horita expanded on this by noting that bad visualization/analytics tools can mitigate against user engagement, thus potentially impacting efficacy in practical settings [10]. They also highlighted that a common factor across many sectors such as education and research, where streamlined data-driven strategies could benefit the decision-making process, is insufficient training in analytics tools [11].

Researchers also explored the contributions of advanced analytics and optimization techniques in decision making. Integrating these BI tools with strategic decision-making helps organizations make informed strategic choices with up-to-date insight [12]. Similarly, Akindote et al. demonstrated the use of geographic information systems (GIS) and analytics to improve decision-making in spatial planning and logistics [13]. Khanra et al. As an example toward the healthcare industry, demonstrating how analytics-oriented techniques can enhance patient care and reduce operational costs [14].

Similarly, implementation of big data analytics in various industries has also been extensively explored. Mansour & Bick delivered a systematic review of big data platforms employed in the healthcare digital transformation context and how artificial intelligence (AI)-based analytics have transformed diagnostics and treatment planning [15]. Tawil et al. analyzed the ways small and medium-sized enterprises (SMEs) in emerging economies utilize data analytics for a competitive edge, emphasizing opportunities and barriers in these contexts [16].

Therefore, while much academic research has contributed to the advances in data analytics, there are still areas that lack the necessary focus. Numerous studies indicate that although analytics adoption is growing, data standardization, integration, and culture acceptance still face challenges. The objective of this systematic review is to compile these insights and offer a complete evaluation of the advantages and disadvantages of data-driven decision-making in organizations.

III. Methodology

PRISMA is selected for this study to conduct a systematic review. This offers a solid framework to analyze a large amount of data and evaluate their significance in organizations' decision-making processes [17]. The most up to date version, PRISMA 2020, brings important new changes which improve the transparency of systematic review and encourage reporting by all including with advances in methodology and terminology [9]. In certain domains, like strategic management, PRISMA is key to spotting gaps in the use of information within public organizations and small businesses, as it is a continuous improvement framework [18]. Additionally, this method has been useful in the industrial domain, to augment decision models and production systems with advanced technologies, such as AI and deep learning [19]. Moreover, PRISMA has been applied in research related to the digital transformation of bibliometric and systematic studies which have sought to resolve challenges associated with data recreation and meta-analysis [20].

A. Identification

The identification phase focuses on the exhaustive search for articles relevant to the topic. Using recognized databases and specific keywords, potential publications are extracted for the review. According to Page et al. (2021), this phase establishes a systematic framework to ensure that relevant studies are included and that initial bias in source selection is minimized [17]. A systematic search for research data was performed in various known databases, including Scopus, Web of Science, IEEE Xplore, ProQuest, and Emerald Insight as illustrated in Fig. 1. The reason for selecting these databases is highly related to the data analytics applications in the organizational decision-making domain. To get sufficient coverage, certain keywords, along with Boolean operators, were utilized, e.g., "Data Analytics" OR "Big Data Analytics" AND "Decision Making" AND "Systematic Review", Although the study primarily focused on organizational and business contexts, some research from the healthcare field was included. This decision was made because their findings provided valuable insights that aligned with the broader goal of understanding how data analytics can enhance decision-making processes across various sectors. Furthermore, to ensure the currency of the findings, we applied a temporal range filter to select for only articles published in the last 5 years (2020-2024) as shown in Fig. 2. In total, 150 articles were identified after this first search.

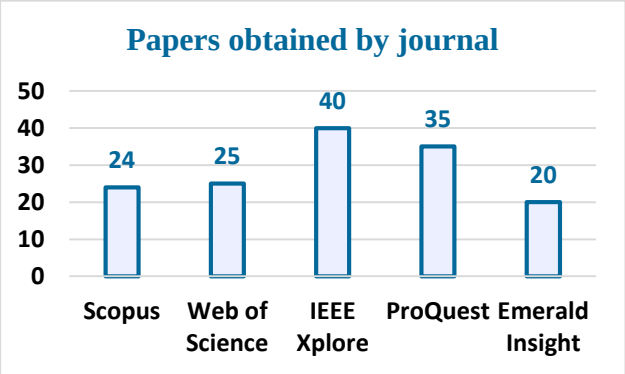


Fig. 1. Articles obtained by journal.

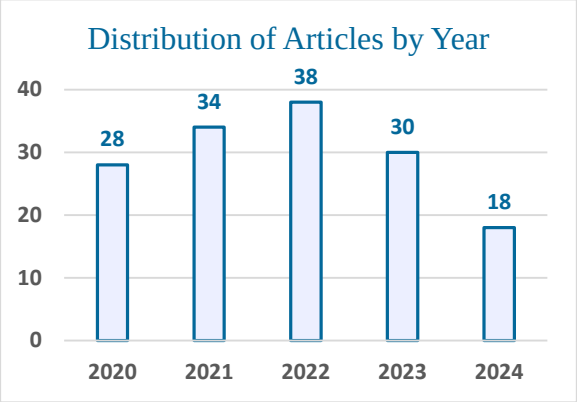


Fig. 2. Articles by publication year.

B. Screening

In this phase, initial studies that are duplicates or irrelevant are removed by following pre-defined criteria. Shamseer et al. (2015) highlight that the current stage is important to reduce the effort of unnecessary data return and that only studies with a relevant focus proceed to the next stages [21]. Then we removed duplicate records, so this time we were left with 104 articles. In the next stage, the titles and abstracts of the rest of the articles were reviewed to determine relevance. Forty articles were rejected if they did not satisfy the established inclusion criteria, which included the following: their exclusive focus on domains such as health or education; failing to directly relate to organizational decision-making. Finally, 64 articles were reviewed in detail at this stage.

C. Eligibility

So, the data from included studies were appraised in detail, to ascertain/assure the applicability of studies for systematic review. According to Galetsi et al. In this step, each publication was reviewed regarding its methodology and results to evaluate the quality of its content [22]. Further inclusion and exclusion criteria were applied during the full-text review of the 64 articles to assess study appropriateness. At this stage, 14 articles were excluded because they lacked sufficient methodological details, while another 18 were removed due to reasons such as not presenting results relevant to the organizational context or containing incomplete data. Ultimately, 32 articles met the eligibility criteria, forming a robust foundation for the analysis and ensuring the study's methodological rigor.

D. Inclusion

Thus, in the systematic review inclusion phase, verifying the findings of relevant papers is crucial to ensure that the studies selected for inclusion within a review align with the review objectives. Anderson et al. (2021) stress the importance of this phase for the synthesis outcomes, where they must be aligned with the optimal criteria and appropriate synthesis techniques to generate trustworthy findings. Their findings highlight the significance of developing clear research questions and ensuring methodological transparency, so the studies included are both relevant and applicable. Such a rigorous approach provides for improved analysis, improving the likelihood that the findings are actionable, thereby making the contribution that much more useful for policy-makers and organizational leaders [23]. Following the methodology described in the previous section, we reviewed several articles, and 32 of them were appropriate for the systematic review. They provide a solid basis of how data analytics may help organizational decision-creating. These share topics varying from one to another including Big Data Maturity Model, drivers of business analytics adoption, and specific applications in various fields such as supply chain management; thus, showcasing the relevance of the same across various organizational frameworks.

E. General Results

- Initial items identified: 150
- Duplicates removed: 46
- Articles discarded by title/abstract: 40
- Articles evaluated in full text: 64
- Articles excluded in full text: 32 (14 for lack of methodological data, 18 for other reasons)
- Final articles included: 32

Finally, we should clarify that of the total of 18 articles excluded for “other reasons”, the specific reasons are as follows: 10 articles were not relevant to the organizational context; 5 had insufficiently described methodologies; 3 were not available in their complete version; 3 were not available in their full version; and 3 were not available in their complete version.

Fig. 3 illustrates the application of the PRISMA Method, and the results obtained in each of its phases. This approach has allowed us to clearly and transparently structure and document the entire review process, from the initial identification of studies to the final selection of relevant articles. Using this paradigm not only guaranteed a more precise and complete interpretation but also enhanced the strength and fidelity of the study. We followed this systematic framework in order not only to source relevant studies but also to continuously reduce the risk of bias and to synthesize the evidence coherently so that the results are a true reflection of the current landscape of the topic investigated.

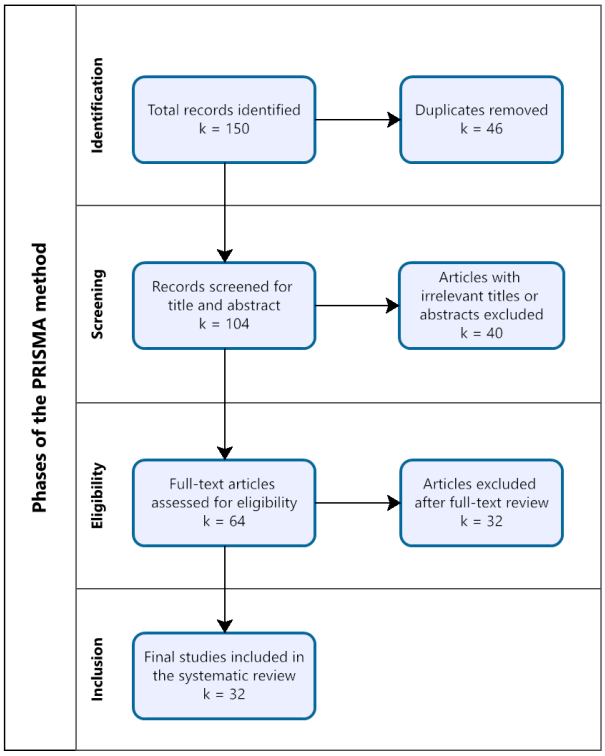


Fig. 3. PRISMA Method phases.

IV. Results

Findings were categorized and analyzed thematically into eight areas according to the final subset of included articles. Through the themes, we give an overview of the key issues and findings, capturing the range of contributions, with these summarized in Table I.

A. Data Quality and its Impact on Analytics

The success of analytical initiatives hinges on the quality of data. Galetsi et al. demonstrated how data standards enable better interoperability and reduce fragmentation issues [1]. Al-Sai et al. emphasized the requirement to measure organizational maturity to implement big data effectively [5]. Lin et al. analyzed how the use of analytical systems can be integrated and usefully implemented when data is broken into parts [24]. Salazar-Reyna et al. Data Accessibility & Cleanliness: [25] highlighted the importance of accessible clean data as the most significant data-based barrier. Ifenthaler & Yau pointed out, "the stronger data literacy, the better analytic practices can be adopted" [26].

B. Organizational and Cultural Resilience

Cultural and organizational resistance remains a major challenge. Horani et al. highlighted that lack of organizational leadership and support hinders the implementation of analytic tools [7]. Kew & Tasir noted that a lack of training in analytical tools limits their adoption in educational settings [11]. Gonzales & Horita highlighted that the lack of intuitive design in visual analytical tools generates rejection among end users [10]. Teniwut & Hasyim identified inefficient workflows as major barriers in supply chains [27].

TABLE I. Articles Grouped By Topic

Thematic	Related Articles	Brief Description of the Subject Matter
Data Quality and its Impact on Analytics	Galetsi et al. [1], Al-Sai et al. [5], Lin et al. [24], Salazar-Reyna et al. [25], Ifenthaler & Yau [26]	It focuses on how data quality affects the integration and effective use of analytical systems in various applications
Organizational and Cultural Resilience	Horani et al. [7], Kew & Tasir [11], Gonzales & Horita [10], Teniwut & Hasyim [27]	It examines the cultural and organizational barriers that hinder the implementation of analytical tools in different sectors.
Improved Decision Making	Komolafe et al. [3], Christenson Jr. & Goldstein [28], Aprijal et al. [29], Yang & Wang [30]	Analyzes how analytical tools support the identification of patterns and the optimization of strategic decisions.
Approaches to Optimization	Akindote et al. [13], Khanra et al. [14], Alnoukari [12], Soylu et al. [31]	Details the use of innovative approaches and advanced tools to improve operational efficiency and decision making.
Innovation in Specific Sectors	Tawil et al. [16], Ayuningtyas et al. [32], Raja et al. [34]	Explores how analytics is driving innovation in emerging industries and in the transformation of existing processes.
Use of Big Data in Decision Making	Rahman [9], Raja et al. [34], Matcha et al. [35], El Falah, Z., et al. [36]	Highlights the impact of big data in improving organizational decisions through personalization and advanced analytics.
Big Data Analytics Adoption Strategies	Patricio-Peralta et al. [2], Al Nuaimi & Awofeso [4], Astudillo et al. [6], Di Berardino & Vona [18]	Presents key strategies for effectively adopting big data in enterprise environments.
Critical Factors in the Adoption of Big Data Analytics	Dwilaga [19], Grander et al. [37], Hu, L., & Shu, Y. [38], Orlu et al. [39]	It examines the organizational and technological factors necessary for successful integration of analytics into complex systems.

C. Improved Decision Making

Analytical tools have transformed decision-making in different sectors. Komolafe et al. demonstrated that these tools generate competitive advantages by identifying patterns in emerging markets [3]. Christenson Jr. & Goldstein highlighted the positive impact of analytics on risk mitigation and optimization of strategic processes [28]. Aprijal et al. found that analytics in manufacturing and retail reduce operating costs and improve efficiency [29]. Yang & Wang analyzed how systematic evaluations optimize scalability in big data environments [30].

D. Approaches to Optimization

Innovative tools and approaches have overcome specific problems. Akindote et al. explored the integration of GIS and analytical tools to optimize decision-making systems [13]. Khanra et al. observed that analytical tools significantly improve operational efficiency in healthcare systems [14]. Alnoukari highlighted how business intelligence tools help organizations optimize strategic decisions through advanced data analysis [12]. Soylu et al. showed how these tools increase transparency in public administration [31].

E. Innovation in Specific Sectors

Analytics has generated significant innovations in multiple sectors. Mansour & Bick explored how big data has scaled digital transformation in the healthcare sector [15]. Tawil et al. highlighted barriers and opportunities for data-driven decision-making in emerging economies, fostering innovation in various sectors [16]. Ayuningtyas et al. highlighted applications of analytics in emerging sectors to optimize organizational processes [32]. Rakesh et al. proposed approaches to improve scalability in big data, facilitating the use of these technologies in varied sectors [33].

F. Use of Big Data in Decision Making

Big data has been shown to improve the quality of organizational decisions. Rahman proposed integration frameworks to maximize the impact of business intelligence [9]. Raja et al. identified how big data increases operational efficiency in healthcare systems [34]. Matcha et al. found that learning analytics dashboards facilitate decision-making in educational settings by providing more personalized information [35]. El Falah, Z., et al. showed how intelligent approaches to data analytics can transform the ability of organizations to process and apply big data in their strategic decisions [36].

G. Big Data Analytics Adoption Strategies

Regarding strategies for big data adoption, Patricio-Peralta et al. emphasized the need for trained personnel to maximize the marketing impact [2]. Al Nuaimi & Awofeso pointed out that infrastructure is key to optimizing supply chains [4]. Astudillo et al. proposed user-centric methodologies to integrate analytics in organizations [6]. Di Berardino & Vona identified analytical frameworks that facilitate strategic decision-making in corporate environments [18].

H. Critical Factors in the Adoption of Big Data Analytics

Key factors include organizational maturity and cultural support. Dwilaga proposed optimized decision models to overcome problems in complex industrial processes [19]. Grander et al. highlighted how advanced analytical tools can generate significant value in decision support systems [37]. Hu, L., & Shu, Y. developed intelligent approaches to analyze complex data, improving the effectiveness of strategic decisions [38]. Orlu et al. explored strategies to improve decision-making under uncertainty [39].

The findings of this study provide valuable insights for organizations seeking to navigate the complexities of data-driven decision-making. Only with the knowledge of the main barriers, like fragmented data systems and cultural resistance, can leaders work their way around these and craft strategies that would ensure that they foster a data-driven culture and still allocate funds towards building robust data infrastructures. These actions would not only surmount the active challenges but would also enable organizations to reap the full benefits of analytics: faster, more nuanced and integrative decision-making aligned with strategic objectives. With these strategies, organizations can promote innovation and maintain their competitive edge, enabling them to achieve real growth in the data-driven world.

V. Discussion

It dives into the insights that were identified and arranged into 8 Paramount Sections -- The Relation of Data Quality to Analytics, Organizational and Culture Resistance Towards Analytics Adoption, Tools and Techniques in Optimizing Analytics, The Providing of Strategies for Big Data Adoption, Industry-Driven Innovations, How Big Data Impacts Decision Making, Major Elements of Factors in Big Data Analytics Adoption and Influence, New Horizons of Decision Making. The point is, these categories afford a total framework for data-driven decision-making impact between industries. The discussion also reviews barriers to analytics adoption by organizations, while focusing on the advances made possible through more advanced tools and techniques. As a whole, shaping the advantages and realities in tens of thousands of organizations, these findings provide a comprehensive perspective on how to use data analytics, advancing towards whether organizations tune the best practices of data analytics or hinder the process of data analytics to sustain innovative outcomes in these highly data dependent enterprises.

A. Data Quality and its Impact on Analytics

The results reaffirm that data quality is a critical foundation for analytical initiatives to succeed. Galetsi et al. [1] and Lin et al. [24] highlight that fragmented data systems and a lack of interoperability are significant impediments to organizations' ability to seamlessly integrate data into their analytics systems. Al-Sai et al. big data requires an organization to be matured [5], as well, and Ifenthaler and Yau [26] argue that a crucial first step toward adopting analytics practices is increasing data literacy. But, notwithstanding those attempts, issues, such as setting adequate data quality thresholds or data access, remain Salazar-Reyna et al. [25] continue to get in the way of setting up effective analytics systems. These issues cannot be overcome in isolation: solving them requires a system-wide response that brings together technological and organizational solutions to maximize the potential of analytics.

B. Organizational and Cultural Resilience

Resistance at both the cultural and organizational levels continue to pose a major challenge to adopting analytics across various industries. Horani et al. [7] identified the lack of strong leadership and organizational support as critical barriers that hinder successful analytics adoption. Similarly, Kew and Tasir [26] found that inadequate training remains a significant obstacle, particularly in educational institutions where analytics has the potential to transform decision-making processes. To overcome these barriers, fostering a culture that values data-driven decision-making, supported by consistent leadership and tailored training programs, is essential for ensuring the effective implementation of analytics. These findings underscore the need for targeted strategies to address both leadership and skill development to facilitate the broader integration of analytics practices. Gonzales & Horita [10] stated how non-intuitive designs of analytical tools discourage the use of those tools and Teniwut & Hasyim [27] reported that disorganized workflows amplify the resistance to change in supply chains. These findings underscore the need for better training, stronger leadership, and redesigned technologies that make adoption easier.

C. Improved Decision Making

The transformation of decision-making through data analytics Komolafe et al. [3] showed analytics support competitive advantage based on patterns in born-global markets. According to Christenson Jr. & Goldstein [28], its impact is to provide better risk mitigation and optimize the strategic process. In addition, furthermore, demonstration of these techniques in industrial contexts has been made by Aprijal et al. [29] and Yang & Wang [30] describing relevant operational cost and scalation benefits for the solution. However, certain segments are struggling to implement these solutions, indicating the requirement for more bespoke solutions.

D. Approaches to Optimization

Such innovative tools and strategic approaches have been critical for the optimization of analytical systems. Akindote et al [13], Soylu et al. [31] also looked into how GIS and big data can be integrated to enhance transparency and decision-making in public administration. Business Intelligence tools play an important role in strategic decision solidifying, Alnoukari [12]. Khanra et al [14], on the other hand, demonstrated that these tools can increase efficiency in healthcare systems. However, problems related to interoperability and standardization limit effective implementation.

E. Innovation in Specific Sectors

Data analytics has generated disruptive innovations in specific sectors. Mansour & Bick [15] evidenced how big data is driving digital transformation in healthcare, while Tawil et al. [16] highlighted its impact in emerging economies. Ayuningtyas et al [32] and Rakesh et al. [33] discussed how these technologies optimize organizational processes and promote scalability. While these advances are promising, infrastructure and cultural barriers continue to limit their widespread adoption.

F. Use of Big Data in Decision Making

Big data has significantly improved the quality of organizational decisions. Rahman [9] developed frameworks that maximize the impact of business intelligence, while Raja et al. [34] evidenced operational improvements in healthcare systems. Matcha et al. [35] highlighted the benefits of analytic dashboards in educational settings. Finally, El Falah et al. [36] demonstrated how intelligent approaches in data analytics strengthen the strategic capability of organizations. However, the lack of alignment between strategic objectives and analytical capabilities persists as a challenge.

G. Big Data Analytics Adoption Strategies

Patricio-Peralta et al [2] stressed the importance of having trained personnel to maximize the impact of big data in marketing. Al Nuaimi & Awofeso [4] identified infrastructure as a critical factor in supply chains, while Astudillo et al. [6] and Di Berardino & Vona [18] proposed user-centric approaches and analytical frameworks that foster strategic decision-making. Although these strategies are fundamental, resistance to change and lack of specialized resources limit their effectiveness.

H. Critical Factors in the Adoption of Big Data Analytics

Key factors include organizational maturity, cultural support, and strategic focus. Dwilaga [19] proposed optimized models for industrial processes, while Grander et al. [37] highlighted the value generated by advanced tools. Hu, L., & Shu, Y. [38] evidenced that intelligent approaches improve the effectiveness of strategic decisions, and Orlu et al. [39] explored strategies to mitigate uncertainties in big data contexts. However, challenges persist in widespread implementation due to the lack of integration between departments.

I. Geographical Distribution of Scientific Contributions

Fig. 4 provides an overview of the distribution of articles by continent, highlighting the scientific contributions to the field of data analytics and decision-making. This analysis is key to understanding regional dynamics and priorities in research. The Americas leads with 15 articles, with the United States standing out as the largest contributor with 9 publications, reaffirming its prominent position in this field. Europe, contributing 12 articles, demonstrates a strong and diverse presence in the field, with significant contributions from countries like the United Kingdom and Germany. Asia, with 8 articles, showcases remarkable growth, driven by emerging economies such as India and China, solidifying its expanding role in this domain. Oceania, with 3 articles, maintains a steady but smaller contribution, while Africa, represented by only 1 article, highlights a clear gap and an opportunity to encourage research and development in the region. This distribution not only reveals regional disparities but also underscores the importance of fostering global collaborations to bridge these gaps and accelerate progress in underrepresented areas.

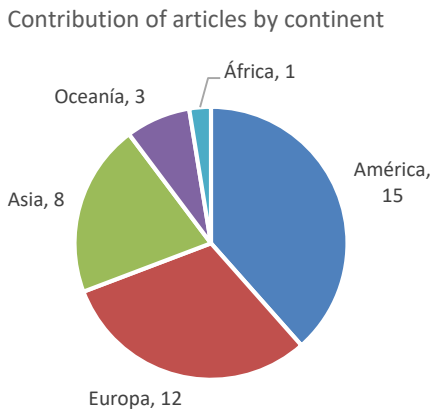


Fig. 4. Contribution of articles.

J. Methodological Diversity in Data Analytics and Decision Making

One main point of interest is the disparity between the number of techniques (65) compared to the number of articles analyzed (39) as detailed in Table II, which looks at the techniques and methodologies used in the reviewed articles. This difference is a testament to the effective diversity of methodology run in data analytics and decision-making. One of the main reasons this happens is the routine use of combining complementary methods in a single study. This method allows researchers to tackle issues from different angles and, as a result, create more resilient and all-encompassing solutions. Such a study could combine predictive analytics and machine learning algorithms to identify trends and predict future outcomes, while at the same time applying data visualization tools to expose the insight in a more digestible and telling manner.

TABLE II. TECHNIQUE / METHODOLOGY

Technique / Methodology	Quantity
Systematic Review	15
Big Data Analysis	10
Decision Modeling	8
Data Mining	7
Predictive Analytics	6
Machine Learning	5
Analytics Dashboards	4
Narrative Analysis	3
Decision Support Systems	3
Maturity Assessment	2
Assessment Framework	2

Additionally, this diversity in methodology is a reflection of the inherently interdisciplinary nature of the field, with methodologies borrowed from domains such as statistics, artificial intelligence, and organizational management [8]. Most of the studies use some secondary/supplementary methods to augment or triangulate the analysis obtained using the primary method. This is especially prevalent in studies that seek to compare the relative effectiveness of different implementations or to investigate different applications in the same setting. Another reason for this use of multiple approaches is the need to customize solutions according to the industry, for example, healthcare, manufacturing, or logistics which usually face complex and interdependent challenges that require more of a holistic approach.

K. Study Limitations

To be transparent about what was found and the scope of that finding, the table below sets out the main limitations identified. However, recognizing these limitations is key to appropriately assessing the results and directing future research efforts.

To aid the interpretation of the results and support transparency about the boundaries of the study, Table III below summarizes the key limitations we identified. It is necessary to consider these limitations in correctly interpreting the results and guiding the future at the same time.

The benefits and challenges uncovered in this study serve as a roadmap for organizations striving to embrace data analytics effectively. Acknowledging benefits like faster decision-making, enhanced accuracy and better alignment with strategic objectives enables leaders to advocate for analytics investments. Additionally, by learning about common pain points such as fragmented data systems and cultural resistance, organizations can better prepare for roadblocks and work proactively to counter them. A comprehensive approach that harmonizes technology with strategic functions can cultivate an ecosystem where data-driven insights thrive. This not only helps in the effective adoption of analytics but also promotes the innovative spirit and maintains the competitiveness of organizations in today's world of data.

TABLE III. STUDY LIMITATIONS

Limitation	Description
Restricted Temporal Coverage	The review focuses on articles published between 2020 and 2024, which may exclude earlier studies that could offer additional perspectives or alternative approaches to data analytics adoption and impact.
Database Selection and Publication Bias	The review uses well-known databases (Scopus, Web of Science, IEEE Xplore, ProQuest, and Emerald Insight). This might lead to the exclusion of relevant studies from grey literature or other unindexed sources, potentially affecting the representativeness of the findings.
Reliance on the Quality of Included Studies	The robustness of the conclusions depends on the methodological rigor of the 32 articles reviewed. Any limitations in the methodology of the included studies could influence the reliability and validity of the overall synthesis.
Lack of Empirical Validation	The study is based primarily on a systematic literature review without supplementary empirical or experimental validation to confirm the effectiveness of the proposed recommendations or conclusions in real-world settings.
Limited Applicability to Certain Sectors	Although the review includes studies from various sectors (e.g., healthcare, marketing, logistics, manufacturing, education, e-commerce), many examples are concentrated in healthcare, marketing, and supply chain management. This may limit the generalizability of the findings to sectors with different challenges.

VI. CONCLUSION

The systematic review shows that data analytics significantly contributes to changing the way decisions are being made in organizations. Synthesizing a body of 32 studies published between the years of 2020–2024, the review distilled eight key themes—including data quality and organizational resilience, to approaches for optimization and strategies for adoption—that together depict the facilitators and impediments to adopting data-driven decision-making.

The key findings highlight the enhanced accuracy, timeliness, and competitive advantage through data analytics, albeit under considerable challenges. These aspects are said to be the likes of data fragmentation, insufficient interoperability, organizational resistance as well as the lack of holistic plans integrating both the technical and socio-cultural factors.

An important value added of this review is its full template of themes revealing the factors enabling or constraining successful integration of analytics. The synthesis highlights the need to improve data quality, create accountability for data management and use, and establish strong and holistic approaches to addressing barriers identified to support data driven development.

For practical validation, future studies should compare and contrast the effectiveness of varying data analytics methodologies. This work will further shape and guide the current understanding of how and when to implement analytics solutions in various organizational contexts through concrete recommendations.

To conclude, data analytics have proven to be essential for enhancing organizational efficiency and making strategic data-driven decisions. However, to achieve these advantages, it is important to address the challenges identified through specific empirical investigation and focus on integrative, adaptive strategies.

VII. Future Work

This study has illuminated both the significant progress and persistent challenges in leveraging data analytics to enhance organizational decision-making. Building on these findings, future research should focus on key areas that can drive meaningful advancements and ensure sustainable growth in this field.

One of the essential factors is the uniformity of the quality of data, according to which the quality of data will be ensured, particularly for organizations that are working through different industries and geographies. It is also essential to understand the socio-cultural dynamics that shape how these analytics tools are adopted, as this is often the determining factor of where such initiatives succeed or fail. One way to leverage its transformative potential in sectors that have not applied data analytics (for example education, environment, management).

Another exciting opportunity is building intuitive, frictionless tools made for the needs of developing economies. These tools can democratize access to advanced analytics and bridge technological gaps. Real-time processing of the data collected by IoT devices through incorporation into analytical processes can lead to new innovative opportunities for improving the efficiency of decision-making processes in organizations. How might we develop strong socio-technical frameworks to help with this problem, another area where more research is needed, as theory has yet to grapple with the cultural complexity that can result through this process, which puts together technology, organizational processes, and human factors in an interdisciplinary context?

While this systematic review aims to synthesize the literature available, it is worth noting that most of the studies being reviewed offer solutions without strong empirical validations of these or direct comparisons with other approaches. Although our work was not an independent experimental validation—in line with the systematic review format—we recognize the need for empirical validation in establishing the real-world utility of data analytics approaches. Future work might include a meta-analysis of validation results from the primary studies, or even empirical comparative studies of well-justified approaches under comparable conditions. Such initiatives would deepen our insights about the relative advantages and disadvantages of each approach and provide more refined guidelines of best practices about how to employ data analytics in organizational decision making.

There is also continued important research on how to optimize analytical processes. Hybrid methods have been found useful in multiple domains, such as using GIS with analytics for supply chain optimization [13], implementing advanced tools for operational improvements in healthcare [14], and using a decision support system that can reduce cost and performance-focused management in logistics environments [27]. When developing targeted, sector-specific solutions that build upon the current tools available as well as novel approaches, these insights should serve as a guide.

Finally, the eight themes identified in this study provide a roadmap for advancing knowledge and application in data analytics. Each theme, ranging from data quality to big data adoption strategies, offers distinct opportunities for exploration, ensuring that future research addresses challenges holistically while maximizing the benefits of analytics in decision-making. By focusing on these areas, researchers can ensure that analytics evolves as an inclusive, adaptable, and transformative force across diverse organizational and societal contexts. These proposals are presented in detail in Table IV, allowing a clear and organized visualization of the research priorities in each thematic area.

TABLE IV. Future Research Directions by Thematic Area

Thematical	Details
Data Quality and its Impact on Analytics	Explore automated tools to ensure real-time data quality. Investigate global standards to improve system interoperability. Evaluate case studies of fragmented data in complex organizations.
Organizational and Cultural Resilience	Design training programs to reduce resistance to the use of analytical tools. Study the socio-cultural factors that influence the adoption of analytics in different regions.
Improved Decision Making	Develop hybrid models that combine machine learning with traditional decision-making approaches. Evaluate the impact of analytics on long-term strategic decisions in key industries.
Approaches to Optimization	Test the integration of emerging technologies, such as blockchain, into analytics systems. Design frameworks for the simultaneous implementation of analytics and GIS in industrial contexts.
Innovation in Specific Sectors	Analyze case studies in emerging sectors such as agriculture and renewable energy. Evaluate the feasibility of advanced analytics in low resource sectors.
Use of Big Data in Decision Making	Study the effectiveness of public policies to foster the adoption of big data in emerging economies. Propose strategies to reduce implementation costs in small and medium enterprises.
Big Data Analytics Adoption Strategies	Investigate how big data can improve prediction and planning in global crises. Develop specific tools for the integration of big data in governmental decision-making processes.
Critical Factors in the Adoption of Big Data Analytics	Identify and measure new key indicators of success in big data adoption. Explore how cultural differences influence adoption rates in different organizational contexts.

References

[1] Galetsi, P., Katsaliaki, K., & Kumar, S. (2019). Values, challenges and future directions of big data analytics in healthcare: A systematic review. Social Science & Medicine. https://doi.org/10.1016/j.socscimed.2019.112533

[2] Patricio-Peralta, C., Mondragon, J. Z., Terrones, L. S., & Ramirez Villacorta, J. (2024). Big data analysis and its impact on the marketing industry: a systematic review. Indonesian Journal of Electrical Engineering and Computer Science. https://doi.org/10.11591/ijeecs.v35.i2.pp1032-1040

[3] Komolafe, A. M., Aderotoye, I. A., Abiona, O. O., Adewusi, A. O., Obijuru, A., Modupe, O. T., & Oyeniran, O. C. (2024). Harnessing business analytics for gaining competitive advantage in emerging

markets: A systematic review of approaches and outcomes. International Journal of Management & Entrepreneurship Research. https://doi.org/10.51594/ijmer.v6i3.939

[4] Al Nuaimi, D., & Awofeso, N. (2024). The Value of Applying Big Data Analytics in Health Supply Chain Management. F1000Research. https://doi.org/10.12688/f1000research.156525.1

[5] Al-Sai, Z., Husin, M., Syed-Mohamad, S. M., Abdullah, R., Zitar, R. A., Abualigah, L., & Gandomi, A. (2022). Big Data Maturity Assessment Models: A Systematic Literature Review. Big Data and Cognitive Computing. https://doi.org/10.3390/bdcc7010002

[6] Astudillo, B. K., Santórum, M., & Aguilar, J. (2020). A Methodology for Data Analytics Based on Organizational Characterization Through a User-Centered Design. Journal of Organizational Analytics. https://doi.org/10.1007/978-3-030-51828-8_20

[7] Horani, O., Khatibi, A., Al-Soud, A., Tham, J., & Al-Adwan, A. (2023). Determining the Factors Influencing Business Analytics Adoption at Organizational Level: A Systematic Literature Review. Big Data and Cognitive Computing. https://doi.org/10.3390/bdcc7030125

[8] Szukits, Á., Móricz, P. Towards data-driven decision making: the role of analytical culture and centralization efforts. Rev Manag Sci 18, 2849–2887 (2024). https://doi.org/10.1007/s11846-023-00694-1

[9] Rahman, M. M. (2024). Systematic Review of Business Intelligence and Analytics Capabilities in Healthcare Using PRISMA. International Journal of Health and Medical. https://doi.org/10.62304/ijhm.v1i04.207

[10] Gustavo Romão Gonzales and Flávio Horita. 2020. Supporting visual analytics in decision support system: a systematic mapping study. In Proceedings of the 19th Brazilian Symposium on Human Factors in Computing Systems (IHC '20). Association for Computing Machinery, New York, NY, USA, Article 34, 1–10. https://doi.org/10.1145/3424953.3426483

[11] Kew, S.N., & Tasir, Z. (2022). Learning Analytics in Online Learning Environment: A Systematic Review on the Focuses and the Types of Student-Related Analytics Data. Tech Know Learn 27, 405–427. https://doi.org/10.1007/s10758-021-09541-2

[12] Alnoukari, M. (2020). Examining the Organizational Impact of Business Intelligence and Big Data Based on Management Theory. Business Analytics Journal. https://doi.org/10.37380/jisib.v10i3.637

[13] Akindote, O. J., et al. (2023). Comparative Review of Big Data Analytics and GIS. International Journal of Geographic Information Systems. https://doi.org/10.30574/wjarr.2023.20.3.2589

[14] Khanra, S., et al. (2020). Big Data Analytics in Healthcare: A Systematic Literature Review. Journal of Healthcare Informatics. https://doi.org/10.1080/17517575.2020.1812005

[15] Mansour T., Bick M. A Systematic Literature Review of Big Data Analytics in Healthcare Digital Transformation. JDS, 6(1), 3–17, (2024). https://doi.org/10.33847/2686-8296.6.1_1

[16] Tawil, A. -R. H., Mohamed, M., Schmoor, X., Vlachos, K., & Haidar, D. (2024). Trends and Challenges towards Effective Data-Driven Decision Making in UK Small and Medium-Sized Enterprises: Case Studies and Lessons Learnt from the Analysis of 85 Small and Medium-Sized Enterprises. Big Data and Cognitive Computing, 8(7), 79. https://doi.org/10.3390/bdcc8070079

[17] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. The BMJ. https://doi.org/10.1136/bmj.n71

[18] Di Berardino, D., & Vona, S. (2023). Discovering the Relationship Between Big Data and Decision Making. European Scientific Journal, ESJ. https://doi.org/10.19044/esj.2023.v19n19p1

[19] Dwilaga, A. T. (2023). Decision Model and Industry Optimization in Production: A Systematic Literature Review. Sainteks: Jurnal Sains dan Teknik. https://doi.org/10.37577/sainteks.v5i1.528

[20] Almasri, H., Zakuan, N., Amer, M., & Majid, M. R. (2021). A developed systematic literature review procedure with application in the field of digital transformation. Studies of Applied Economics. https://doi.org/10.25115/eea.v39i4.4559

[21] Rethlefsen, M.L., Kirtley, S., Waffenschmidt, S. et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 10, 39 (2021). https://doi.org/10.1186/s13643-020-01542-z

[22] Kahale, L. A., Hamadeh, G. N., & Ghalayini, W. (2021). Assessing the methodological quality of systematic reviews: insights from the AMSTAR-2 tool. F1000Research. https://doi.org/10.12688/f1000research.28686.1

[23] Anderson, R., Booth, A., Eastwood, A., Rodgers, M., Shaw, L., Thompson Coon, J., Briscoe, S., Cantrell, A., Chambers, D., Goyder, E., Nunns, M., Preston, L., Raine, G., & Thomas, S. (2021). Synthesis for health services and policy: case studies in the scoping of reviews. Health Services and Delivery Research, 9(15). https://doi.org/10.3310/hsdr09150

[24] Lin, J.S., Murad, M.H., Leas, B. et al. A Narrative Review and Proposed Framework for Using Health System Data with Systematic Reviews to Support Decision-making. J GEN INTERN MED 35, 1830–1835 (2020). https://doi.org/10.1007/s11606-020-05783-5

[25] Salazar-Reyna, R., Gonzalez-Aleu, F., Granda-Gutierrez, E.M.A., Diaz-Ramirez, J., Garza-Reyes, J.A. and Kumar, A. (2022), "A systematic literature review of data science, data analytics and machine learning applied to healthcare engineering systems", Management Decision, Vol. 60 No. 2, pp. 300-319. https://doi.org/10.1108/MD-01-2020-0035

[26] Ifenthaler, D., & Yau, J. Y. K. (2021). Supporting Teaching Staff Through Data Analytics. Technology, Knowledge and Learning. https://doi.org/10.14742/ascilite2021.0105

[27] Teniwut, W.A., & Hasyim, C.L. (2020). Decision support system in supply chain: A systematic literature review. Uncertain Supply Chain Management, 8, 131-148. https://doi.org/10.5267/j.uscm.2019.7.009

[28] Christenson Jr., A.P., & Goldstein, W.S. (2022). Impact of data analytics in transforming the decision-making process. Business & IT. https://doi.org/10.14311/bit.2022.01.09

[29] Aprijal, R., Siregar, I., Siahaan, A., & Marlina, L. (2024). Utilization of Data Analytics to Enhance Operational Efficiency in Manufacturing Companies. Journal of Computer Networks, Architecture and High Performance Computing. https://doi.org/10.47709/cnahpc.v6i2.3723

[30] Yang, F., Wang, M. A review of systematic evaluation and improvement in the big data environment. Front. Eng. Manag. 7, 27–46 (2020). https://doi.org/10.1007/s42524-020-0092-6

[31] Soylu, A., Corcho, Ó., Elvesæter, B., Badenes-Olmedo, C., Yedro-Martínez, F., Kovacic, M., Posinkovic, M., Medvešček, M., Makgill, I., Taggart, C., Simperl, E., Lech, T., & Roman, D. (2022). Data Quality Barriers for Transparency in Public Procurement. Inf., 13, 99. https://doi.org/10.3390/info13020099

[32] Ayuningtyas, A., et al. (2023). Big Data Analysis and Its Utilization for Business Decision-Making. World Journal of Business Analytics. https://doi.org/10.58812/wsist.v1i01.177

[33] Raut, R., Narwane, V., Mangla, S., Yadav, V., Narkhede, B., & Luthra, S. (2021). Unlocking causal relations of barriers to big data analytics in manufacturing firms. Ind. Manag. Data Syst., 121, 1939-1968. https://doi.org/10.1108/IMDS-02-2020-0066

[34] Raja, Rakesh, Mukherjee, Indrajit, Sarkar, Bikash Kanti, A Systematic Review of Healthcare Big Data, Scientific Programming, 2020, 5471849, 15 pages, 2020. https://doi.org/10.1155/2020/5471849

[35] Matcha, W., et al. (2020). Learning Analytics Dashboards: A Systematic Review. Educational Data Analytics Journal. https://doi.org/10.1109/TLT.2019.2916802

[36] EL FALAH Zineb, RAFALIA Najat and ABOUCHABAKA Jaafar, "An Intelligent Approach for Data Analysis and Decision Making in Big Data: A Case Study on E-commerce Industry" International Journal of Advanced Computer Science and Applications(IJACSA), 12(7), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120783

[37] Grander, Gustavo & Silva, Luciano & Santibanez Gonzalez, Ernesto. (2021). Big data as a value generator in decision support systems: a literature review. Revista de Gestão. ahead-of-print. https://doi.org/10.1108/REGE-03-2020-0014

[38] Lei Hu and Yangxia Shu, "Enhancing Decision-Making with Data Science in the Internet of Things Environments" International Journal of Advanced Computer Science and Applications(IJACSA), 14(9), 2023. http://dx.doi.org/10.14569/IJACSA.2023.01409120

[39] Orlu, G. U., Abdullah, R. B., Zaremohzzabieh, Z., Jusoh, Y. Y., Asadi, S., Qasem, Y. A. M., Nor, R. N. H., & Mohd Nasir, W. M. H. b. (2023). A Systematic Review of Literature on Sustaining Decision-Making in Healthcare Organizations Amid Imperfect Information in the Big Data Era. Sustainability, 15(21), 15476. https://doi.org/10.3390/su152115476

DyGAN: Generative Adversarial Network for Reproducing Handwriting Affected by Dyspraxia

Jesús Jaime Moreno Escobar[1], Hugo Quintana Espinosa[2], Erika Yolanda Aguilar del Villar[3]
Centro de Investigación en Computación, Instituto Politécnico Nacional, México[1]
Escuela Superior de Ingeniería Mecánica y Eléctrica, Zacatenco, Instituto Politécnico Nacional, México[2,3]

***Abstract*—Dyspraxia primarily affects coordination and is categorized into two forms: 1) Motor, and 2) Verbal ororal. This study focuses on motor dyspraxia, which influences individuals in learning movement-related tasks. Consequently, the *DyGAN* initiative employs deep convolutional aversarial generation networks, using deep learning to create characters resembling human handwriting. The methodology in this study is structured into two main stages: 1) the creation of a first-order cybernetic model, and 2) the execution phase. Using four independent variables and three dependent variables, eight outcomes were analyzed using variance analysis. DyGAN is a twin Deep Convolutional Neural Networks and it is highly sensitive to the Learning Rate. It scored a 67% on the proposal, suggesting that characters can sound written by a human. The project will feature writers from different backgrounds and will help augment data for writing resources for dyspraxia, potentially benefiting those struggling with writing difficulties and improving our understanding of education. The model is designed to be widely applicable. Future work could customize the model to mimic the way a specific child writes, with neural networks, for example.**

***Keywords*—Children with neurodevelopmental disorders; dyspraxia; generative adversarial network; deep learning; deep convolutional neural network; human handwriting**

I. Introduction

Since 2020, the World Health Organization (WHO) estimates that there are 1 billion people living with some form of disability, which means that approximately 15% of the world's population has difficulties in their psychosocial functioning and frequently requires assistance services. On the other hand, in Mexico, according to data from the National Institute of Statistics and Geography (INEGI), there are 6.2 million people who have some form of disability, which represents 4.9% of the total population.The most well-known activities for types of disabilities are six: 1) Walking, climbing, or descending, 2) Seeing, 3) Hearing, 4) Speaking or communicating, 5) Remembering or concentrating, and 6) Difficulty with bathing, dressing, or eating. People with disabilities may have more than one disability.

Dyspraxia, also known as Developmental Coordination Disorder (DCD), is a condition that affects physical coordination, which can make daily tasks more challenging for children. In the school setting, children with dyspraxia often struggle with activities that require fine and gross motor skills. This can include difficulties with handwriting, cutting with scissors, participating in sports, and even simple tasks such as tying shoelaces or buttoning clothes. These challenges can cause frustration and a sense of inadequacy among affected children, as they may find themselves behind their peers in performing seemingly simple tasks [1], [2].

The academic implications of dyspraxia are significant. Children with this condition may have difficulty with tasks that require motor planning and coordination, such as writing legibly and quickly, which is essential for taking notes and completing written assignments. This can result in lower academic performance not because of lack of understanding or intelligence but due to the physical difficulties associated with the disorder. Furthermore, the effort required to complete these tasks can be exhausting, leading to decreased stamina and increased stress, which can further affect a child's ability to learn and engage in the classroom [3], [4].

Socially, children with dyspraxia often face additional hurdles. Their difficulties with coordination and motor skills can make it difficult to participate in group activities and sports, which are crucial for social development and peer relationships. These children may be perceived as clumsy or awkward, leading to possible teasing or bullying from classmates. This social isolation can have profound effects on your self-esteem and general mental health. In addition, teachers may not always be aware or understand the needs of children with dyspraxia, leading to inadequate support and accommodations in the classroom. It is crucial for educators to receive proper training to recognize and support students with dyspraxia, ensuring that they have the tools and understanding necessary to help these children thrive academically and socially [5], [6]. Then, according to Pinos-Medramno et al. in [7], the school performance of children with dyspraxia in Basic Education has a negative impact on their normal development and therefore on their learning process, particularly in their writing. Addressing these challenges requires innovative solutions, such as the application of Artificial Intelligence (AI).

Artificial intelligence can be divided into two main types: 1) AI based on its functionality and 2) AI based on capacity. For the development of a DyGAN, capacity-based AI is utilized, which is further divided into three branches:

1) Artificial Narrow Intelligence (ANI): Its function is to focus on performing a single task, but it has limited memory. At this stage, it should be prepared to act in a single role, minimizing its performance as much as possible.
2) Artificial General Intelligence (AGI): It has the ability to mimic human thinking because of its high cognitive level.
3) Artificial Super Intelligence (ASI): It is the most powerful AI, as it has the capacity for faster learning

and achieving autonomy, which allows it to replicate human behavior and even surpass human thinking ability. However, as of 2023, it is still in the development phase.

For the conceptualization and advancement of a DyGAN, Generative Adversarial Networks (GANs) are utilized due to their proficiency in synthesizing character images derived from a pre-established database, potentially curated for general-purpose applications. The deployment of GANs is notably advantageous as they comprise two neural networks that engage in a competitive dynamic, thereby facilitating the generation of highly realistic imagery. This approach enables the efficient production of a diverse array of character images, which is particularly advantageous for applications necessitating extensive visual datasets. It is imperative to underscore that, at their core, Neural Networks are employed, categorized into five principal architectures:

1) Transformer Neural Networks: These are self-aware neural networks that have been developed for text and are currently driving significant advances in natural language processing.
2) Recurrent Neural Networks: This is a type of artificial neural network that has a sequential structure or time-series data. They are used in applications such as language translation, speech recognition, and image captioning.
3) Siamese Neural Networks: Their functionality is based on the use of two conditions for evaluation. After the evaluation is performed, the output is passed to a classifier which generates the result. They are primarily used for document evaluation.
4) Convolutional Neural Networks: They have many layers, each dedicated to detecting different visual features. Filters are applied to training images with different resolutions, and the output of each layer is obtained by convolving the image, which is then used as input for the next layer.
5) Generative Adversarial Networks: Their use is to generate images from existing datasets, with one network called the generator and the other called the discriminator, competing with each other to generate new instances that resemble those in the training data distribution.

Each of these architectures has distinct characteristics and strengths, making them suitable for a variety of tasks within the domains of artificial intelligence and machine learning. Consequently, the integration of Generative Adversarial Networks (GANs) with these fundamental neural network architectures can significantly enhance the performance and capabilities of a Dynamic Generative Adversarial Network (DyGAN) in generating high-fidelity character images. This advancement facilitates various applications, including digital handwriting analysis, educational tools for children with learning disabilities, and other scenarios where the generation of realistic character images is imperative. For the development of a DyGAN, GANs are employed to generate character images from a pre-established database, which could be designed for general use.

This work is therefore divided into four additional sections. The second section provides an analysis of related work on Generative Adversarial Networks (GANs). Then, this section explains the theory surrounding GANs, including essential definitions for understanding neural networks. The third section summarizes the methodology of this proposal, so that the following section can conduct experimentation aimed at finding optimal hyperparameters to improve the results. Finally, conclusions are developed, focusing on the research findings and possible improvements or future directions that the work could take.

II. Method

The study of dyspraxia dates back to the late 19th century, with the British physician Sir William John Little being the first to study it in depth and name it as such [8]. By 1937, Samuel Orton declared it one of the six most common developmental disorders and showed a distinctive impairment of praxis. He also titled it *Reading, Writing, and Speech Problems in Children*. He was one of the first to break away from the concept of simple recovery reading and treat all aspects of language as related, doing so in a language clinic [9]. Later, in 1972, Anna Jean Ayres categorized Dyspraxia as a Sensory Integration Disorder. During this period, she wrote several books, two of which stand out: 1)S̃ensory Integration and Learning Disorders (1972) and 2)S̃ensory Integration and the Child (1979). In addition, she published multiple academic articles that addressed her theory and techniques for clinical application and founded the *Ayres Clinic*, based in Torrance, California, USA, where she evaluated and treated children using her developed approach. Sensory integration therapy emphasizes detailed evaluation, understanding the unique sensory challenges and styles of each child, which forms the basis for providing the child with appropriate learning opportunities, processing, and using sensory information to improve the child's performance skills [10].

A. Related Work

The exploration of advanced technologies for neurodivergent learning begins with the study *Exploring the Efficacy of an IoT Device as a Sensory Feedback Tool in Facilitating Learning for Neurodivergent Students* in [11]. This research investigates how IoT devices can provide real-time sensory feedback to enhance the learning experiences of neurodivergent students. These devices offer customized responses that improve engagement and learning outcomes by creating adaptive learning environments. Similarly, the *Web-based Assessment and Training Model for Dyslexia, Dyscalculia, Dysgraphia, Dyspraxia, ADHD and Autism* in [12] provides a digital platform for individualized assessments and training. The Web-based model integrates educational games and cognitive exercises, offering flexibility and scalability for diverse educational settings. Both studies underscore the potential of digital tools to support personalized and inclusive education.

Building on the theme of interactive learning technologies, the study *Improving Cognitive Learning of Children with Dyspraxia Using Selection-Based Mid-Air Gestures in the Athynos Game* [13] explores the use of Mid-Air Gestures-Based Interactions in the Athynos game to improve cognitive learning for children with dyspraxia. This approach not only

engages children more effectively but also improves their cognitive and motor skills. This research aligns with *ATHYNOS: Helping Children with Dyspraxia Through an Augmented Reality Serious Game* in [14], which uses augmented reality to create interactive learning experiences targeting specific motor and cognitive skills. Both studies highlight the effectiveness of immersive and interactive technologies in supporting children with motor impairments, emphasizing the role of serious games in educational and therapeutic settings.

The advancement of computational diagnostic systems is exemplified by *TestGraphia, a Software System for the Early Diagnosis of Dysgraphia* in [15], which uses advanced algorithms to analyze handwriting and identify patterns of dysgraphia. This objective and quantifiable method aids in early diagnosis, which is critical for timely interventions. Complementing this approach is *Preventing Dyspraxia: A Project for the Creation of a Computational Diagnostic System Based on the Theory of Embodied Cognition* in [16], which leverages computational techniques to assess motor and cognitive functions, providing a holistic view of dyspraxia. Both studies emphasize the importance of early detection and intervention, using computational tools to enhance diagnostic accuracy and support effective therapeutic strategies.

The integration of wearable haptics and Virtual Reality (VR) in rehabilitation is explored in *Wearable Haptics and Immersive Virtual Reality Rehabilitation Training in Children with Neuromotor Impairments* in [17]. This study shows that combining haptic feedback with virtual reality significantly enhances the rehabilitation process by providing engaging and interactive experiences. Similarly, *Integration of Serious Games and Wearable Haptic Interfaces for Neuro Rehabilitation of Children with Movement Disorders: A Feasibility Study* in [18] examines the feasibility of using serious games and haptic interfaces for neurorehabilitation, reporting improvements in motor functions and user participation. These studies demonstrate the potential of immersive technologies to create effective rehabilitation environments for children with neuromotor impairments.

Focusing on learning disabilities, *Rotoscopy-Handwriting Prototype: Using computer animation technique to assist handwriting teaching for children with dyspraxia* in [19] and *using the rotoscopy technique to assist handwriting teaching for children with dyspraxia* in [20] both present innovative methods employing rotoscopy and computer animation to teach handwriting. These systems provide visual and kinesthetic feedback, helping children with dyspraxia improve their handwriting skills through interactive exercises. These studies highlight the value of specialized software in addressing specific educational challenges associated with learning disabilities.

The study *Puzzle Time - VR Runner* [21] presents a VR game designed to support the development of cognitive and motor skills in children. By combining cognitive puzzles with physical activities, the game encourages physical movement and cognitive engagement, making it a useful tool for educational and therapeutic purposes. This approach is echoed in *Case Study: Using a Novel Virtual Reality Computer Game for Occupational Therapy Intervention* in [22], which explores the use of a VR game for occupational therapy. The positive results in motor skills and patient participation reported in these studies underscore the potential of VR games as effective tools to enhance cognitive and motor skills.

Lastly, the application of advanced machine learning techniques is highlighted in *Integrated Transfer Learning Based on Group Sparse Bayesian Linear Discriminant Analysis for Error-Related Potential Detection* in [23]. This study uses transfer learning to improve the detection accuracy of error-related potentials in EEG data, demonstrating the utility of machine learning in neurorehabilitation settings. Additionally, *Contemporary Speech/Speaker Recognition with Speech from Impaired Vocal Apparatus* in [24] addresses the challenges of speech recognition for individuals with impaired vocal apparatus, emphasizing the importance of inclusive technologies that accommodate diverse needs. Both studies show advances in machine learning and speech recognition technologies that contribute to better diagnostic and therapeutic outcomes.

Together, these studies illustrate the significant progress that is being made in the use of advanced technologies to support neurodivergent learning and rehabilitation. From IoT devices and web-based platforms to immersive VR and specialized software, these innovations are creating more inclusive, engaging, and effective educational and therapeutic environments. Using these technologies, educators and clinicians can provide better support and interventions, ultimately improving outcomes for people with learning and developmental disorders.

B. Deep Convolutional Generative Adversarial Networks

Deep Convolutional Generative Adversarial Networks (DCGANs) constitute a particular type of neural network designed to generate data from an existing dataset as observed in Fig. 1. The primary goal of DCGANs is to examine and understand the distribution of data within the training set, with the purpose of reliably generating new data that follow this distribution [25].

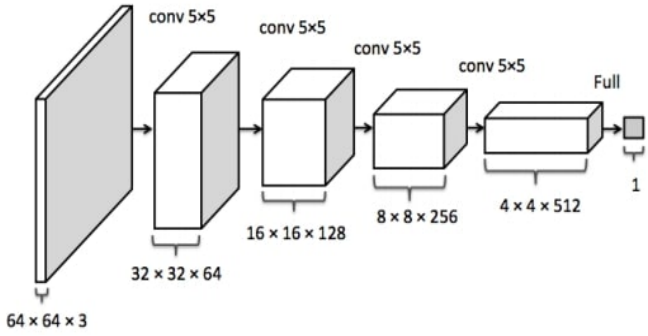


Fig. 1. Representation of Deep Convolutional Generative Adversarial Networks (DCGANs) [26].

A GAN network consists of two neural components: 1) a *Generator* (G), and 2) a *Discriminator* (D). The primary function of the generator is to understand the distribution of the training dataset and produce information that fits that distribution. On the other hand, the Discriminator evaluates the probability of authenticity of a given data point, determining whether it comes from the real training set or if it is synthetic, generated by G. Fig. 2 shows not only the architecture, but also the interaction between G and D.

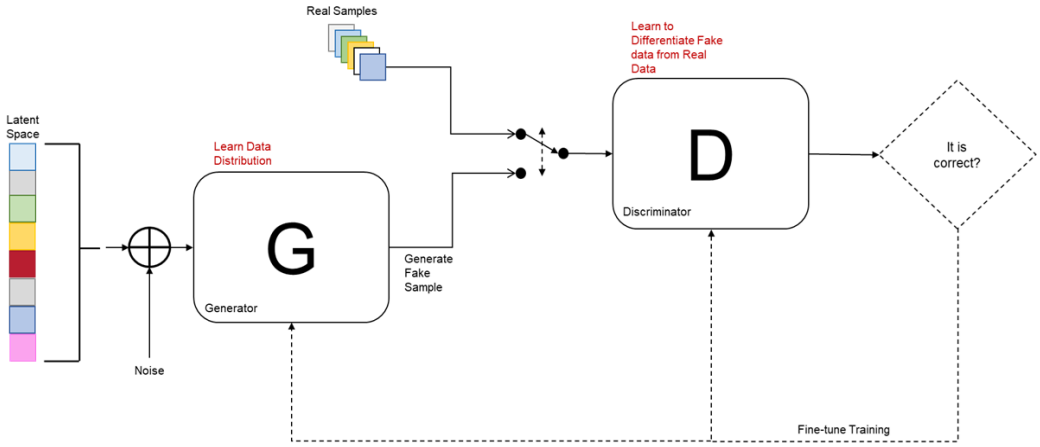


Fig. 2. Architecture of a DCGAN [27].

In this manner, **G** and **D** compete against each other through a Minimax game, where each adversary seeks to maximize its actions while simultaneously minimizing those of the opponent. The mathematical representation of this Minimax game in GANs is expressed by Eq. 1.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \quad (1)$$

The objective is to train the discriminator **D** to accurately classify the data as authentic, maximizing $\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)]$, or synthetic, maximizing $\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]$, where the output probability should approach zero. Concurrently, the generator **G** is optimized to deceive the discriminator **D** through the generation of data that closely resemble the training set, thus minimizing $\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]$. Within this framework, the MinMax game is engaged solely in the latter part of the equation. As Goodfellow elucidated in [28] the aforementioned equation may not provide an adequate gradient for **G** to learn effectively. This limitation arises because, during the initial phases, the data synthesized by **G** are of insufficient quality, leading **D** to reject them with high confidence, given their stark divergence from the real training data. Then Eq. 2 provides a summary of how the Generator and the Discriminator compute the gradient throughout the training process.

$$V_{\text{GAN}}(D,G) = \begin{cases} D: \max_D \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] \\ +\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1-D(G(\mathbf{z})))] \\ G: \max_G \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(D(G(\mathbf{z})))] \end{cases} \quad (2)$$

In addition, Eq. 2 describes the objective functions for Generative Adversarial Networks, where we can identify the two main components in it: the Discriminator (D) and the Generator (G).

$$\textbf{Discriminator}(D): \max_D \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1-D(G(\mathbf{z})))] \quad (3)$$

The Discriminator's objective is to maximize the expected value of correctly distinguishing between real data (x) from the data distribution and generated data (G(z)) from the Generator, Eq. 3. It seeks to maximize the logarithmic probability that real data are classified as real and the logarithmic probability that generated data are classified as fake.

$$\textbf{Generator}(G): \max_G \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(D(G(\mathbf{z})))] \quad (4)$$

The Generator's objective is to maximize the expected value of the log probability that the Discriminator classifies the generated data as real, Eq. 4. Essentially, the Generator tries to fool the Discriminator by generating data that is indistinguishable from real data [27]. The Fig. 3 shows the mathematical interaction of the GAN components for the generation of an image or synthetic data.

C. DyGAN

In Fig. 4, the DyGAN first-order cybernetic model, an MNIST database and numbers are included. The algorithm was designed in such a way that the user enters a character and, in turn, connects to an MNIST database containing characters. In the output, the character is displayed on the screen. The system is controlled by feedback, as efficiency is important in determining whether accuracy has been achieved. This is the principle of DCGAN, which is a Deep Convolutional Generative Adversarial Network. The discriminator and the generator are involved in the process.

From Fig. 5, the Generator and Discriminator are fundamental processes that enable the functionality of DCGAN. The generator creates images from pixels, in this case, with a size of 28, which matches the one used in the MNIST database. In each epoch, the Generator produces an image and the Discriminator evaluates whether it is real or fake. If the discriminator indicates that it is fake, the process continues over several epochs until it can confuse the discriminator or match the generated images with those in the database. The following algorithm describes the training and inference process of the GAN designed for the MNIST dataset to aid children with dyspraxia. The GAN model consists of a

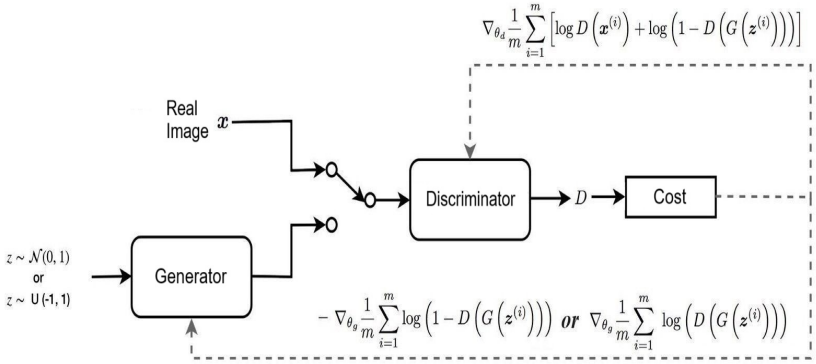


Fig. 3. Mathematical diagram of the generator and discriminator components.

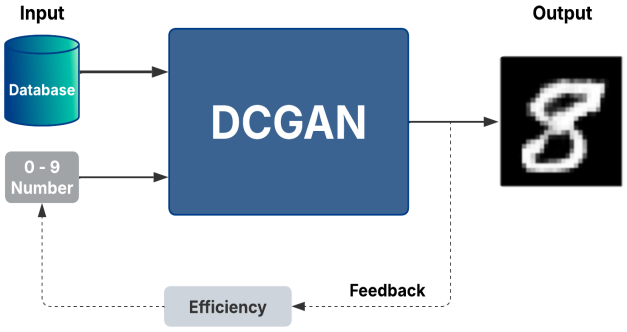


Fig. 4. DyGAN first-order cybernetic model over MNIST database, only numbers are included.

generator and a discriminator, where the generator creates new images from random noise, and the discriminator evaluates the authenticity of the images. Stages of the training and inference methodology for a GAN on MNIST to help children with dyspraxia Here we will preprocess the data and train our model followed by deployment of models.

Algorithm 1: GAN Training and Inference for MNIST

Initialize: Generator G, Discriminator D, learning rate α, batch size m **Input:** MNIST dataset X, random noise Z

Training Phase **for** *number of training iterations* **do**

Sample minibatch of m noise samples $\{z^{(1)}, \ldots, z^{(m)}\}$ from noise prior $p_z(z)$ Sample minibatch of m examples $\{x^{(1)}, \ldots, x^{(m)}\}$ from data distribution $p_{data}(x)$

Update Discriminator: $\theta_d \leftarrow \theta_d + \alpha \nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} [\log D(x^{(i)}) + \log(1 - D(G(z^{(i)})))]$

Update Generator: $\theta_g \leftarrow \theta_g + \alpha \nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} [\log D(G(z^{(i)}))]$

Inference Phase **Input:** Trained Generator G, new random noise z **Output:** Generated image $G(z)$ Sample noise z from noise prior $p_z(z)$ Generate image using G: $\hat{x} = G(z)$ **return** $\hat{x}$

Algorithm 1 provides a comprehensive description of the training and inference process. In the Training Phase, the generator and discriminator are updated iteratively. The generator G learns to create realistic images by minimizing the discriminator's ability to distinguish between real and generated images, while the discriminator D aims to maximize its accuracy in this task. In the Inference Phase, the trained generator uses new random noise z to produce new images, providing a useful tool for evaluating therapeutic progress in children with dyspraxia through generated visual aids. This methodology ensures a systematic and effective training process, facilitating the creation of high-quality synthetic data that can be utilized in therapeutic settings.

III. Experimental Results

First, since we propose three independent variables: 1) batch size, 2) learning rate, and 3) activation function, we carried out a total of eight experiments, as shown in Table I. Based on this, an Analysis of Variance (ANOVA) was performed. The analysis in an ANOVA table is associated with certain factors being investigated, including the activation function, batch size, and learning rate, along with their residuals. The Sum of Squares tells us how much variance each source contributes to the total data set. Degrees of freedom (df) refer to how many values or quantities in a statistical distribution, for every factor. The sum of squares is divided by the corresponding degrees of freedom to calculate the mean square, which measures the average variance due to each source. Thus, the F-Ratio is calculated as a ratio of mean square of source/mean square residuals to tell if its variances are significantly different. P-Value - is the probability value, and it is used to determine the significance of an observed effect. Since we can obtain four dependent variables or test variables: 1) Discriminator Loss (D Loss), 2) General Accuracy, 3) Adversarial Loss, and 4) Discriminator Accuracy (D Accuracy), we can infer the behavior of the system in four independent branches. Table I also shows the best results for each dependent variable highlighted in gray.

Table II shows that it is evident that none of the factors, activation function, batch size, or learning rate, demonstrate statistically significant effects on overall efficiency over Discriminator Loss, as indicated by their high P-values. However, it should be noted that the *Learning Rate*, although not statistically significant, has a value of P closest to zero among the

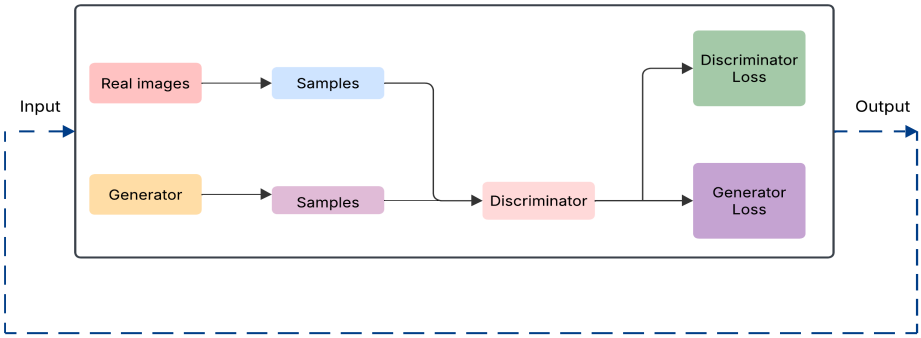


Fig. 5. DyGAN first-order cybernetic extended model over MNIST database, only numbers are included.

TABLE I. EXPERIMENTAL DESIGN BASED ON THREE INDEPENDENT VARIABLES (*italics*) WITH FOUR DEPENDENT OR OUTPUT VARIABLES (**BOLDS**)

Activation Function	*Batch Size*	*Learning Rate*	**D Loss**	**General Accuracy**	**Adversarial Loss**	**D Accuracy**
RELU	64	2e-4	0.626676	0.601562	0.623441	0.71875
RELU	64	2e-3	0.688536	0.652653	0.652653	0.59375
RELU	32	2e-4	0.684453	0.5625	0.762182	0.40625
RELU	32	2e-3	0.629332	0.671875	0.93706	0.21875
TANH	64	2e-4	0.626756	0.664062	0.897452	0.296875
TANH	64	2e-3	0.674438	0.523438	0.846442	0.234375
TANH	32	2e-4	0.583811	0.6875	0.718293	0.625
TANH	32	2e-3	0.694146	0.5625	0.791022	0.375

TABLE II. ANOVA TABLE WITH DIFFERENT SOURCES OF VARIABILITY, EFFECT OVER DISCRIMINATOR LOSS

Sources	**Sum of Squares**	**df**	**Mean Square**	**F-Ratio**	**P-Value**
A) Activation Function	0.000310578	1	0.000310578	0.17	0.7015
B) Batch Size	0.0000760391	1	0.0000760391	0.04	0.8485
C) Learning Rate	0.00339307	1	0.00339307	1.85	0.2450
Residuals	0.00732558	4	0.00183064		
Total (Corrected)	0.0111023	7			

TABLE III. ANOVA TABLE WITH DIFFERENT SOURCES OF VARIABILITY, EFFECT OVER GENERAL ACCURACY

Sources	**Sum of Squares**	**DF**	**Mean Square**	**F-Ratio**	**P-Value**
A) F.A	0.000326274	1	0.000326274	0.5	0.8286
B) Batch Size	0.000227484	1	0.000227484	0.04	0.8564
C) L.R	0.00138228	1	0.00138228	0.23	0.6592
Residuals	0.0244519	4	0.00611296		
Total(Corrected)	0.0263879	7			

three factors, making it the most significant factor compared to the activation function and batch size. This suggests that while the effects are not statistically significant in general, the Learning Rate has a relatively stronger influence on the outcome measure. From Fig. 6, we also notice that the learning rate factor (LR) is more significant when conducting the experiments, as it influences the way DyGAN learns and is a factor to consider.

The ANOVA of Table III analyzes the variability of the General Accuracy through contributions attributable to various factors. Since the Type III sum of squares has been chosen, the contribution of each factor is assessed by eliminating the effects of the other factors. The P-values indicate the statistical significance of each of these factors. Since no P-value is found to be less than 0.05, it is again concluded that none of the factors exerts a statistically significant effect on General Accuracy at a confidence level of 95%.

IV. CONCLUSIONS

We presented the Deep Convolutional Generative Adversarial Networks to lay down theoretical background for the proposal and discuss how it works, mathematically-based solutions, tools, and applications involving on it. Furthermore, a methodology for creating an application capable of producing personalized fonts with DCGANs was described. The performance of the proposed model was evaluated by ANOVA to confirm its precision. In summary, the Learning Rate is found to be one of the governing parameters affecting DyGAN. The proposal was 67% correct, which means that the possible characters may have been written by a human. This proposal is intended to contribute to the creation of dyspraxia in writing expanding resources for learning through data augmentation with potential support for those who have dyspraxia. The architecture is intended for anyone with a writing issue, and in this respect it has an obvious impact on the education building.

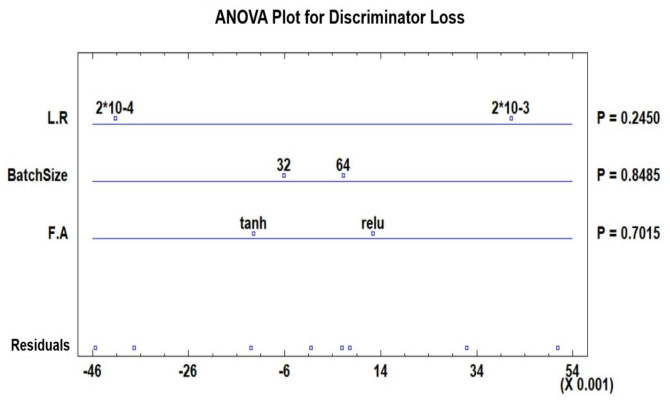


Fig. 6. Analysis of residuals on the variable discriminator loss.

Our application aims to provide a model that is deployable and that can be used effectively by all audiences. Future work for DyGAN can also be a case study to have unique characters from each person in which the model imitates handwriting of particular child, and extension with other models/architects as well including artificial neural networks.

ACKNOWLEDGMENT

The research described in this work was carried out at the Centro de Investigación en Computación (CIC) together with the Escuela Superior de Ingeniera Mecánica y Eléctrica (ESIME) of the Instituto Politécnico Nacional, Campus Zacatenco. It should be noted that this research is part of a degree thesis entitled *DPGAN: Aplicación de DCGAN en la generación de caracteres únicos a través de la inteligencia artificial* supported by *Jesús Sebastián Ruiz Cruz*, work directed by Dr. Jesús Jaime Moreno Escobar.

REFERENCES

[1] C. Missiuna, L. Rivard, and D. Bartlett, "Children with developmental coordination disorder: At home, at school, and in the community," *CanChild Centre for Childhood Disability Research*, 2012.

[2] A. Kirby and D. Sugden, "Understanding, diagnosing, and managing dcd," *London: Routledge*, 2011.

[3] R. Lingam, L. Hunt, J. Golding, M. J. Jongmans, and A. Emond, "The functional impact of developmental coordination disorder and its co-occurring conditions in school-aged children," *Child: care, health and development*, vol. 36, no. 4, pp. 478–487, 2010.

[4] R. H. Geuze, "Characteristics of developmental coordination disorder: Evidence from the literature," *Journal of Child Neurology*, vol. 22, no. 6, pp. 646–658, 2007.

[5] S. Summers, "Dyspraxia and its effect on students," *British Journal of Special Education*, vol. 41, no. 1, pp. 1–8, 2014.

[6] J. Cairney, S. Veldhuizen, and P. Szatmari, "Developmental coordination disorder and its consequences: An overview," *Human Movement Science*, vol. 29, no. 1, pp. 1–3, 2010.

[7] V. F. Pinos Medrano, N. F. Medrano Núñez, and P. Alarcón Salvatierra, "La dispraxia y sus efectos en el aprendizaje," *Dominio de las Ciencias*, vol. 3, no. 2, pp. 380–400, 2017. [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=6325867

[8] B. S. Schifrin and L. D. Longo, "William John Little and cerebral palsy: A reappraisal," *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 90, no. 2, pp. 139–144, Jun. 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030121150000261X

[9] M. Ahearn, "Who Were Orton and Gillingham? | Academy of Orton-Gillingham Practitioners and Educators," Dec. 2016. [Online]. Available: https://www.ortonacademy.org/resources/who-were-orton-and-gillingham/

[10] "Anna Jean Ayres | Sensory Integration, Autism & Dyspraxia | Britannica." [Online]. Available: https://www.britannica.com/biography/Anna-Jean-Ayres

[11] A. Sims, Z. Ali, and K. Meehan, "Exploring the efficacy of an iot device as a sensory feedback tool in facilitating learning for neurodivergent students," in *2023 IEEE World AI IoT Congress (AIIoT)*, 2023, pp. 0755–0761.

[12] B. M, S. K G, A. D. P, S. K. S, S. K. S, S. S, and R. S, "Web based assessment and training model for dyslexia dyscalculia dysgraphia dyspraxia adhd & autism," in *2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2022, pp. 1753–1757.

[13] B. Dhanalakshmi, R. Dhanagopal, D. Raguraman, and T. Thamdapani, "Improving cognitive learning of children with dyspraxia using selection based mid-air gestures in athynos game," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, 2020, pp. 231–237.

[14] D. Avila-Pesantez, L. Vaca-Cardenas, L. A. Rivera, L. Zuniga, and L. Miriam Avila, "Athynos: Helping children with dyspraxia through an augmented reality serious game," in *2018 International Conference on eDemocracy & eGovernment (ICEDEG)*, 2018, pp. 286–290.

[15] G. Dimauro, V. Bevilacqua, L. Colizzi, and D. Di Pierro, "Testgraphia, a software system for the early diagnosis of dysgraphia," *IEEE Access*, vol. 8, pp. 19 564–19 575, 2020.

[16] R. Sperandeo, L. L. Mosca, Y. M. Alfano, V. Cioffi, A. D. D. Sarno, A. Galchenko, D. Iennaco, T. Longobardi, E. Moretto, S. Dell'Orco, B. Muzii, and N. M. Maldonato, "Preventing dyspraxia: a project for the creation of a computational diagnostic system based on the theory of embodied cognition," in *2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 2019, pp. 511–512.

[17] I. Bortone, D. Leonardis, N. Mastronicola, A. Crecchi, L. Bonfiglio, C. Procopio, M. Solazzi, and A. Frisoli, "Wearable haptics and immersive virtual reality rehabilitation training in children with neuromotor impairments," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 7, pp. 1469–1478, 2018.

[18] I. Bortone, D. Leonardis, M. Solazzi, C. Procopio, A. Crecchi, L. Bonfiglio, and A. Frisoli, "Integration of serious games and wearable haptic interfaces for neuro rehabilitation of children with movement disorders: A feasibility study," in *2017 International Conference on Rehabilitation Robotics (ICORR)*, 2017, pp. 1094–1099.

[19] M. F. Othman and W. Keay-Bright, "Rotoscopy-handwriting prototype: Using computer animation technique to assist the teaching of handwriting for children with dyspraxia," in *2011 Eighth International Conference on Information Technology: New Generations*, 2011, pp. 464–469.

[20] ——, "Using rotoscopy technique to assist the teaching of handwriting for children with dyspraxia," in *2010 Third International Conference on Advances in Computer-Human Interactions*, 2010, pp. 175–178.

[21] N. Gouveia, B. Patrão, and P. Menezes, "Puzzle time — vr runner," in *2017 4th Experiment@International Conference (exp.at'17)*, 2017, pp. 121–122.

[22] S. Tresser, "Case study: Using a novel virtual reality computer game for occupational therapy intervention," *Presence*, vol. 21, no. 3, pp. 359–371, 2012.

[23] J. Wang, T. Yu, and Z. Huang, "Integrated transfer learning based on group sparse bayesian linear discriminant analysis for error-related potentials detection," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 2020, pp. 487–492.

[24] S. Selva Nidhyananthan, R. Shantha Selvakumari, and V. Shenbagalakshmi, "Contemporary speech/speaker recognition with speech from impaired vocal apparatus," in *2014 International Conference on Communication and Network Technologies*, 2014, pp. 198–202.

[25] "Generative Adversarial Networks (GAN), una introducción." [Online]. Available: https://es.linkedin.com/pulse/generative-adversarial-networks-gan-una-introducci%C3%B3n-moralo-garc%C3%ADa

[26] "Deep Convolutional GAN - DCGAN - in PyTorch and TensorFlow," Jul. 2021. [Online]. Available: https://learnopencv.com/deep-convolutional-gan-in-pytorch-and-tensorflow/

[27] "dcgan_api_v2_1.md." [Online]. Available: http://personal.cimat.mx:8181/~mrivera/cursos/aprendizaje_profundo/dcgan/dcgan.html

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

Enhanced Cyber Threat Detection System Leveraging Machine Learning Using Data Augmentation

Umar Iftikhar, Syed Abbas Ali
Department of Computer & Information System Engineering, NED University of Engineering & Technology, Karachi, Pakistan

***Abstract*—In the modern era of cyber security, cyber-attacks are continuously evolving in terms of complexity and frequency. In this context, organizations need to enhance Network Intrusion Detection Systems (NIDS) for anomaly detection. Although the existing Machine Learning models are in place to cater to the situations but new challenges emerge rapidly which affects the performance and efficiency of existing models specifically the unreachability of large datasets and unorganized data. This results in degraded efficiency for the identification of complex attacks. In this paper, data augmentation has been done of NSL-KDD which is a standard dataset for Intrusion Detection Systems (IDS) specifically for IoT-based devices. The improvement in performance and efficiency of NIDS has been performed by training the augmented dataset using the K-Nearest Neighbor (KNN) ML model.**

***Keywords*—Anomaly detection; cyber threat intelligence; generative adversarial networks; data augmentation; Wasserstein GAN with gradient penalty**

I. Introduction

This section gives some background regarding the research presented in this paper along with a detailed problem statement followed by the objectives.

A. Background

The domain of Network security has become one of the most immensely important domain in the modern technological landscape because of various challenging threats, numerously increasing the risk factors. In this context, continuous counter-security measures have become the utmost need of modern-day technology that declares war against intruders and stabilizes the computer networks that are responsible for the protection of user information without compromising the entire network. Network Security has provided significant benefits with time in terms of various aspects like evaluation of security problems, practicing attack and defense, and enhancement of confrontation of network information. Network security plays a vital role and always needs room for improvement in terms of the efficiency and accuracy of intrusion detection techniques [10]. The use of wireless technology and the transmission of large amount of information over the networks has significantly increased security issues specifically in the emerging field of IoT. Network security has become the main point of concern and an effective and efficient Intrusion Detection System (IDS) has now become an essential need for traditional and IoT networks that provide countermeasures against rising threats in small and medium enterprises.

Because of these reasons, Intrusion Detection Prevention Systems (IDPS) have become crucial for network security as they play a vital role in tackling threats. IDPS have become more powerful than ever ensuring the security of networks.

B. Problem Statement

The modern-day NIDS has various crucial shortcomings and drawbacks that degrade its ability to effectively handle cyber-attacks. Majorly it lacks in authenticity and timely responsiveness of network threats in existing datasets that are used for training of NIDS. The frequent staling of these datasets is the main reason for identifying new threats as IDS models mainly rely on these datasets.

Another major limitation regarding the current deployment of NIDS is expensive middleware applications that only monitor a certain portion of the entire network, whereas neglects monitoring of other segments. Data augmentation methodologies play a pivotal role in the training of machine learning models. On the other hand, various existing research has claimed the enhancement in the adaptability of models but still, the utilization of these augmentation techniques in cyber security poses problems.

C. Objectives

The main objective of this research is to produce a vigorous cyber threat detection system equipped with the tendency of auto-encoder based data synthesis and attack surface vector modeling [11]. Generative Adversarial Networks (GANs) [2] have the ability to enhance the attack detection performance of NIDS when datasets like NSL-KDD are incorporated into it. Utilization of GANs to produce the synthetic data, NSL-KDD can be extended using more assorted and accurate examples of network traffic that enhance the quality of trained models of the provided datasets.

To authenticate the designed methodology, the NSL-KDD dataset has been selected as a vigorous baseline for this research. The dataset chosen has been proven globally in the field of networks due to its rational evaluation, reproducibility, and trustworthy capability of accomplishing theoretical amendments. The results generated by the selection of the above-mentioned dataset impact a concrete transferable theory that is not affected by dataset-specific limitations. Thus, the framework proposed in this research is more capable of being implemented on complex datasets resulting in the evolution of various network traffic scenarios.

The idea of using generative model approaches and the NSL-KDD in combination will enhance the capability to cater the challenges such as data misbalancing and overfitting in IDS. Training of augmented datasets with denser samples, the

performance can be enhanced against cyber-attacks, thus enhancing the model's efficiency in detecting network threats.

II. Related Work

In this section, all the related work was discussed in detail including NSL-KDD Dataset, various ML Models, GAN for augmentation of data, KNN Model, and gaps in existing research.

A. Overview of NSL-KDD Dataset

The NSL-KDD dataset is an enhanced version of the KDD Cup 1999 dataset that was developed to address the problems like redundancy and class imbalance [12]. It is capable of providing improved and reliable standards for evaluating the performance of Network Intrusion Detection Systems (NIDS). The dataset in KDD Cup 1999 consists of a large number of duplicate instances that create noise while training the machine learning models [14]. Due to this reason, NSL-KDD has an edit advantage in which the majority of duplicated records are considered redundant and are eliminated. This elimination of duplicate records provides more enhanced and refined data of network traffic, improving the capability to assess the performance of NIDS.

There are 41 features in the NSL-KDD that are capable enough to capture the characteristics of network traffic. These characteristics include connection details, timing, transferred bytes, and payload content. These characteristics are further divided into structural, content, time-based, and host-based groups that help in attaining improved statistical and machine-learning methodologies used for intrusion detection. The enhanced features of KDD-NSL such as addressing redundancy and balanced class distribution, improve the capabilities to detect common and rare attacks in NIDS [3][4]. Vast research has been carried out using this dataset for the development and testing of new algorithms, which further makes this dataset more relevant in the field of network security and intrusion detection.

B. Machine Learning Models for Network Security

Machine Learning Models are now playing a vital role in the domain of network security as they provide capabilities like detection of more unusual activities, attacks, and adjustments of cybersecurity systems. There are a large number of models for anomaly detection, some of the most popular used models are Auto encoders, Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Random Forest Classifiers, Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), Generative Adversarial Networks (GANs).

Training the Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) data augmentation [9] with the KNN model gives an enhanced result in the detection of cyber-attacks. The NSL-KDD dataset is being considered as the standard dataset for intrusion detection, sometimes agonizes from class imbalance. WGAN-GP, being a benchmark generative model addresses this issue by creating high-quality synthetic samples for diminished classes. This augmentation improves the diversity of the dataset, guaranteeing the training of ML models on illustrative data. The enhanced features of KNN such as instance classification depending upon neighborhood patterns make a perfect technique for this augmented dataset.

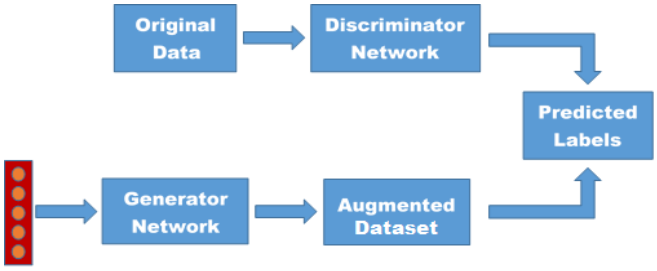


Fig. 1. Workflow of GAN.

The essence of this approach enhances its tendency to meet high accuracy and robustness in defending against cyber-attacks. WGAN-GP produces the data with very less overfittings, sustaining the integrity of the distribution, which is vital in identifying attacks. The flexibility and interoperability of KNN ensure detection in diversified attack scenarios. Combining the WGAN-GP augmented dataset with the KNN model, the system becomes capable of addressing limitations of traditional detection methods like bad generalization and imbalance bias, which leads to a more enhanced cyber-attack detection system for real-world applications.

Generative Adversarial Network (GAN) is one of the most proven model for generative learning that is divided into three main areas which are deep learning, generative models, and adversarial learning techniques [8]. GANs efficiently utilize deep learning networks for solving complex patterns.

The objective function of the original WGAN is given in Eq. (1).

$$\mathcal{L}_{WGAN} = \mathbb{E}_{x\sim pdata}[\boldsymbol{D}(\boldsymbol{x})] - \mathbb{E}_{z\sim pz}[\boldsymbol{D}(\boldsymbol{G}(\boldsymbol{z}))] \quad (1)$$

where,

x is a sample from the real data distribution pdata,

z is a noise vector sampled from a prior distribution pz,

D(x) is the output of the critic for a real sample,

D(G(z)) is the output of the critic for a generated sample.

The Gradient Penalty of WGAN-GP can be mathematically defined in Eq. (2).

$$\mathcal{L}_{GP} = \mathbb{E}_{x\sim pz}[(||\nabla_x \boldsymbol{D}(\hat{\boldsymbol{x}})||_2 - \mathbf{1}^2] \quad (2)$$

where,

$\hat{x}$ is a random sample,

$\nabla_x D(\hat{x})$ is the gradient of the critic D concerning the input,

$||\nabla_x D(\hat{x})||_2$ is the norm of gradient.

The WGAN-GP loss function is expressed in Eq. (3).

$$\boldsymbol{L} = \mathbb{E}[\boldsymbol{D}(\boldsymbol{x})] - \mathbb{E}[\boldsymbol{D}(\boldsymbol{G}(\boldsymbol{z}))] + \boldsymbol{\lambda}\mathbb{E}[\left(||\nabla \boldsymbol{D}(\hat{\boldsymbol{x}})||_2 - \mathbf{1}\right)^2 \quad (3)$$

Eq. (3) is comprised of two parts; in the first part, the critical loss is referred while in the second part is related to the WGAN gradient penalty. This ensures maintaining stability while

training. The diagram in Fig. 1 explains the architecture of a WGAN-GP [13], presenting the generator and discriminator's roles in the production of high-quality synthetic data while differentiating it from real data.

C. Generative Adversarial Network in Data Augmentation

When it comes to data augmentation, GANs are proven to be the most relevant technique that increases the range and volume of data serving the machine learning model.

GANs can produce suitable data distribution over the original dataset. Also, it can compensate the unusual situations where data is limited or unbalanced. Synthetic data, refers to different samples that are not present in data, for example, other viewpoint images in image data or samples of some neglected categories in categorical data. This permits to enhancement of the dataset by inculcating realistic synthetic samples and GANs in addition improves the overall performance of machine learning models and provides a better ability for generalization of data as illustrated in Fig. 1. GAN usage for augmentation of data is extensive and produced various advancements in multiple fields by synthetic data creation [6][7]. They can also be deployed in the analysis of network traffic, where they can be utilized in producing synthetic data that aids in creating enhanced models and is capable of identifying abnormalities in network systems.

D. K-Nearest Neighbor (KNN)

K-Nearest Neighbors (KNN) is a non-parametric, controlled learning algorithm that is popular for its classification and regression tasks. This algorithm functions by detecting the 'k' nearest data points known as neighbors within the featured space over a given input point. Also, it provides predictions either based on the majority class for classification or the average of their values for regression. Euclidean, Manhattan, or Minkowski distance metrics are being used to calculate the distance between two points.

KNN is specifically beneficial in the detection of abnormality as it tends to detect data points that significantly diverge from their nearest neighbors. Abnormalities usually appear as data points with lesser or more detached neighbors as compared with normal data points. KNNs are very useful in multidimensional data handling, implementation simplicity, and flexibility.

The mathematical equation of KNN can be given in Eq. (4):

$$\boldsymbol{d}(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{x_i} - \boldsymbol{y_i})^{\boldsymbol{2}}} \quad (4)$$

where,

d is the distance between two points **x** and **y** in an n-dimensional space.

The formula for calculating the anomaly score is given in Eq. (5).

$$\boldsymbol{Anomaly\ Score}(\boldsymbol{x}) = \boldsymbol{d}\left(\boldsymbol{x}, \boldsymbol{x_{(k)}}\right) \quad (5)$$

where,

$x_{(k)}$ is the k[th] nearest neighbor of (x).

E. Gaps in Existing Research

In the area of GANs, a wide research gap is still there. The research performed in this paper will try to cover these areas. The first and most important part is augmentation of data using GAN. The majority of studies on GAN focused on GAN in a general manner, rather than emphasizing the functionality. This research has been carried out with a focus on how GANs can be utilized extensively to handle new types of data.

The methodology proposed in this paper focuses on practical ability by dimensionality reduction with WGAN-GP and Auto-encoder. The performed research ensures uninterrupted translation to the operational environment. Also, it addresses the real-world scenarios for network traffic data in various ways. This research will result in providing effective ways to process the real-world data, with enhanced features like reduced dimensionality, and augmentation of data, as, the WGAN-GP's framework is specifically trained to produce realistic network traffic.

Observations show continuous improvements in the performance with the proposed methodology that is almost similar to real-world scenarios. Various network environments can be adaptable to proposed design of WGAN-GP as it allows retraining the generator with domain-specific traffic data. For example, this design can be deployed on IoT and cloud-based infrastructures by modifying the training process for reflection of their unique characteristics.

III. Methodology

This section describes the complete methodology applied to the research.

A. Data Description

The dataset used for experimentation is the NSL-KDD dataset. It contains 84,952 entries with 28,318 reserved entries for validation.

While considering the features and labels, the dataset contains similar properties. Every single record in the dataset has several parameters depicting multiple aspects of the network traffic. These features provide help in specific identification of the attacks in the behavior of the network. Preprocessing of data has been performed to identify the classification task to the binary decision [1] that is either "normal" or "attack". Due to this binary transformation, intrusion detection has been more simplified specifying whether the ongoing activity is malicious or not.

B. Data Preprocessing

The preprocessing of data is the most important part as it transforms the data to be used by machine learning algorithms. Other processes that are associated with data preprocessing are data collection, data sanitization, and data normalization. The details of the process include data loading, data splitting, data validation, data testing, handling of numerical and categorical features, and reduction of dimensionality using auto-encoders. The data preprocessing process is further divided into the following steps.

The first step is to import the NSL-KDD dataset, which is divided into two parts. One is referred to as the original dataset

containing 84,952 entries. The other one is referred to as a reserved dataset containing 28,318 entries. The remaining entries are referred to as test datasets for evaluating the model's performance on new data. This technique empowers our research to minimize overfitting effects to train models for any unseen instances.

The next step is scaling the features for numeric features and one-hot encoding for categorical features. In numeric features, the median strategy was used for imputing missing values. This was done to refrain from the mean value of numeric data so that it does not bend towards outliers. In categorical features, a one-hot encoder class was used for each feature. The preprocessing technique was used to assign the use of a column transformer which introduces automation of their application to training and test datasets.

The encoding dimension is constant throughout the experiment, to preserve the true spirit of feature representation and analyze the behavior of varying factors. Out of 100%, a 30% augmentation level has been represented to create a balance between sufficient variability and training data, preserving the integrity of the original labels. Also, diverse representations will be learned by the model without being overwhelmed with noise at 30%. This will further enhance the generalization capabilities. The quality of the synthetic data is evaluated using the Classification Model Utility approach. This method involves training a model on the original dataset and another on the combined original + synthetic dataset, then testing both models on the same validation dataset. Improved performance on the validation dataset indicates the synthetic data's quality and utility [5]. The augmentation level increased to 50% enhances variability without introducing out-of-distribution issues.

The next step is the generation of data. It comprises three phases that include optimization of the constructed model, new synthesized data was produced using WGAN-GP.

The next step is the setup of the WGAN-GP model. This setup comprises two major components, the generator and the discriminator. The generator produces data that resembles with input dataset. The discriminator determines whether the generated data is fake or real. While training the WGAN-GP, numerous iterations were performed through many epochs. The generated data was still transformed and in scaled format, it was necessary to untransformed it to bring it back into its original feature space.

The final step is the merging of synthesized data with existing data.

The hyper-parameters and their functionality are as follows:

- Learning Rate: Controls the weight update per iteration, enhancing the speed and stability of training.
- Num Epochs: Determines the count for seeing the entire dataset by the model of the training process.
- Critic Iterations: Number of updates for the critic (discriminator) per generator update, affecting model stability.
- Lambda GP: It is the coefficient of gradient penalty that ensures that makes sure normalization of the gradient to stabilize the training.

The deviation in hyper-parameters shows a major difference in WGAN-GP performance. Table I depicts the response of the model over the selected algorithm and parameter. The fluctuated parameters during the experiment are shown in Table I.

TABLE I. VARIATION IN HYPER-PARAMETER

Instance	Hyper-Parameter Configuration		
	Learning Rate	***Num of Epochs***	***WGAN Augment %***
X1	2.00E-04	100	30%
X2	1.00E-04	100	30%
X3	5.00E-05	100	30%
X4	2.00E-04	50	30%
X5	1.00E-04	50	30%
X6	5.00E-05	50	30%

C. Performance Metrics

The parameters applied on WGAN-GP for examining the performance are as follows:

1) Accuracy (ACC): It is considered the most important parameter in evaluating the model's performance. This metric evaluates the number of samples for correct prediction over the number of all samples. The formula for calculating this metric is given in Eq. (6).

$$\boldsymbol{ACC} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \quad (6)$$

2) Recall: This parameter refers to the ability of the model to predict positive samples. This is calculated by dividing the number of samples that are categorized as true positive overall positive samples. The formula for calculating this metric is given in Eq. (7).

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad (7)$$

3) Precision: In this parameter, true positive identified the number of samples over several samples that are predicted as positive. Eq. (8) calculates the precision.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad (8)$$

4) F1_score: In this parameter, the recall and precision are combined into a single metric. This is called the harmonic mean of recall and precision. Eq. (9) calculates the F1_score.

$$\boldsymbol{F1\_score} = \frac{\boldsymbol{2\times Precision\times Recall}}{\boldsymbol{Precision+Recall}} \quad (9)$$

5) Matthews correlation coefficient (MCC): This performance parameter is considered the best metric for binary classification. It combines all parts of the confusion matrix. The equation for calculating this metric is given in Eq. (10).

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \quad (10)$$

D. Model Training

While training the models, various ML Models have been considered on encoded information generated from auto-encoder. K-Nearest Neighbor (KNN) has been shortlisted from various models. The parameters for selecting this model were its applicability in classifying the tasks.

Algorithm 1: Machine Learning Model for Predictions

```
Initialize (train_set, validate_set)
Step 1: Import K-Nearest Neighbor (KNN) instances from the scikit-learn library
Step 2:
    If (train_set: == X_train_encoded) then
        Train model on X_Train_Encoded
    Else
        Train model on X_Train_Augmented
    End
Step 3: Predict the imported models on Validate_Set
Step 4: Evaluate the models on the following performance metrics
    a. Accuracy
    b. Precision
    c. Recall
    d. F1 score
    e. MCC
Step 5: Export the performance metrics in the CSV file
End
```

The machine learning model can be categorized into two stages. First is training of the model and second is validation of the model. The step-by-step processing of the research algorithm is presented in Algorithm 1. This algorithm gives a method for comparing the accuracy of the models and various parameters for evaluation using the encoded validation dataset to assess the possibility of identifying cyber-attacks.

K-Nearest Neighbor (KNN) ML model is used in this research. KKN Algorithm is known for its simplicity and instance-based machine learning which makes it suitable for tasks like classification and regression. This algorithm operates in a manner that compares a new data point with its nearest neighbors in the featured space. The "K" is the number of considerable neighbors. The label of the new point is determined by classifying the most common class from corresponding K neighbors. Regression is achieved by averaging the neighbors' values that are used to predict the output.

Another reason for using KNN was its non-parametric feature. In this feature, assumptions were not made regarding data distribution. The other features of this model like simplicity and effectiveness for handling non-linear relationships make this algorithm the most suitable choice for tasks like pattern recognition, recommendation systems, and anomaly detection. However, the performance of KNN depends on the selection of K and the distance metric (e.g., Euclidean, Manhattan), along with the size and quality of the dataset.

E. WGAN-GP Augmented Data Training and Validation

The blending of WGAN-GP synthesized data with the NSL-KDD dataset [30], poses a great impact on the performance of the model. The model training starts with the preprocessing of raw datasets, which is the most important process that ensures that the data is ready for model training. Also, further processes like imputation, scaling, and encoding were performed on the augmented data.

After the completion of preprocessing on augmented data, the next step was to update of machine learning model. The subset of the model used in the training process is replicated for the update process. The training of the model was performed on the same metrics used in the dataset training process. Those are accuracy, precision, recall, F1-score, and Mathews Correlation Coefficient (MCC). Using these metrics, performance was assessed after training the model.

Once the KNN model training is completed, the next step is to apply the trained model to encoded validation data. The metrics for analyzing the performance are the same as in previous steps. The use of the validation model trained from the WGAN-GP augmented dataset is the most important factor in assessing the generalization capabilities of the trained model from the augmented datasets.

By observing the performance of the model on the validation dataset, the quality of the synthetic dataset was accounted for.

IV. Experimental Results and Discussion

Table II shows the performance evaluation of the trained model with original data and augmented data. The datasets were evaluated on five performance metrics as described in Algorithm 1 step 4. The values of performance parameters present significant and effective results. These results can be used to draw a meaningful conclusion.

On the original dataset, the provided metrics reflect the mixed performance of the model. 64.77% of accuracy specifies the classification of the instances in the model is an average of two-thirds. Nevertheless, considering accuracy alone can mislead, especially while dealing with imbalanced datasets. The precision value of 0.7376 depicts that the prediction of a positive outcome by the model is 73.76% correct, resulting in a reduced false positive rate. As far as recall is considered, the 0.6477 value highlights the capturing of actual positive cases by the model up to 64.77%, leaving a significant number of false negatives. The F1 score, which provides a balanced measure between precision and recall, with 0.5594, reflects an imbalance relation between these two metrics. This reflects the struggle of the model to maintain an optimal trade-off between precision and recall. Moreover, MCC which covers all aspects of the confusion matrix, is 0.26467. This depicts the model has limitations in prediction, performing slightly better than random guessing.

With an augmented dataset, the performance metrics show variable performance for various hyper-parameter configurations. These performances are presented in the form of graphs in Fig. 2 to 6. Instance X1 provides average recall at 59.11%, but looking at precision (45.12%) and F1 score (44.69%) gives a high false positive rate, whereas negative MCC (-0.0328) means that the performance of the model is overall weak. On the other hand, instance X2 results as the most improved, enhanced, and efficient model. The highest level of accuracy (79.74%), precision (0.8352), recall (0.7974), and F1

score (0.7988) having the strongest MCC (0.6297), depicts the overall best performance providing balanced and reliable predictions. X3 performance is considered to be an average performance having an accuracy of 60.21%, precision at 0.5895, and recall at 0.6021. However, lower F1_score (0.4886) and MCC (0.0722) show that there is room for improvement.

By analyzing the provided parameters of X4, the performance of the model seems to be poor with an accuracy of 38.21%, precision at 0.3656, and recall at 0.3821. The results of F1_score (0.3728) and negative MCC (-0.3152), indicate the predictions very near to random guessing. The worst performance results can be seen in the X5 instance with very low accuracy (13.09%), precision of 0.1489, and recall (0.1309). The F1_score (0.1373) and highly negative MCC (-0.7379) show extremely unreliable predictions. The X6 instance results same as X1 depicting average recall (0.5899), low precision (0.4625), F1 score of 0.4485, and nearly zero MCC (-0.0321), indicating poor predictive power.

In a nutshell, instance X2 is considered to be the most significant as compared to other hyper-parameter configurations providing a strong balance between precision and recall. X1 and X6 can be considered as average while, X4 and X5 require major improvements.

TABLE II. OVERALL PERFORMANCE OF DIFFERENT HYPER-PARAMETERS WITH KNN AS A CLASSIFIER

Instance	Data Model	Hyper-Parameter Configuration				
		Accuracy	Precision	Recall	F1 Score	MCC
Original Dataset	Original Dataset	64.77%	0.737635	0.647655	0.559395	0.264674
X1	Augmented Dataset	59.11%	0.451186	0.591118	0.446918	-0.03282
X2		79.74%	0.835227	0.797432	0.798767	0.62967
X3		60.21%	0.589498	0.60214	0.488586	0.072174
X4		38.21%	0.365578	0.382094	0.372804	-0.31518
X5		13.09%	0.148879	0.130908	0.137281	-0.73789
X6		58.99%	0.462506	0.589905	0.44854	-0.03214

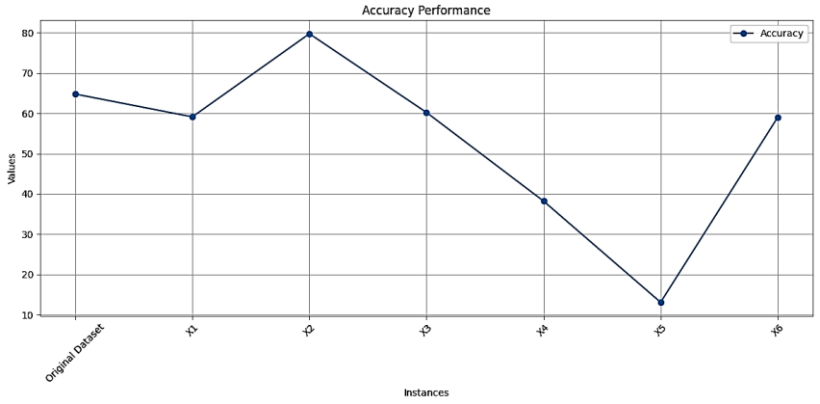


Fig. 2. Accuracy performance.

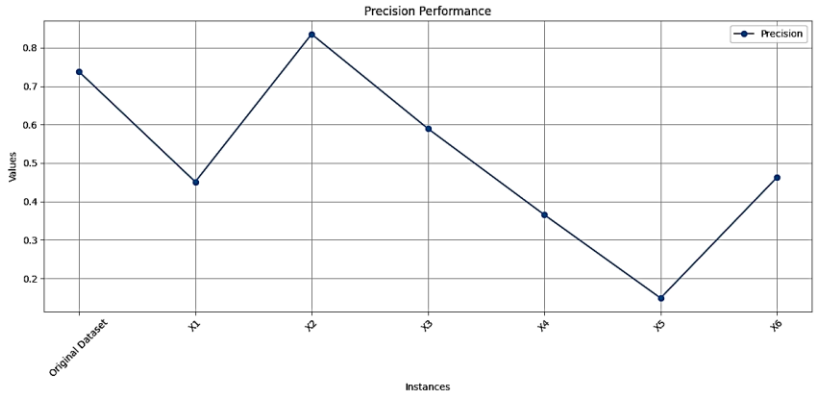


Fig. 3. Precision performance.

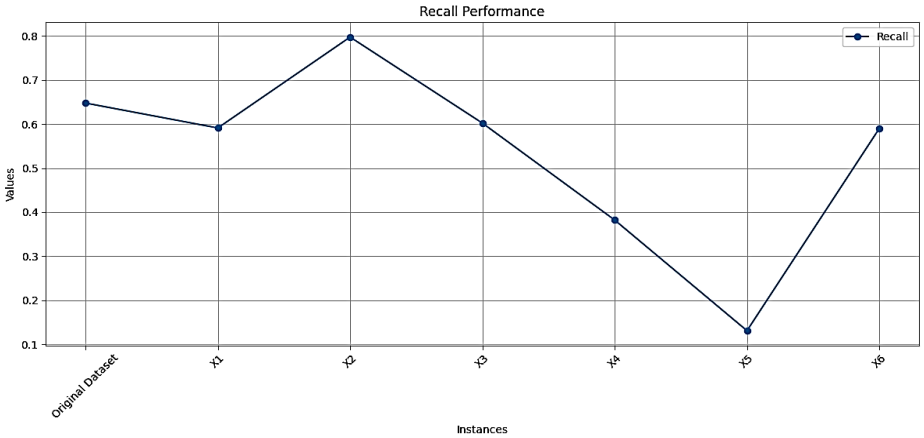


Fig. 4. Recall performance.

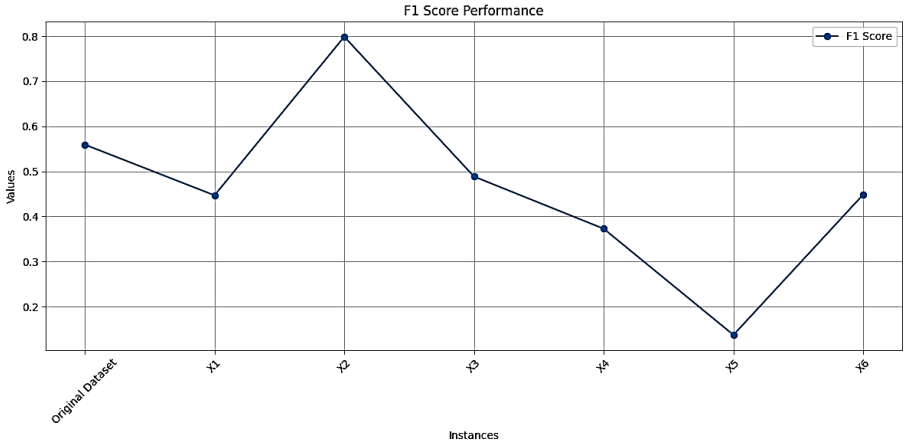


Fig. 5. F1_score performance.

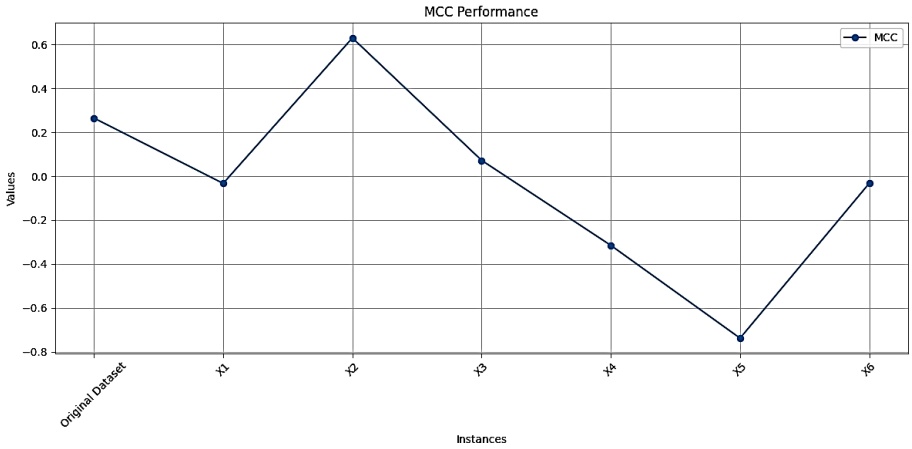


Fig. 6. MCC performance.

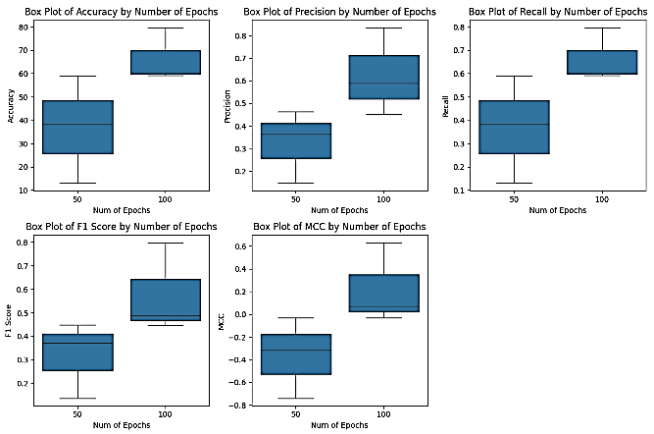


Fig. 7. Correlation heatmap between epoch and hyperparameters.

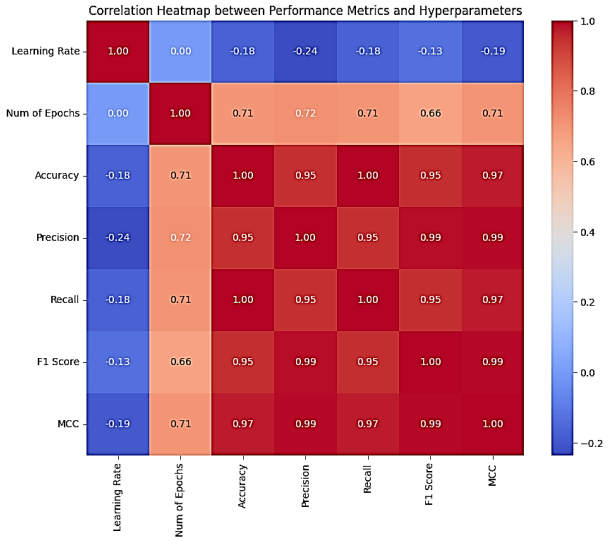


Fig. 8. Correlation heatmap between performance metrics and hyperparameters.

V. Conclusion

The performance analysis of KNN using various hyper-parameter configurations shows unique trends in deciding which configuration is suitable for enhanced performance of machine learning classifier. KNN was found to be more sensitive in selecting hyper-parameters, resulting in great variability, with some configurations resulting in poor performance. Dataset augmentation improves performance in general. However, it is essential to how hyper-parameter values are combined. The optimistic combination of hyper-parameter values for data augmentation will leverage the performance of the machine learning algorithm. Also, from the graphs shown in Fig. 2 to 6, it can be concluded that the increment in epoch value increases the chance of performance as compared to lower epoch values. By analyzing the heat map correlation matrix in Fig. 7 and 8, a value of 0.7 implies a strong positive correlation between the two variables. This concludes that variables have a direct relationship: as one variable increases, the other tends to increase as well.

The research in this paper presents a focused approach to enhancing the capabilities of Network Intrusion Detection Systems (NIDS) through improvement in learning capacity using advanced generative methods like WGAN-GP for augmentation of the dataset. In addition, resource-constrained environments can be considered for future experiments that will help in determining the practical implementation of these techniques in real-time IoT-based applications. Future enhancement of this research can be done by comparing different augmentation techniques on diversified datasets that will enhance the adaptability and robustness of NIDS.

References

[1] C. Strickland, “DRL-GAN: A hybrid approach for binary and multiclass network intrusion detection,” Sensors, vol. 24, no. 9, p. 2746, 2024.

[2] A. Mari, D. Zinca, and V. Dobrota, “Development of a machine-learning intrusion detection system and testing of its performance using a generative adversarial network,” Sensors, vol. 23, no. 3, p. 1315, 2023.

[3] M. Arafah, “Evaluating the impact of generative adversarial models on the performance of anomaly intrusion detection,” IET Networks, vol. 13, no. 1, pp. 28–44, 2023.

[4] E. Goud, “Enhancing DDoS attack detection in SDNs with GAN-based imbalanced data augmentation,” International Journal on Recent and Innovation Trends in Computing and Communication, vol. 11, no. 9, pp. 541–551, 2023.

[5] Z. Wang, C. Han, W. Bao, and H. Ji, “Understanding the effect of data augmentation on knowledge distillation,” arXiv preprint, arXiv:2305.12565, 2023.

[6] S. Bourou, A. Saer, T. Velivassaki, A. Voulkidis, and T. Zahariadis, “A review of tabular data synthesis using GANs on an IDS dataset,” Information, vol. 12, no. 9, p. 375, 2021.

[7] Y. Sun, M. Li, L. Li, H. Shao, and Y. Sun, “Cost-sensitive classification for evolving data streams with concept drift and class imbalance,” Computational Intelligence and Neuroscience, vol. 2021, 2021.

[8] A. Aggarwal, M. Mittal, and G. Battineni, “Generative adversarial network: An overview of theory and applications,” International Journal of Information Management Data Insights, vol. 1, no. 1, p. 100004, 2021.

[9] A. Shafee, M. Baza, D. A. Talbert, M. M. Fouda, M. Nabil, and M. Mahmoud, “Mimic learning to generate a shareable network intrusion detection model,” in 2020 IEEE 17th Annual Consumer Communications & Networking Conference (CCNC), IEEE, 2020, pp. 1–6.

[10] R. Vinayakumar, M. Alazab, K. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, “Deep learning approach for intelligent intrusion detection system,” IEEE Access, vol. 7, pp. 41525–41550, 2019.

[11] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, “Modeling tabular data using conditional GAN,” in Advances in Neural Information Processing Systems, 2019, pp. 7335–7345.

[12] J. Lee and K. Park, “GAN-based imbalanced data intrusion detection system,” Personal and Ubiquitous Computing, pp. 1–8, 2019.

[13] R. Abdulhammed, M. Faezipour, A. Abuzneid, and A. AbuMallouh, “Deep and machine learning approaches for anomaly-based intrusion detection of imbalanced network traffic,” IEEE Sensors Letters, vol. 3, no. 1, pp. 1–4, 2018.

[14] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, “A detailed analysis of the KDD CUP 99 data set,” in Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), Ottawa, ON, Canada, 2009, pp. 1–6.

Data Analytics for Product Segmentation and Demand Forecasting of a Local Retail Store Using Python

Arun Kumar Mishra[1], Megha Sinha[2]
Department of Computer Science and Engineering, University College of Engineering and Technology (UCET), Vinoba Bhave University, Hazaribag -825301, Jharkhand, India[1]
Department of Computer Science and Engineering, Sarala Birla University, Ranchi-835103, Jharkhand, India[2]

***Abstract*—In today's competitive business environment, understanding customers' expectations and choices is a necessity for the successful operations of a retail store. Forecasting demand also plays an important role in maintaining inventory at an optimum level. The work utilises data analytics for product segmentation and demand forecasting in a local retail store. Python is being used as a programming language for data analytics. Historical sales data of a local store has been used to categorise products into different segments. Statistical techniques and a k-means clustering algorithm have been used to understand different segments of the product. Machine learning algorithms and time series models have been used to forecast future sales trends. The business insights allow the retail store to meet customers' expectations, manage inventory at an optimum level and enhance supply chain efficiency. The present work seeks to illustrate how data-driven tactics can enhance operational decision-making in retail.**

***Keywords*—Data analytics; product segmentation; demand forecasting; multicriteria ABC classification; seasonality**

I. Introduction

To ensure business growth and maintain operational efficiency, understanding customer choices and forecasting demand for various products have become important in the current competitive retail environment. The problems a local retail store faces include but are not limited to inventory management, optimisation of sales strategy and fulfilling customers' expectations in changing market trends. In this scenario, product segmentation and demand forecasting help overcome these hurdles to run a successful business.

Based on common characteristics such as sales performance, revenue generation, demand trends and consumer preferences, products are classified into different groups. This process is nothing but product segmentation. It helps retailers customise marketing strategies, optimise inventory, and enhance overall resource allocation. Previous sales data is utilised to predict future demand trends in demand forecasting. It helps retailers make proactive decisions in procurement, inventory management, and supply chain operations.

Python, an object-oriented programming language, provides a rich set of libraries and tools for data analytics. Python provides extensive solutions for addressing intricate retail difficulties, encompassing data pre-treatment, visualisation, machine learning, and time series modelling. Pandas, NumPy, Scikit-learn, and Prophet are particularly adept at product clustering, trend analysis, and prediction modelling.

This article examines the utilisation of Python-based data analytics methods for efficient product segmentation and demand forecasting in a small retail establishment. The study seeks to analyse previous sales data to Determine specific product segments for focused marketing and inventory approaches and construct predictive models to anticipate future demand, reducing stockouts and excess inventory.

This study's findings emphasise that data-driven techniques can enhance decision-making processes in retail, resulting in greater efficiency, customer satisfaction, and profitability.

II. Literature Review

Generally, uniform control measures for all inventory products are inadvisable. The high-value items may be essential to the viability of the firm. The study in [1] talked about ABC and multicriteria ABC analysis. In ABC analysis, products are classified into three classes: A, B and C. Class A items entail significant stock-out expenses and necessitate stringent control measures. It emphasised that multicriteria ABC analysis was crucial for comprehending different product categories: volume drivers, margin drivers, regular movers, and slow movers. In multicriteria analysis, categories A_B and A_C denote volume driver items, B_A and C_A indicate margin driver items, categories B_B, B_C, and C_B imply regular items, category C_C reveals slow-moving items, and A_A encompasses both margin and volume driver items. The research performed multicriteria ABC analysis on the online retail dataset utilising data analytics methodologies. The study in [2] proposed using a three-phased Multi-Criteria Inventory Classification (MCIC) integrating the Analytical Hierarchy Process (AHP), Fuzzy C-Means (FCM) algorithm, and a newly proposed Revised-Veto (RVeto) phase to adhere to the ABC Classification principles and enhance its application and adaptability. Classification based on several criteria is essential to meeting management's needs in the current context. The study in [3] presented a semi-supervised explainable methodology that integrated semi-supervised clustering with explainable artificial intelligence. The semi-supervised method integrated intelligent initialisation with a constrained clustering process that directed the classification procedure towards Pareto-distributed items. At the

same time, explainable artificial intelligence was employed to generate comprehensive micro and macro explanations of inventory categories at both the item and class levels. Implementing the suggested method for the automatic classification of chemical items within a distribution organisation has demonstrated its efficacy in delivering precise, transparent, and thoroughly elucidated ABC classifications. The study in [4] presented an optimal multi-criteria ABC inventory classification for supermarkets to manage commodities based on unit price, lead time, and annual usage. Of the 442 objects, 30 were categorised as group "A," 31 as group "B," and 27 as group "C" under the new ABC classification; nevertheless, all these things were categorised in group "A" in the conventional ABC classification. The study in [5] indicated that AI-based methods exhibited more accuracy than multiple discriminant analysis (MDA). The statistical study specified that SVM facilitated superior classification accuracy compared to alternative AI methodologies. This discovery indicated the potential for employing AI-driven methodologies for multi-criteria ABC analysis within enterprise resource planning (ERP) systems. The study in [6] aimed to present a case-based multiple-criteria ABC analysis that enhances the traditional method by incorporating other factors, such as lead time and SKU criticality. It offered greater managerial flexibility. Decisions from instances served as input, with preferences for alternatives represented naturally using weighted Euclidean distances. It facilitated easy understanding for the decision-maker. The study in [7] examined current portfolio models in procurement that categorise purchases into several product classifications. Case studies from two European automotive OEMs and two vehicle industry suppliers and benchmarking interviews at Toyota, Japan, were used to establish a connection between these product categories and various supplier types. Further, it tried correlating the product categories and supplier types with the specification process—specifically, associating the specification types with their respective generators. The study in [8] conducted supplier segmentation within the automobile sector and proposed four techniques for supplier relationships. Additionally, a four-phase approach for analysing, selecting, and managing decisions on a dynamic relationship strategy with suppliers in the automotive sector was delineated. The study in [9] reviewed the available literature, focusing on market conditions, supplier characteristics, buyer characteristics, and the connections between buyers and suppliers. The study in [10] formulated an innovative methodology for supplier segmentation. Fuzzy logic was utilised to divide suppliers in a broiler firm.

The study in [11] performed a comparative analysis of machine learning algorithms for demand forecasting under uncertainty. The research utilised a synthetic dataset. The machine learning algorithms compared were Linear Regression, Decision Tree Regression, Random Forest Regression, Support Vector Machine Regression (SVR), XG Boost Regression on the parameters of Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The study in [12] proposed a model that integrated time series analysis, boosting and deep learning for demand forecasting. It achieved a significant enhancement in accuracy relative to state-of-the-art studies. The testing utilised authentic data from Turkey's SOK Market. The article compared the Decision Tree Classifier, Gaussian Naive Bayes, and K-Nearest Neighbours (KNN). The Gaussian Naive Bayes technique exhibited the greatest accuracy in demand estimation. The study in [13] focused on demand forecasting and consumer satisfaction within the retail sector. It emphasised the importance of precise demand estimation for merchants. The discussion encompassed machine learning methodologies for forecasting product demand. The paper considered variables for prediction, including time, location, and historical data.

III. Present Work

In this paper, product segmentation was performed using ABC and Multicriteria ABC analysis. First, data was collected, and then it was prepared for the segmentation exercise. Then, segmentation was performed using Python's inventorize package. After that, classification algorithms were applied to it, and performance was evaluated. The flow of work is shown in Fig. 1.

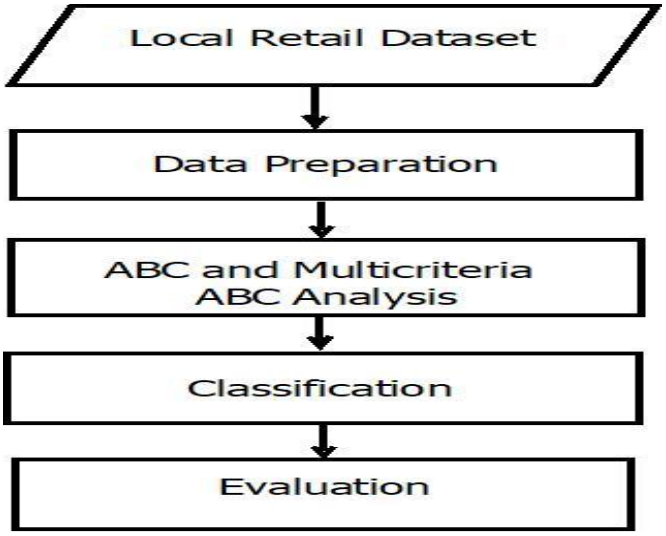


Fig. 1. Workflow for product segmentation.

The same dataset was analysed for demand trends, and a comparative analysis of different machine learning algorithms was done to forecast the demand for A-class products. Fig. 2 shows the workflow for this.

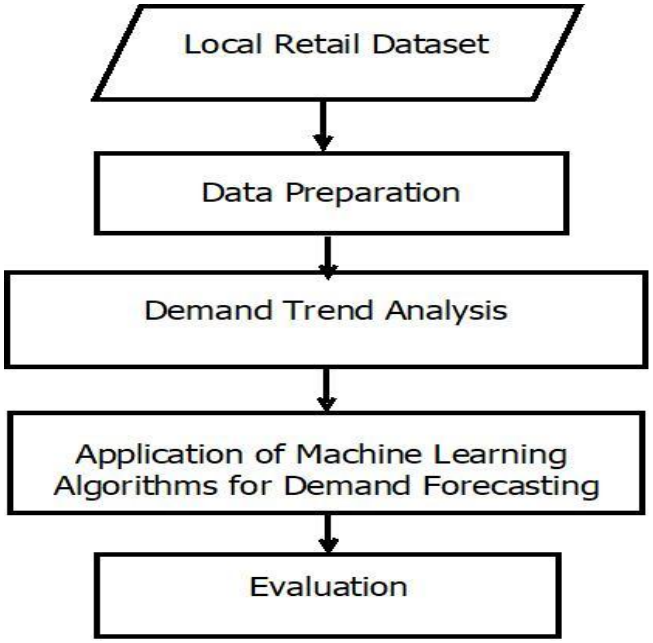


Fig. 2. Workflow for demand trend analysis and forecasting.

Two years monthly sales data of a local retail store situated at Hazaribag was captured and stored in a file named 'Sales_data.xlsx'. The dataset sample is displayed in Fig. 3.

```
[ ] data=pd.read_excel('Sales_data.xlsx')
```

```
[ ] data.head()
```

	Month	SKU	Quantity	Price	Revenue
0	2022-04	Mask	12.0	50.00	600.00
1	2022-04	N95 Mask	12.0	50.00	600.00
2	2022-04	Refill	5.0	508.54	2542.70
3	2022-04	Refilling ABC & Servicing	5.0	508.54	2542.70
4	2022-04	SHOES	12.0	452.38	5428.56

Fig. 3. Local retail dataset sample.

The dataset has five columns: 'Month', 'SKU', 'Quantity', 'Price', and 'Revenue'. It has 4339 records and was analysed for null values and duplicates. After removing records with null values and eliminating duplicates, the dataset comprised 4089 rows. It was further analysed for quantity value. Only those records were kept in which the quantity value was greater than 0. For analysis purposes, one new column, 'Date', was added using the column 'Month', converting it to a datetime object and dropping it for further analysis. Fig. 4. shows the sample prepared dataset. The pertinent columns of the dataset, namely 'SKU', 'Quantity', and 'Revenue', were retained for ABC and multicriteria ABC analysis. Additionally, data was consolidated according to 'SKU'.

```
[ ] data.set_index('Date', inplace=True)
```

```
[ ] data.head()
```

Date	SKU	Quantity	Price	Revenue
2022-04-01	Mask	12.0	50.00	600.00
2022-04-01	N95 Mask	12.0	50.00	600.00
2022-04-01	Refill	5.0	508.54	2542.70
2022-04-01	Refilling ABC & Servicing	5.0	508.54	2542.70
2022-04-01	SHOES	12.0	452.38	5428.56

Fig. 4. Sample local retail dataset.

The inventorize module in Python was utilised to conduct an ABC analysis based on volume. Further, the ABC analysis was performed on revenue. After that, multicriteria analysis based on 'Revenue' and 'Quantity' was performed on the dataset. The dataset obtained after this analysis was used to perform a comparative study of machine learning algorithms viz. KNN, Decision Tree, Random Forest and Naïve Bayes algorithm for classification. The product mix categorisation was treated as the labelled output, while the rest of the columns were used as the basis for classification. First of all, data was split into training and test data. The test size of the data was kept at 20% of the total data. The models were trained and then tested. Confusion matrices were plotted for each model, and accuracy scores were calculated.

As shown in Fig. 4, the prepared dataset was used to understand the demand trend. A graph was plotted for monthly sales over time to know monthly sales trends. Seasonal decomposition was performed to learn more about seasonality trends, and a graph was plotted. Further analysis was carried out to compare machine learning algorithms, viz. Linear regression, Decision tree, Random forest, SVR and XG Boost regressor for demand forecasting of class A items of local retail store. Next, demand forecasting was performed using the said machine learning algorithms. The dataset was split into train and test data with 20% data size. A comparison of predictions was done for all these items. The performance metrics included MAE, MSE and RMSE. To visualise the results in a single frame, grouped bar graphs were plotted for MAE, MSE, and RMSE. Then, the graph was plotted to visualise the comparative performance of the machine learning algorithm in this study.

IV. Results and Discussions

The result counts of ABC analysis on volume and revenue are displayed in Fig. 5 and Fig. 6, respectively.

Then, multicriteria analysis based on 'Revenue' and 'Quantity' was performed on the dataset. Fig. 7 displays the product mix count.

The dataset obtained after this analysis was used to perform a comparative study of machine learning algorithms, viz., KNN, Decision Tree, Random Forest, and Naïve Bayes algorithm for classification. Fig. 8 shows the confusion matrices for these algorithms.

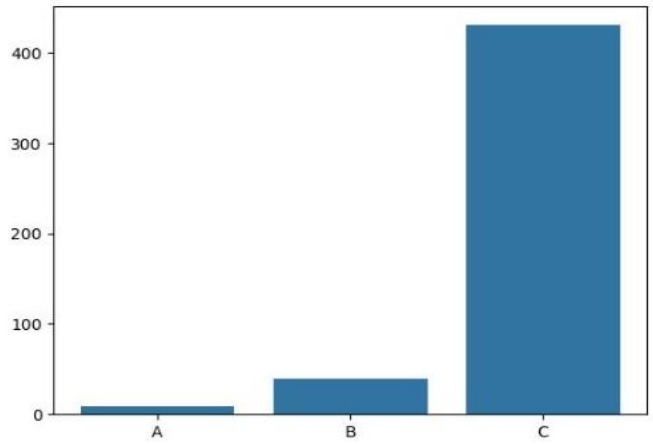


Fig. 5. ABC Count by volume.

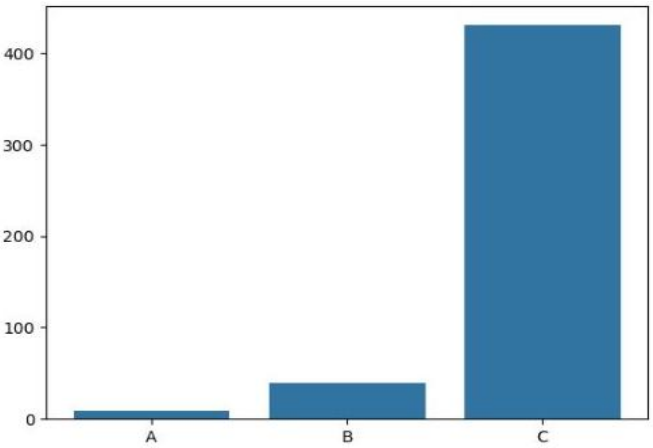


Fig. 6. ABC Count by revenue.

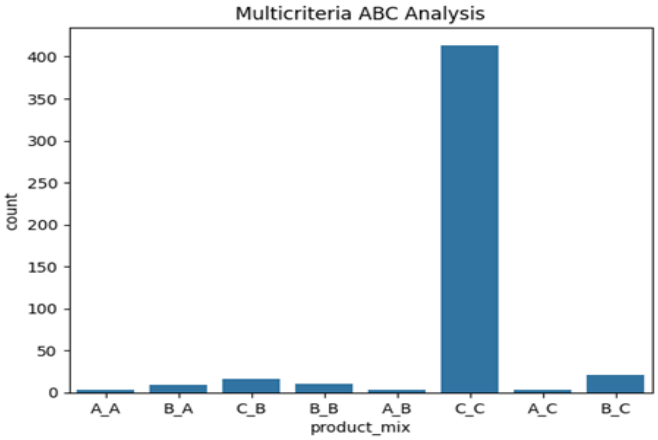


Fig. 7. Multicriteria ABC analysis.

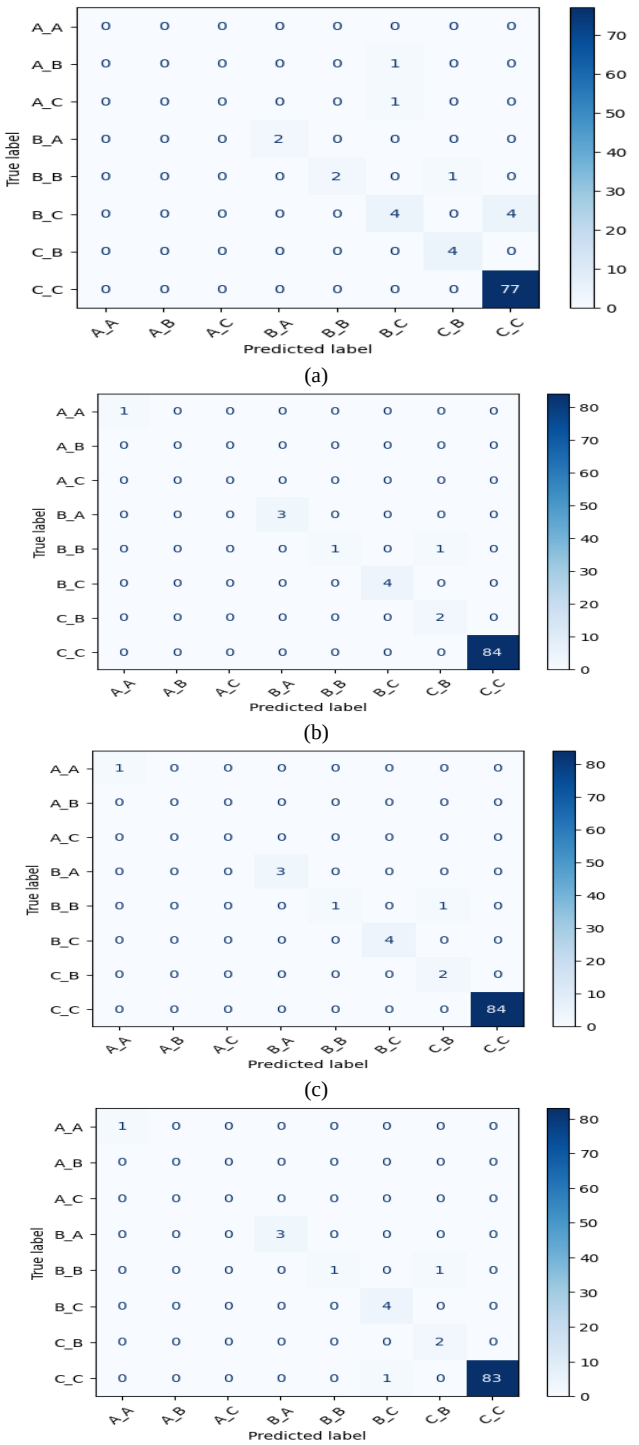


Fig. 8. Confusion matrix (a) KNN (b) Decision Tree (c) Random Forest and (d) Naïve Bayes classification algorithms.

The comparative chart for the accuracy scores has been displayed in Fig. 9.

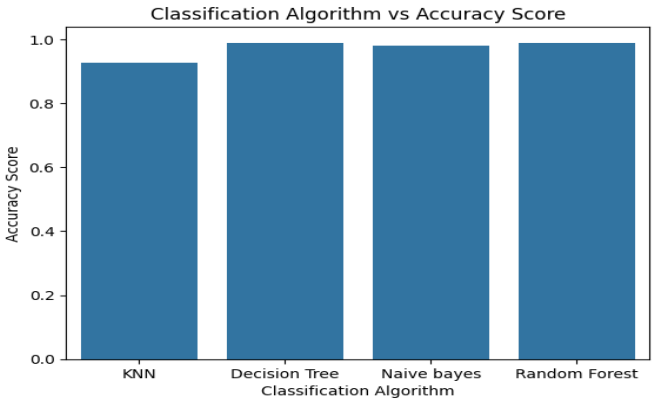


Fig. 9. Accuracy scores comparison for classification.

Decision Tree and Random Forest classification algorithms with identical accuracy scores outperformed KNN and Naïve Bayes algorithms.

Next, a graph plotting monthly sales over time shows monthly sales trends. Fig. 10 displays the seasonal decomposition for the same.

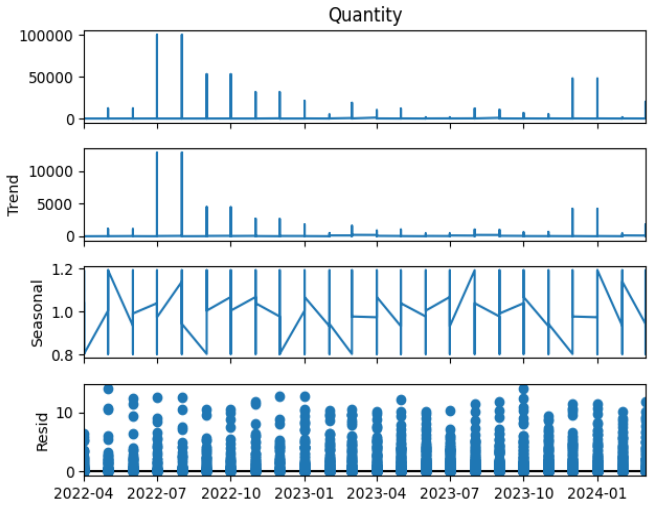
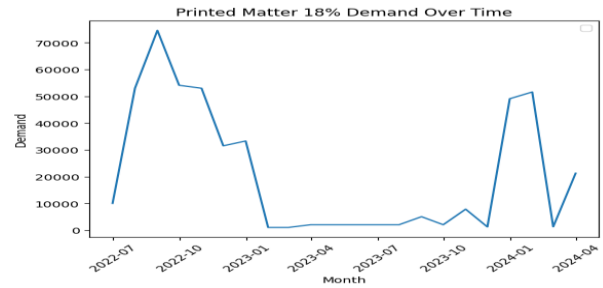
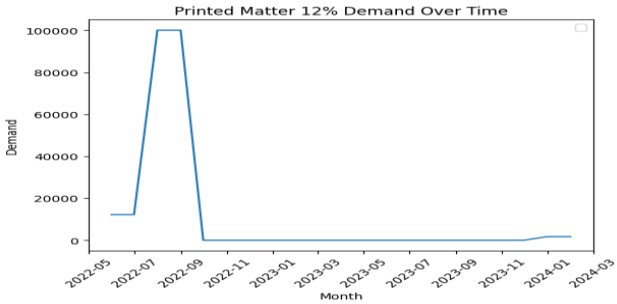


Fig. 10. Seasonality trend for monthly sales data.

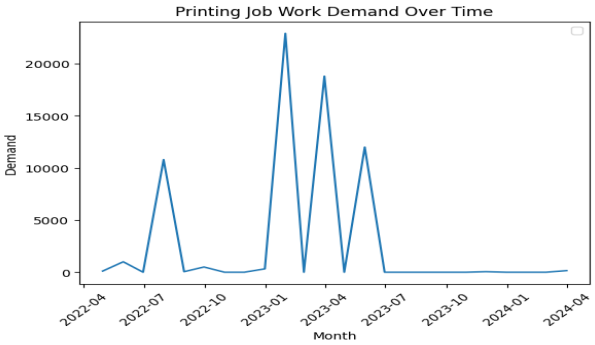
There are irregular spikes in this plot, suggesting an occasional high level of demand at irregular periods. It is clear from the trend plot that there is variability in the dataset. It indicates that the variability is inconsistent enough to make an upward or downward trend over time. The seasonal component of the graph indicates a continuous seasonal influence. However, that influence is very thin. It is shown by minor variations in numbers within a recurring period. The residuals in the plot are scattered, showing greater fluctuations during instances of spikes in the data. It suggests that the spikes may not be described by the trend or seasonality. Further analysis was carried out to compare machine learning algorithms, viz. Linear regression, Decision tree, Random Forest, SVR and XG Boost regressor for demand forecasting of class A items of local retail store. Fig. 11 shows the demand pattern for these items.



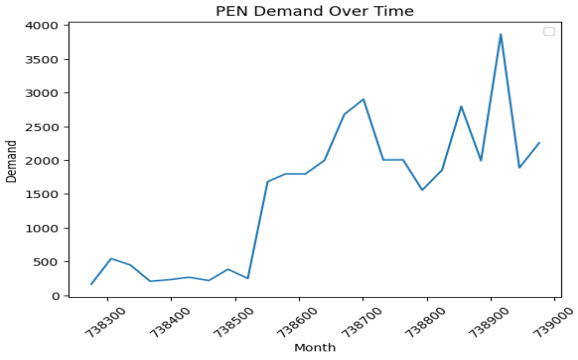
(a)



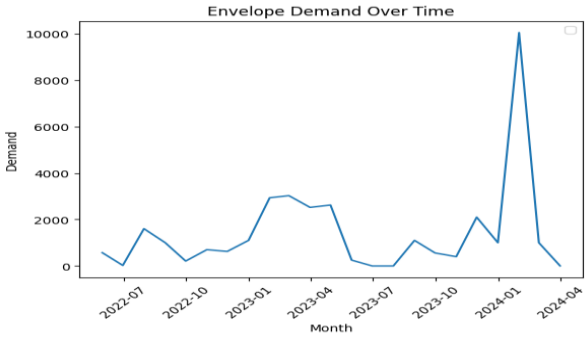
(b)



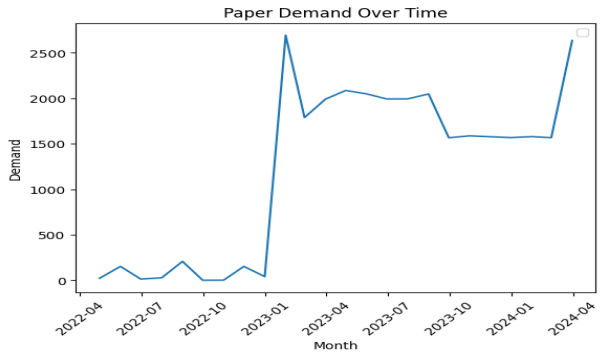
(c)



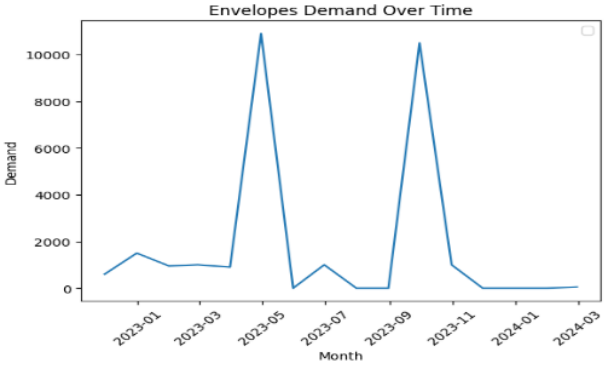
(d)



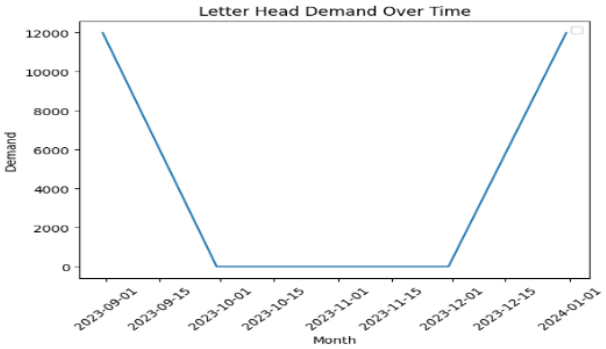
(e)



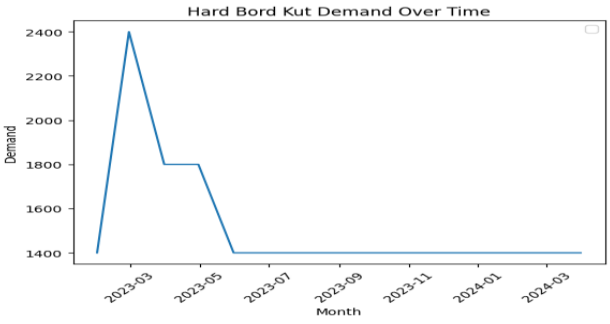
(f)



(g)



(h)



(i)

Fig. 11. Demand pattern for (a) Printed Matter 18% (b) Printed Matter 12% (c) Printing Job Work (d) PEN (e) Envelope (f) Paper (g) Envelopes (h) Letter Head (i) Hard Bord Kut.

Demand for these products was forecasted using machine learning algorithms, and MAE, MSE, and RMSE were calculated. To visualise the results in a single frame, grouped bar graphs have been plotted for MAE, MSE, and RMSE, as shown in Fig. 12, 13, and 14, respectively.

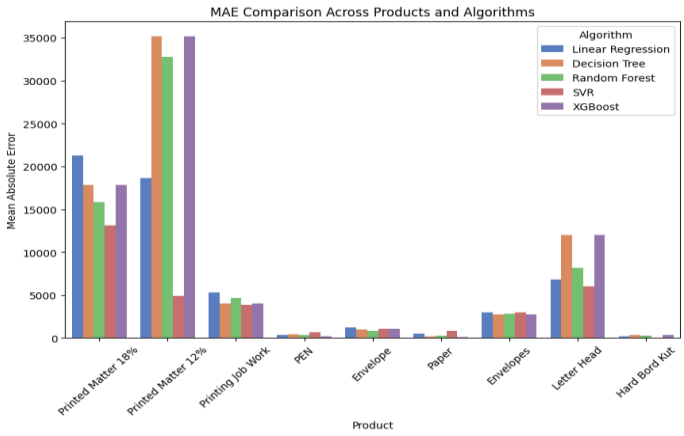


Fig. 12. Comparative performance of machine learning algorithms based on MAE for all class 'A' items.

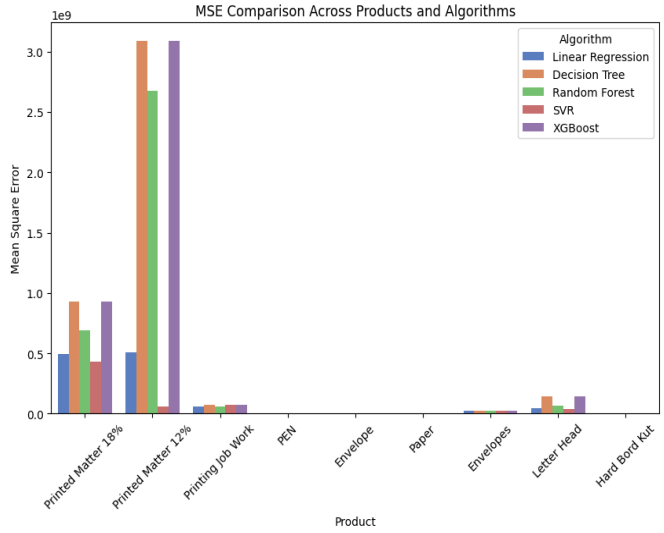


Fig. 13. Comparative performance of machine learning algorithms based on MSE for all class 'A' items.

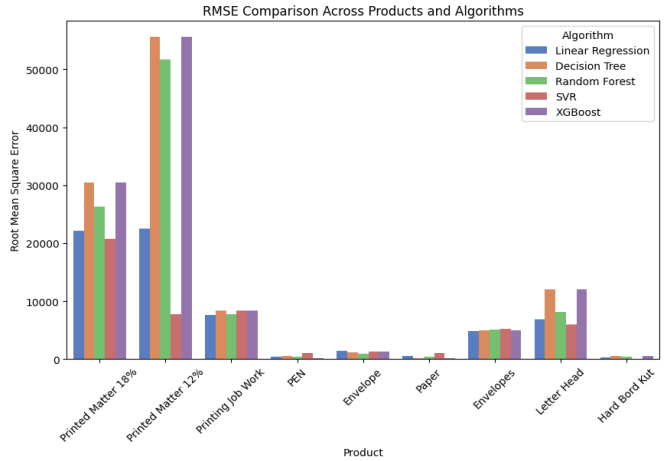


Fig. 14. Comparative performance of machine learning algorithms based on RMSE for all class 'A' items.

Fig. 15 shows the comparative performance of the machine learning algorithm compared in this study.

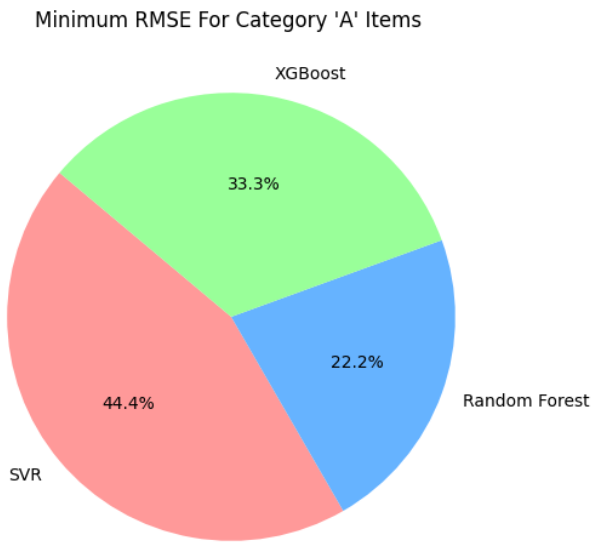


Fig. 15. Comparative analysis of best-performing machine learning algorithms for demand forecasting.

It can be seen that SVR outperformed other algorithms for 44.4% of class A items, XGBoost performed better than other algorithms for 33.3% of items, and Random Forest outperformed other algorithms for 22.2% of class A items.

V. Conclusion

Given that clients seek a diverse range of products while desiring more excellent value for their expenditure, it is imperative to comprehend the several categories of products, including volume drivers, margin drivers, and frequent and slow movers. Multi-criteria ABC analysis serves as an effective tool for conducting this segmentation. This study aims to evaluate the efficacy of classification algorithms in conducting multi-criteria ABC analysis on a retail dataset. Products have been classified based on 'Quantity' and 'Revenue' parameters into A_A, B_A, C_B, B_B, A_B, C_C, A_C and B_C categories. In the contemporary internet company landscape, categorising products and locating providers of vital commodities has become crucial for business viability. Consequently, strategic collaborations may be established for commodities that generate volume and margin. The 'inventorize' module was an effective Python tool for conducting multi-criteria analysis based on quantity and income. It may assist in determining the critical elements to retain. The results indicate that, among the classification algorithms evaluated based on accuracy score, Random Forest and Decision Tree Classifier exhibited comparable performance. Further, a data-driven approach was applied to identify demand trend analysis and demand forecasting on class A items of a local retail store. Machine learning algorithms were also compared to forecast these items. SVR outperformed other algorithms for nearly half of the products in this respect.

References

[1] A. K. Mishra, & M. Sinha. Data Analytics for Multi-criteria ABC Analysis and Supplier Segmentation in Making a Competitive Supply Chain Using Python. In PROCEEDINGS OF 10th INTERNATIONAL SYMPOSIUM ON FUSION OF SCIENCE AND TECHNOLOGY (ISFT-2024) JANUARY 4-8, 2024. https://www.jcboseust.ac.in/assets/files/sovenir_PROCEEDING_ISFT_2024_1.pdf

[2] Fatih Yiğit, Sakir Esnaf, "A New Fuzzy C-Means and AHP-Based Three-Phased Approach for Multiple Criteria ABC Inventory Classification." IMSS'19 Sakarya University - Sakarya/Turkey, 9-11 September 2019, pp. 633-642

[3] A. A. Qaffas, M. A. B. Hajkacem, C. -E. B. Ncir and O. Nasraoui, "Interpretable Multi-Criteria ABC Analysis Based on Semi-Supervised Clustering and Explainable Artificial Intelligence," in IEEE Access, vol. 11, pp. 43778-43792, 2023, doi: 10.1109/ACCESS.2023.3272403.

[4] Aregawi Yemane, Alehegn Melesse Semegn and Ephrem Gidey," ABC Classification for Inventory Optimization (Case Study Family Supermarket)", Industrial Engineering & Management, Research - (2021) Volume 10, Issue 5, ISSN: 2169-0316

[5] Min-Chun Yu (2011). Multi-criteria ABC analysis using artificial-intelligence-based classification techniques. Expert Syst. Appl.. 38. 3416-3421. 10.1016/j.eswa.2010.08.127.

[6] Ye Chen, Kevin W. Li, D. Marc Kilgour, Keith W. Hipel, A case-based distance model for multiple criteria ABC analysis, Computers & Operations Research, Volume 35, Issue 3, 2008, Pages 776-796, ISSN 0305-0548,https://doi.org/10.1016/j.cor.2006.03.024.

[7] R. Nellore, and K. Söderquist (2000) 'Portfolio approaches to procurement: Analysing the missing link to specifications', Long Range Planning, 33, 245-267.

[8] G. Svensson (2004) 'Supplier segmentation in the automotive industry: A dyadic approach of a managerial model', International Journal of Physical Distribution and Logistics Management, 34, 12-38.

[9] M. Day, G. M. Magnan and M. M. Moeller (2010) 'Evaluating the bases of supplier segmentation: A review and taxonomy', Industrial Marketing Management, 39, 625-639.

[10] J. Rezaei and R. Ortt, (2013) 'Multi-criteria supplier segmentation using a fuzzy preference relations based AHP', European Journal of Operational Research, 225, 75-84.

[11] Arun Kumar Mishra, Megha Sinha, & Sudhanshu Kumar Jha. (2024). Comparative analysis of machine learning algorithms for demand forecasting under uncertainty. Computer Science & IT Research Journal, 5(8), 1817-1827. https://doi.org/10.51594/csitrj.v5i8.1409.

[12] Z. H. Kilimci, A. O. Akyuz, M. Uysal, S. Akyokus, M. O. Uysal, B. Atak Bulbul & M. A. Ekmis (2019). An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain. Complexity, 2019(1), 9067367.

[13] A. I. Arif, S. I. Sany, F. I. Nahin, & A. S. A. Rabby (2019, November). Comparison study: product demand forecasting with machine learning for shop. In 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 171-176). IEEE.

YOLOv7-b: An Enhanced Object Detection Model for Multi-Scale and Dense Target Recognition in Remote Sensing Images

Yulong Song, Hao Yang, Lijun Huang, Song Huang
School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China

***Abstract*—To address the challenges of dense object distribution, scale variability, and complex shapes in remote sensing images, this paper proposes an improved YOLOv7-b model to enhance multi-scale target detection accuracy and robustness. First, deformable convolution (DCNv2) is introduced into the YOLOv7 backbone to replace the standard convolutions in the last two ELAN modules, thereby providing more flexible sampling capabilities and improving adaptability to irregularly shaped targets. Next, a Bi-level Routing Attention (BRA) module is integrated after the SPPCSPC module, employing both coarse- and fine-grained routing strategies to focus on densely distributed targets while suppressing irrelevant background. Finally, training and evaluation are conducted on the large-scale DIOR remote sensing dataset under unified hyperparameter settings and evaluation metrics, allowing a systematic assessment of the overall model performance. Experimental results show that, compared with the original YOLOv7, the improved YOLOv7-b achieves significant enhancements in Precision, Recall, mAP@0.5, and mAP@0.5:0.95, with mAP@0.5 and mAP@0.5:0.95 reaching 85.72% and 66.55%, respectively. Visualization further demonstrates that YOLOv7-b provides stronger recognition and localization for densely arranged, small-scale, and morphologically complex targets, effectively reducing missed and false detections. Overall, YOLOv7-b delivers higher detection accuracy and robustness in multi-scale remote sensing target detection. By combining deformable convolution with a dynamic sparse attention mechanism, the model excels in detecting highly deformable objects and dense scenes, offering a more adaptive and accurate solution for small-target detection, dense target recognition, and multi-scale detection in remote sensing imagery.**

***Keywords*—YOLOv7-b; remote sensing images; object detection; deformable convolution; bi-level routing attention; multi-scale**

I. INTRODUCTION

Optical remote sensing images [1] are typically captured by satellites or high-altitude aircraft from a nadir viewing angle. Compared with ground-based images, they exhibit significant differences in imaging modes and resolutions. With the rapid development of remote sensing technology and the continuous improvement of imaging accuracy, these images have demonstrated broad application prospects in fields such as military reconnaissance, disaster assessment, environmental monitoring, and urban planning [2] [3]. However, multi-scale remote sensing images often face challenges such as large target scale variations, complex deformations, and severe background noise, which pose higher requirements for the accuracy and robustness of target detection algorithms [4] [5].

Traditional remote sensing image object detection methods encompass template matching [6], shape and texture-based approaches [7], segmentation-based techniques [8], and visual saliency-based methods [9]. These typically rely on predefined rules or object shapes for recognition after feature extraction or segmentation [10] [11] [12]. While effective in specific scenarios, they struggle with the diverse appearances of objects in complex backgrounds and numerous types in remote sensing images. Additionally, these methods are susceptible to noise and occlusion, limiting their applicability in large-scale and multi-scenario contexts.

In recent years, deep learning techniques have been extensively utilized in remote sensing image object detection due to their powerful feature learning capabilities [13] [14] [15]. Object detection algorithms can be divided into two-stage detectors (e.g., Faster R-CNN, Mask R-CNN) and one-stage detectors (e.g., YOLO series, SSD, FCOS) based on whether candidate regions are generated. One-stage detectors, which offer faster inference speed and lower computational cost, are more suitable for remote sensing scenarios with high real-time requirements, while two-stage detectors generally achieve higher accuracy at the cost of greater computational overhead [10]. Overall, the end-to-end training paradigm of deep learning is more adaptable to the diverse target distribution characteristics of remote sensing images, enabling a single model to detect multiple types of objects across different scenarios.

Against this backdrop, various improved algorithms have emerged to enhance the accuracy and efficiency of object detection in multi-scale remote sensing images. For instance, Azimi et al. proposed a method combining joint image cascading and feature pyramid networks, using multi-scale convolutional kernels to extract features at different scales and achieving significant accuracy improvements on the DOTA dataset [16]. Deng et al. redesigned the Faster R-CNN architecture by integrating a multi-scale target generation network and a feature fusion module, demonstrating excellent performance on the NWPU VHR-10 and SAR-Ship datasets [17]. Liu et al. introduced an adaptive multi-scale feature enhancement and fusion module, which notably improved detection accuracy on the DOTA and HRSC2016 datasets [18].

For small target detection, Liu et al. enhanced YOLOv2 using a feature-map concatenation strategy to improve detection performance on small-scale targets [19]. Chen et al. proposed a multi-scale spatial and channel attention mechanism to enable

deep neural networks to focus more accurately on key target regions [20]. Ying et al. combined multi-scale convolutional kernels with attention mechanisms to significantly improve detection precision for small targets in complex backgrounds [21]. Additionally, Li et al. developed a fast detection method based on YOLOv3, achieving 93.5% accuracy in multi-scale target detection for high-resolution remote sensing images [22]. Zhang et al. introduced a Dual Multi-Scale Feature Pyramid Network (DM-FPN) tailored for small and densely distributed targets, yielding notable results on the DOTA dataset [23]. Meanwhile, Cheng et al. integrated multi-feature fusion with an attention mechanism in MFANet to achieve high detection accuracy across multiple public datasets [24].

Despite advances in remote sensing image detection, three key challenges remain. First, dense target packing leads to excessive suppression of detection boxes during Non-Maximum Suppression (NMS), causing significant missed detections. This requires better integration of multi-level semantics and contextual information to reduce false suppression from overlapping boxes. Second, objects with large scale differences, such as small vehicles and large facilities in the same image, often result in occlusion of smaller objects by larger ones, exacerbated by limited network focusing and localization capabilities. Lastly, high-altitude imaging results in small targets occupying only a few pixels, increasing the proportion of small targets and the likelihood of missed detections. This necessitates enhanced small-target detection and recognition.

The different comparative results observed across datasets primarily stem from variations in dataset characteristics, including target scale distribution, class imbalance, background complexity, and annotation quality. In remote sensing object detection tasks, the arrangement of objects, morphological variations, and the degree of scene interference significantly impact the performance of detection algorithms. For example, in densely arranged object scenarios, the BRA module's dynamic attention mechanism effectively separates foreground and background information, thereby improving detection accuracy. Meanwhile, in datasets with highly deformed objects, the adaptive sampling mechanism of DCNv2 enhances the model's ability to accommodate geometric deformations, improving target localization precision. As a result, the proposed algorithm tends to perform better on datasets that contain complex deformations and densely packed objects, whereas its performance gain over standard YOLOv7 might be relatively limited on simpler datasets with more uniform target scales and fewer background distractions. Experimental results also indicate that the proposed method achieves a more significant improvement on complex multi-scale datasets such as DIOR compared to certain simpler datasets, further validating its effectiveness in handling challenging remote sensing object detection tasks. Future research can further explore the adaptability of this method across different types of remote sensing datasets and optimize its generalization capability to ensure stable detection performance in diverse scenarios.

To address these challenges, this paper proposes a multi-scale remote sensing image detection approach named YOLOv7-b. It integrates Deformable Convolution (DCNv2) and Bi-level Routing Attention (BRA) into the YOLOv7 framework. Specifically, YOLOv7-b incorporates Efficient Layer Aggregation Network structures (ELAN and E-ELAN) into the backbone to efficiently train deeper networks by managing gradient paths. DCNv2 is employed to capture deformed targets, addressing issues such as scale variability, viewpoint changes, and local deformations. Additionally, the BRA module is inserted at the feature-fusion stage to enhance attention on densely distributed targets and salient features, significantly reducing missed detections in high-density scenes—a common weakness in conventional detectors.

To validate the proposed algorithm, experiments were conducted on the DIOR dataset, which features high resolution, diverse target categories, and wide scene coverage. Results show that YOLOv7-b significantly enhances detection accuracy and inference speed, especially for dense targets like vehicles and ships. Additionally, the model's localization performance was analyzed across various IoU thresholds and compared extensively with the original YOLOv7. The findings indicate that integrating DCNv2 and BRA modules effectively improves the network's feature extraction and dense-target recognition capabilities.

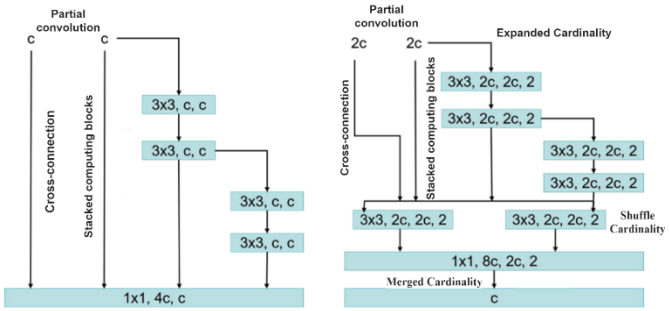
This paper enhances the YOLOv7 architecture with several significant advancements in object detection. The authors integrate Deformable Convolution (DCNv2) into the YOLOv7 backbone to improve feature representation for deformed and multi-scale targets using adaptive offsets and modulation. Additionally, the Bi-level Routing Attention (BRA) module is introduced to address dense target interference and enhance focus on key regions through a sparse attention mechanism. Extensive experiments on the DIOR remote sensing image dataset show that the proposed YOLOv7-b model achieves superior accuracy for multi-scale, high-density, and deformed targets while maintaining high inference speed. These results highlight the model's practical value and potential for real-world applications.

The remainder of this paper is organized as follows: Section II describes the materials and methods, detailing the improved model architecture and experimental setup. Section III presents the experimental results and analysis, including ablation and comparative studies that validate the effectiveness of each module. Section IV summarizes the main findings and contributions, and discusses future work. We hope this research provides valuable insights for remote sensing image target detection and promotes broader applicability in this field.

II. Materials and Methods

A. Model Construction and Improvement

1) Adaptive backbone network: The backbone network of YOLOv7 introduces an efficient layer aggregation structure based on ELAN and E-ELAN. By meticulously controlling the shortest and longest gradient propagation paths within the network, it enables deeper models to learn and converge effectively. The overall framework is shown in Fig. 1.



(a) ELAN (b)E-ELAN

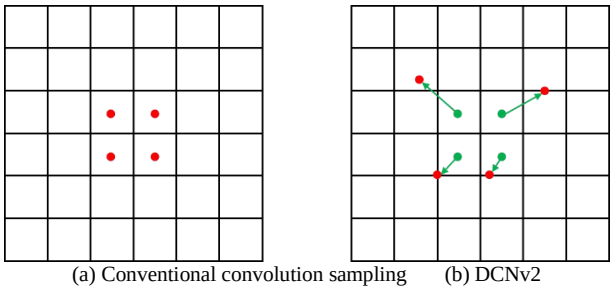
Fig. 1. The structural diagrams of ELAN and E-ELAN.

As can be observed from the figure, the ELAN module employs a gradient path design strategy, which offers multiple advantages. Firstly, adjusting the gradient propagation paths allows the weights of each computational unit to acquire diverse information, thereby achieving higher parameter utilization efficiency. Secondly, since the gradient path directly determines the mode of information transfer and is applied to the weight update process of each computational unit, this strategy ensures that the model maintains a stable learning capability during training and effectively circumvents common degradation issues. Moreover, efficient parameter utilization enables the network to achieve faster inference speeds without the need for additional complex structures, while simultaneously maintaining a high level of accuracy.

However, in larger-scale applications, if computational blocks are stacked without limitation, ELAN can maintain a stable state even under different gradient path lengths or numbers of computational blocks. However, this stability may be broken in extreme stacking situations, thereby leading to a decrease in the efficiency of parameter utilization. To address this issue, researchers further proposed the E-ELAN module, which enhanced the network capability of ELAN through means such as expansion, shuffling, and merging cardinality. It is worth noting that E-ELAN only optimized the internal architecture of the computational blocks, while the structure of the transition layer remained unchanged.

Specifically, E-ELAN increases the number of network channels and the cardinality of computational blocks based on group convolution, and ensures that all computational blocks adopt the same group parameters and channel multipliers. Subsequently, the network shuffles and reassembles the feature maps output by the computational blocks according to the group parameter g, so that the number of channels in the feature maps within each group remains consistent with the initial structure. Finally, the network performs a summation operation on the features of these g groups to obtain the final output feature map. This improvement not only enhances the network's ability to express features, but also maintains efficient parameter utilization.

2) Introduction and improvement of deformable convolutions: One of the key challenges faced in the task of multi-scale remote sensing image detection is the geometric deformation caused by factors such as scale variation, pose variation, and local deformation of objects. To effectively address this issue, Deformable Convolutional Networks (DCNs) introduce learnable offsets on the basis of the standard grid sampling of conventional convolutions, enabling the network to adaptively locate according to the actual shape of the object, thereby significantly improving detection accuracy in object detection tasks. However, DCNs have a fatal problem: while expanding the region of interest, they also include irrelevant areas, thereby weakening the network performance. To solve this problem, the optimized version DCNv2 further enhances the modeling capability of deformable convolution from two aspects. First, by expanding the sampling range of deformable convolution, sampling is allowed in a larger area, thereby improving the ability to learn offsets. Second, a "modulation mechanism" is introduced, which requires each sample not only to learn the offset but also to learn the feature amplitude, thereby flexibly adjusting the network's spatial distribution and the degree of attention to different samples. Fig. 2 shows the difference between the sampling points of ordinary convolution and the sampling points of DCNv2 after the introduction of the modulation mechanism.



(a) Conventional convolution sampling (b) DCNv2

Fig. 2. Comparison of new ordinary convolution sampling and DCNv2 sampling.

Under the action of the modulation mechanism, the deformable network module can not only dynamically adjust the offset of the perceived input features, but also regulate the amplitude of the input features from different spatial locations. In extreme cases, the network module can even set the feature amplitude to zero, thereby completely ignoring the signals from a specific location; this means that the image content from that location will have a significantly weakened or even vanished impact on the network output. Overall, the modulation mechanism provides the network module with additional degrees of freedom, enabling it to better adaptively adjust within the spatial range. Taking the convolution kernel operating on K sampling points as an example, let w_k and p_k represent the feature values of the input feature map x and the output feature map y at position p, then the modulated deformable convolution can be expressed as Eq. (1).

$$y(p) = \sum_{K=1}^{K} w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k \qquad (1)$$

Here, Δp_k and Δm_k represent the learnable offset and modulation parameter at the k-th position, respectively, where $\Delta m_k \in [0,1]$ and Δp_k can be any real number. Compared with the first version of DCNs, the upgraded DCNv2 can not only learn the offset of sampling points but also their weights, thereby

significantly enhancing the flexibility and accuracy of object detection.

Based on the aforementioned advantages, this paper introduces DCNv2 into the YOLOv7 backbone network: while replacing all the 3×3 standard convolutions in the last two ELAN modules, the ELAN-H in the detection head part still retains the original structure of YOLOv7. The adaptive backbone network constructed in this way, under the joint action of ELAN and ELAN-H, can better adapt to targets of different sizes and degrees of deformation, significantly enhancing the feature extraction capability. The overall structure is shown in Fig. 3.

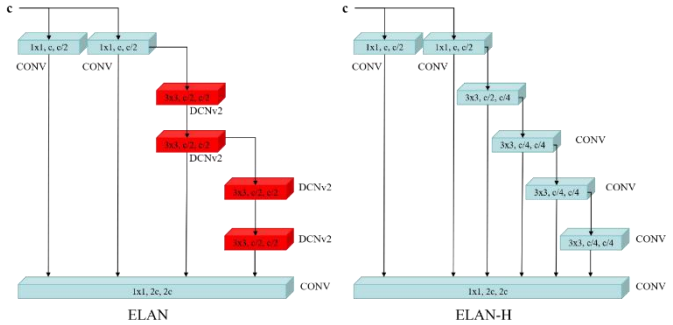


Fig. 3. The structural diagram of the new ELAN and ELAN-H.

B. Bidirectional Routing Self-Attention Mechanism

1) Fundamental principles and implementation of BRA: Identify applicable sponsor/s here. If no sponsors, delete this text box (sponsors).

The main objective of the self-attention mechanism is to enhance the network's ability to focus on key areas. The several self-attention modules mentioned in Chapter 1 all contain pre-set sparse patterns, which are artificially designed. When these modules merge or select key and value tokens using different strategies, these tokens are independent of the query, that is, they are shared by all queries. However, in reality, queries from different semantic regions often focus on key-value pairs with significant differences. Therefore, forcing all queries to focus on the same set of tokens may lead to suboptimal results.

Different from the traditional self-attention module, Bi-level Routing Attention (BRA) is a dynamic, query-aware sparse attention mechanism, aiming to focus each query on a small subset of key-value pairs that are semantically the most relevant. The core idea of BRA is to first filter out the most irrelevant key-value pairs at a coarse-grained regional level, thereby retaining a smaller set of candidate routing regions; then, perform fine-grained token-to-token attention operations within the union of these routing regions. Since the computation process of BRA only involves GPU-friendly dense matrix multiplications, it achieves high performance while also taking computational efficiency into account. The specific steps can be divided into the following three stages:

a) Regional division and input projection: Assuming the input is a two-dimensional feature map $X \in R^{H\times W\times c}$, it is first divided into $S\times S$ non-overlapping regions, with each region containing $\frac{HW}{S^2}$ feature vectors. This process transforms the feature map X into $X' \in R^{\frac{H}{S}\times\frac{W}{S}\times c}$. Subsequently, queries (Q) ,keys (K) and values (V) are generated through linear projection, as shown in Eq. (2)-(4).

$$Q = X'W^q \tag{2}$$

$$K = X'W^k \tag{3}$$

$$V = X'W^v \tag{4}$$

Here, $W^q, W^k, W^v \in R^{c\times c}$ are the projection weights for queries, keys, and values, respectively.

b) Directed routing from region to region: Based on the regional division, BRA constructs a directed graph to capture the relationships between regions. Specifically, the average query (Q') and key (K') for each region are first calculated, i.e., $Q', K' \in R^{S^2\times c}$ .Then, the adjacency matrix representing the inter-regional correlation is calculated using Eq. (5).

$$A' = Q'(K')^T \tag{5}$$

The adjacency matrix $A' \in R^{S^2\times S^2}$ represents the semantic correlation between two regions. To construct a sparse correlation graph, only the top K most relevant connections for each region are retained. Specifically, this goal is achieved by calculating the routing index matrix $I' \in R^{N\times S^2}$, as shown in Eq. (6).

$$I' = topkIndex(A') \tag{6}$$

where the i-th row of I' contains the indices of the top k most relevant regions for region i .

c) Impact of the number of relevant regions on feature aggregation and experimental analysis: In the Bi-level Routing Attention (BRA) module, the number of relevant regions k plays a crucial role in determining how the network aggregates information from different spatial areas. A lower k value restricts the receptive field, limiting the model's ability to capture long-range dependencies, while a higher k value increases computational overhead and may introduce unnecessary noise, reducing detection accuracy. To investigate the sensitivity of BRA to different k values, we conduct an ablation study by varying k and analyzing its impact on detection performance.

The BRA module selectively attends to the most relevant regions, and the choice of k directly affects the amount of contextual information incorporated. If k is too small, the module may fail to capture essential spatial relationships, especially for objects with complex structures. Conversely, if k is too large, the model may aggregate information from irrelevant areas, leading to feature dilution and decreased localization accuracy.

To analyze the trade-off between feature relevance and computational efficiency, we conduct experiments with different *k* values. The results are presented in the Table I:

TABLE I ANALYTICAL FEATURE RELEVANCE AND COMPUTATIONAL EFFICIENCY FOR DIFFERENT VALUES OF *K*

k (Number of Relevant Regions)	mAP@50	mAP@50:95	Inference Speed (ms)
1	69.11%	51.03%	29.77 ms
3	73.11%	51.64%	20.15 ms
5	75.27%	46.35%	27.35 ms
7	68.59%	50.02%	25.68 ms
9	69.16%	52.75%	13.21 ms

The results indicate that using a moderate k value (e.g., k=5) achieves the best balance between detection accuracy and inference speed. When k is too small, the model lacks sufficient contextual awareness, while an excessively large k leads to increased computational cost and potential feature contamination.

Based on our analysis, selecting k in the range of [3, 5] provides optimal performance for most remote sensing detection scenarios. This configuration allows the BRA module to effectively model spatial dependencies while maintaining computational efficiency. Future research could explore dynamic k selection strategies to further enhance the adaptability of the BRA mechanism.

d) Fine-grained attention computation: Utilizing the routing index matrix I', fine-grained attention computation can be performed in the union of the first k relevant regions. For each query token of region i, it only focuses on the key-value pairs in these regions. Specifically, the key and value tensors are first obtained through aggregation operations, as shown in Eq. (7) - Eq. (8).

$$K^g = gather(K, I') \tag{7}$$

$$V^g = gather(V, I') \tag{8}$$

Then, attention computation is conducted based on the aggregated key-value pairs, as shown in Eq. (9):

$$Attention(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{C}})V \tag{9}$$

The final output is given by Eq. (10):

$$O = Attention(Q, K^g, V^g) + LCE(V) \tag{10}$$

Here, $LCE(V)$ is the context enhancement term, which is implemented through depthwise separable convolution with a kernel size set to 5.

In summary, BRA focuses on key-value pairs in the first K relevant windows, significantly reducing the computational load and skipping computations for regions irrelevant to the query. Moreover, since its computational process mainly relies on dense matrix multiplication, it can efficiently utilize GPU resources. The specific steps are shown in Fig. 4, where *mm* denotes matrix multiplication.

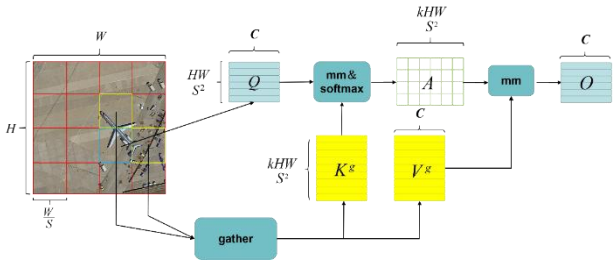


Fig. 4. Schematic diagram of BRA principle.

2) Motivation and justification of the proposed method: The proposed method is specifically designed to address key challenges in remote sensing object detection, particularly the issues related to multi-scale target variations, densely arranged objects, and irregularly shaped targets. To justify the selection of the proposed approach, we first analyze the limitations of existing methods and explain how our modifications effectively resolve these issues.

a) Limitations of existing methods: Traditional object detection models, including standard YOLOv7, face challenges in adapting to highly deformed objects and densely packed target distributions. The fixed receptive field in conventional convolution operations limits the network's ability to flexibly extract relevant features from targets with varying scales and aspect ratios. Moreover, traditional feature aggregation mechanisms struggle to efficiently focus on key areas, leading to misdetections and reduced accuracy in complex remote sensing environments.

b) Motivation for our approach: To overcome these limitations, our method integrates two key improvements:

- Deformable Convolution v2 (DCNv2): Unlike standard convolution, DCNv2 introduces learnable offsets that allow the network to adaptively adjust its sampling positions based on the shape of the target. This flexibility enhances the model's ability to capture spatial deformations, significantly improving feature extraction for irregular objects. Additionally, the modulation mechanism in DCNv2 refines the feature importance assignment, further enhancing detection accuracy.
- Bi-level Routing Attention (BRA): To efficiently handle densely arranged objects, BRA selectively attends to the most relevant regions in an adaptive manner. Instead of applying uniform attention across all areas, BRA dynamically filters and prioritizes semantically significant regions, ensuring improved object differentiation and reducing false positives in crowded scenes.

c) Suitability for remote sensing object detection: The combination of DCNv2 and BRA directly addresses the unique challenges of remote sensing images. Remote sensing targets often exhibit significant variations in scale, shape, and orientation. By incorporating DCNv2, our model gains enhanced spatial adaptability, while BRA strengthens the feature representation of key regions. Experimental results demonstrate that these modifications not only improve detection accuracy but also maintain efficient inference speed,

making the proposed approach highly suitable for real-world remote sensing applications.

3) Fusion strategy: The improved YOLOv7-b model is based on the original YOLOv7, and further optimizes the network's feature extraction and object detection performance by introducing the Bi-level Routing Attention (BRA) module and replacing the Deformable Convolutional Network v2 (DCNv2). The focus of the improvements in this paper is to address the challenges of detecting multi-scale targets, densely arranged targets, and irregular targets commonly found in remote sensing images, thereby enabling the network to perform more excellently in complex remote sensing scenarios.

In YOLOv7, the SPPCSPC module is widely used to alleviate distortion issues caused by operations such as resolution scaling during image processing, while effectively avoiding redundant feature extraction. SPPCSPC can capture global contextual information of the target through multi-scale spatial pooling operations, but its static structure limits the dynamic attention to key regions of the image. Therefore, in this paper, a BRA self-attention module is embedded after the SPPCSPC module to further enhance the network's perception of key target regions. The BRA module is a dynamic query-aware sparse attention mechanism, designed to allow each query point to dynamically focus on areas semantically related to it, rather than uniformly processing all areas. Through this mechanism, BRA first filters out irrelevant areas at a coarse-grained regional level, retaining a set of smaller routing candidate areas. Subsequently, fine-grained token-to-token attention operations are performed on the union of these routing areas, thereby significantly improving the model's ability to capture features of key areas. Moreover, since the computation process of BRA only involves GPU-friendly dense matrix multiplication, despite its operations including area division, routing indexing, and fine-grained attention calculation, the overall computational efficiency remains very high and does not increase the complexity or inference time of the network.

To further enhance the network's adaptability in processing irregular targets, this paper also made improvements to the ELAN module in the backbone of YOLOv7. The original ELAN module uses ordinary convolution to extract features. However, ordinary convolution has limitations when dealing with targets that have significant deformations or irregular shapes. Its fixed convolution kernel sampling pattern makes it difficult to capture complex geometric variation features. Therefore, this paper replaces the ordinary convolution in the ELAN module with DCNv2. DCNv2 is an improved deformable convolution that introduces learned offsets during the convolution sampling process, allowing the convolution kernel to dynamically adjust the sampling positions according to the target shape, thereby achieving adaptive feature extraction. In addition, DCNv2 also incorporates a modulation mechanism. Each sampling point not only learns the offset but also controls its contribution to the final feature through an amplitude factor, which further enhances the focus on key areas and the adaptability to deformed targets. Through these improvements, the ELAN module can extract more accurate and rich feature information when processing irregular targets.

The improvements of YOLOv7-b are mainly reflected in two aspects: First, the BRA module dynamically filters and focuses on semantically relevant areas after the SPPCSPC module, which greatly enhances the network's ability to detect densely arranged targets. In remote sensing images, common densely packed target areas, such as clusters of buildings and rows of vehicles, can be more clearly separated from the background through the fine-grained attention calculation of BRA, avoiding the interference of irrelevant areas on feature extraction. Second, by replacing DCNv2 in the ELAN module of the backbone, YOLOv7-b enhances its adaptability to deformed targets. Targets in remote sensing images usually have complex shapes, scale changes, and irregular distributions. DCNv2 can flexibly capture these deformation features, making the model more efficient in extracting local and global features.

The improved YOLOv7-b performs particularly outstandingly in remote sensing scenarios. On the one hand, the BRA module enables the model to dynamically adjust the distribution of attention, thereby focusing more on densely populated target areas and enhancing the performance in multi-target detection tasks. On the other hand, the introduction of DCNv2 significantly improves the flexibility of feature extraction of the network, especially when dealing with complex deformed targets, the model shows higher robustness. Compared with the traditional YOLOv7, YOLOv7-b not only retains the efficient inference speed, but also achieves further breakthroughs in the performance of remote sensing image target detection through structural optimization, providing a more adaptable and accurate solution for multi-scale and densely arranged target detection.

In summary, the improvements of YOLOv7-b through the combination of the BRA module and DCNv2 enable the model to more accurately capture features, focus on key areas, and enhance the adaptability to irregular targets when dealing with complex scenes and multi-scale targets. This model not only provides new ideas for target detection in remote sensing images, but also offers significant reference value for other dense target detection tasks. Fig. 5 shows the structural diagram of YOLOv7-b

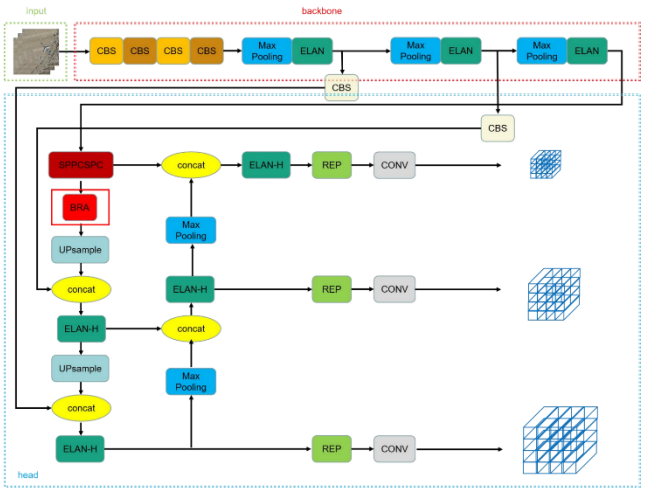


Fig. 5. The structural diagram of YOLOv7-b.

4) Loss function optimization: In this study, the loss function is based on the standard YOLOv7 loss, which consists of three main components: bounding box regression loss, objectness loss, and classification loss. However, to enhance the network's ability to detect remote sensing objects with varying scales and irregular shapes, modifications were made to improve feature learning and localization accuracy.

a) Enhanced bounding box regression loss: To better capture object deformations and varying aspect ratios, we adopt an IoU-aware loss function based on CIoU (Complete IoU) instead of the standard GIoU loss used in YOLOv7. The CIoU loss introduces additional penalty terms related to the aspect ratio consistency and center distance, enabling more precise localization of irregular objects. The modified bounding box regression loss is formulated as follows:

$$L_{box} = 1 - CIoU \tag{11}$$

where CIoU is computed as:

$$CIoU = IoU - \frac{\rho^2(b, b^g) - \alpha v}{c^2} \tag{12}$$

Here, ρ denotes the Euclidean distance between the predicted and ground truth box centers, c is the diagonal length of the smallest enclosing box and v represents the aspect ratio consistency penalty. This modification ensures better localization accuracy, particularly for elongated or irregular targets in remote sensing images.

b) Adaptive objectness loss: The objectness prediction in YOLOv7 is optimized using Focal Loss, which addresses the class imbalance issue by assigning higher weights to hard-to-detect objects. The modified objectness loss is expressed as:

$$L_{obj} = -\alpha_t (1 - p_t)^{\gamma} log(p_t) \tag{13}$$

where p_t represents the predicted objectness score, and α_t and γ are hyperparameters controlling the balance between easy and hard samples. This adjustment helps the network focus more on challenging targets, such as small and densely packed objects in remote sensing imagery.

c) Improved classification loss with label smoothing: To mitigate overconfidence in classification, label smoothing is applied to the classification loss, which is formulated as:

$$L_{cls} = -\sum_{i=1}^{C} y_i' log(\hat{p}_i) \tag{14}$$

where the smoothed label y_i' is given by:

$$y_i' = y_i(1 - \varepsilon) + \frac{\varepsilon}{C} \tag{15}$$

Here, C is the number of classes, y_i is the original one-hot encoded label, and ε is a smoothing parameter. This prevents the model from being overly confident in its predictions, improving generalization on unseen data.

d) Final loss function: The overall loss function for training the improved YOLOv7-b model is defined as:

$$L_{total} = \lambda_{box} L_{box} + \lambda_{obj} L_{obj} + \lambda_{cls} L_{cls} \tag{16}$$

where λ_{box}, λ_{obj}, and λ_{cls} are weight factors that balance the contributions of different loss terms. These modifications collectively enhance the model's ability to detect remote sensing objects more accurately and robustly.

C. Experimental Environment and Settings

1) Introduction to the dataset: This study utilized the DIOR remote sensing image dataset, meticulously collected by experts in the field of earth observation interpretation, to evaluate the effectiveness of the proposed object detection model, as shown in Fig. 6. The dataset comprises a total of 23,463 remote sensing images with a size of 800×800 pixels, covering a spatial resolution range from 0.5 meters to 30 meters, and exhibiting diverse target appearances in various scenes. To ensure the fairness of model training and performance evaluation, the dataset was divided into 5,862 training images, 5,863 validation images, and 11,738 testing images in a ratio of 1:1:2, based on which 190,288 object instances were manually and accurately annotated. The DIOR dataset encompasses 20 common target categories, including Airplane, Airport, Harbor, Bridge, and various sports venues (such as tennis courts, basketball courts, baseball fields, and stadiums). Additionally, it can be observed from the distribution of the dataset that some categories (such as Ship and Vehicle) have a larger proportion of instances, while categories like Dam and Trainstation have relatively few instances, indicating a significant inter-class imbalance. Such a large-scale dataset with multi-scale, multi-resolution, and multi-background scenes not only fully verifies the robustness and generalization ability of the detection algorithm but also provides an important research platform for solving practical problems in the field of remote sensing, such as detecting small-sample categories and identifying multi-object dense scenes.

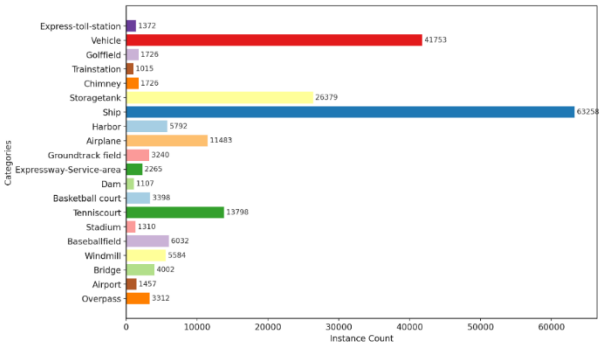


Fig. 6. The distribution diagram of target instances in the dataset.

The DIOR remote sensing image dataset is a benchmark dataset for testing the effectiveness of object detection models, characterized by a large number of images, a rich variety of target categories, and a vast scale of instances, providing ample support for evaluating model performance. The dataset contains 20 target categories, but its distribution is significantly

imbalanced: common categories such as Vehicle and Ship account for the majority of instances, while categories like Trainstation and Express-toll-station are relatively scarce. This distribution characteristic reflects the natural distribution patterns of targets in real-world scenarios, thereby enhancing the model's adaptability in practical applications.

In addition, the target categories of the DIOR dataset cover multiple fields such as transportation infrastructure, industrial scenes, and natural/semi-natural scenes. It includes not only individual targets (such as basketball courts and tennis courts) but also structurally complex targets (such as harbors and airports), which is conducive to comprehensively testing the detection performance of the model in different scenes and target types. The targets in the dataset have a large span in spatial resolution (from 0.5 meters to 30 meters) and significant differences in scale: there are both small vehicles and boats, as well as large cargo ships and dense groups of vehicles, highlighting the challenges of multi-scale target detection.

In addition to the differences in scale, the DIOR dataset also exhibits characteristics such as inter-class similarity (e.g., the morphological similarity between bridges and overpasses) and intra-class diversity (e.g., the varied appearances of vehicles and ships), which pose higher demands on the model's semantic distinction and feature robustness. Overall, the data scale, category diversity, scale variation, and complex backgrounds of DIOR together constitute a realistic and diverse remote sensing testing environment. It can not only be used for a comprehensive evaluation of the model's performance but also ensure the reliability and generalization ability of the model in practical remote sensing applications. Fig. 7 shows examples of some targets in this dataset.

Fig. 7. Examples of target images in the dataset.

2) Experimental settings and evaluation criteria: In the experimental phase, this paper built an advanced object detection model based on the mainstream deep learning framework PyTorch to verify the effectiveness of the proposed method. The flexibility and high efficiency of PyTorch provided a solid technical foundation for the experiments; its excellent support for GPU acceleration also enabled efficient large-scale model training. All experiments were conducted on a high-performance workstation equipped with an Intel Xeon E5-2643 v3 CPU and eight Nvidia Tesla P40 GPUs (each with 24GB of memory), providing ample computing power and memory support for handling complex deep learning tasks.

To ensure the scientific and fair nature of the experiments, all training was completed under the same parameter settings. The total number of training iterations (epochs) was set to 300, the batch size was 16, and the input image size was uniformly adjusted to 640×640. The initial learning rate was set to 0.01 and was gradually adjusted to 0.1 through a learning rate scheduling strategy. Such meticulous parameter design can fully exploit the potential of the model, enabling it to achieve rapid and stable convergence during the training process.

In terms of model performance evaluation, this paper adopts a widely recognized set of indicators, including TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative), Precision (precision), Recall (recall), AP (Average Precision), P-R curve, and mAP (mean Average Precision), to quantify the model's detection capability from multiple dimensions. Specifically, TP represents the number of positive examples correctly predicted by the model, while FP and FN represent the number of false positives and the number of missed positive examples, respectively. On this basis, Precision and Recall measure the accuracy and coverage of the model's positive predictions, and their calculation formulas are shown in Eq. (17).

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \tag{17}$$

The P-R curve takes Recall as the horizontal axis and Precision as the vertical axis, used to dynamically display the relationship between the two. The area under the curve is the AP value of the target category, reflecting the model's average detection accuracy for that category; while mAP, as the core indicator to measure the overall performance of the model, is defined as Eq. (18).

$$mAP = \frac{1}{C}\sum_{i=1}^{C} AP_i \tag{18}$$

The term C represents the total number of target categories, and AP_i is the AP value for the i-th category. The metric mAP is widely used in the field of object detection to comprehensively evaluate the overall performance of a model in multi-category detection tasks.

Typically, mAP is further divided into two forms: mAP@0.5 and mAP@0.5:0.95. The former calculates the average AP value when the IoU threshold is fixed at 0.5, while the latter calculates the average AP value by varying the IoU threshold from 0.5 to 0.95 in steps of 0.05 and then taking the mean. Compared to mAP@0.5, the mAP@0.5:0.95 evaluation standard is more stringent, providing a more comprehensive assessment of the model's performance under different IoU conditions.

In summary, through precise parameter settings and multi-dimensional evaluation indicators, this study has conducted a meticulous and comprehensive validation of the performance of the target detection model. This systematic experimental design not only reflects the high-standard scientific research process,

but also reveals the model's performance under different task conditions with the help of a wealth of indicators, providing solid experimental data and references for subsequent research.

III. Results and Discussion

A. Ablation and Comparative Experimental Results Analysis

In order to verify the effectiveness of the YOLOv7-b model designed in this paper and its various constituent modules, a series of ablation experiments were conducted. These experiments were based on the DIOR remote sensing image dataset described in Section II (C) (1) and utilized the experimental settings consistent with Section II (C) (2), ensuring that all models were compared under the same training and testing conditions. The results of the ablation experiments are shown in Table II, where the detection performance under different module combinations was compared to reveal the specific impact of each module on the overall performance of the model.

TABLE II Ablation Experimental Results of each Module

Method	DCNv2	BRA	mAP@0.5	mAP@0.5:0.95
YOLOv7	√		83.70	63.90
YOLOv7+DCNv2	√		83.82	63.74
YOLOv7+BRA		√	83.91	63.70
YOLOv7-b	√	√	84.25	63.58

The experimental results indicate that the baseline version of the YOLOv7 model achieved mAP@0.5 and mAP@0.5:0.95 of 83.70% and 63.90%, respectively. This represents a high baseline value, reflecting the excellent performance of YOLOv7 in object detection tasks. Upon the introduction of the DCNv2 module, the mAP@0.5 increased from 83.70% to 83.82%. Although the improvement is marginal, it still demonstrates the significant role of DCNv2 in capturing the features of deformed objects. However, the mAP@0.5:0.95 slightly decreased from 63.90% to 63.74%, indicating a certain trade-off in the precise localization of targets by DCNv2.

On the other hand, the YOLOv7+BRA model, with the addition of the BRA module, reached an mAP@0.5 of 83.91%, an improvement of 0.21% over the baseline version. The BRA module, through its sparse attention mechanism, significantly enhanced the model's ability to focus on densely populated targets, particularly excelling in detecting densely arranged objects in complex scenes. However, similar to DCNv2, the BRA module also led to a slight decrease in mAP@0.5:0.95 to 63.70%. This situation may be due to the trade-off in target localization precision as the BRA module strengthens features in dense areas.

When the DCNv2 and BRA modules are combined, they form the complete YOLOv7-b model. The mAP@0.5 of YOLOv7-b reaches 84.25%, which is the highest among all experimental models. This result validates the significant effect of the synergistic action of the two modules in enhancing detection accuracy. DCNv2 improves the network's adaptability to irregular targets, while BRA enhances the model's attention to densely populated targets. However, the performance of YOLOv7-b in terms of mAP@0.5:0.95 is 63.58%, slightly lower than the baseline version. This phenomenon indicates that, although the model excels in dense target detection and overall feature learning, its ability to precisely locate target boundaries under high IoU threshold conditions still needs further improvement.

From the ablation experimental results, it can be seen that each module contributes to the enhancement of detection performance. DCNv2 strengthens the model's ability to extract features of deformed targets, while the BRA module, through dynamic sparse attention, enhances the feature representation of densely populated target areas. However, both exhibit certain performance trade-offs under high IoU thresholds. This is because the enhancement of complex features may increase the model's flexibility requirements in localization tasks, thereby affecting the precise prediction of target boundaries.

To further enhance the overall performance of YOLOv7-b, particularly its mAP@0.5:0.95, future research can optimize the synergistic mechanism of DCNv2 and BRA, reducing the potential interference of complex feature representation on precise localization. Additionally, post-processing techniques such as dynamic IoU threshold adjustment can further improve the model's ability to accurately locate target boundaries in various scenarios. Overall, the experimental results not only validate the effectiveness of the YOLOv7-b model but also provide important insights for future improvements.

In Table II, the mAP@0.5 of the YOLOv7-b model reached 84.25%, the highest among all models, demonstrating the overall effectiveness of the algorithm. However, compared to the mAP@0.5:0.95 of YOLOv7, there was a slight decline. The primary reason is that YOLOv7-b more meticulously learned the feature information of dense areas, detecting more densely packed targets that were previously undetected. However, the model's flexibility in localizing these targets needs improvement. Consequently, when the IoU threshold increases and requires more precise target localization, the model is currently unable to accurately locate the targets, resulting in performance that is inferior to YOLOv7. In subsequent experiments, this paper will further address this issue.

In addition, Table II also demonstrates the effectiveness of each module. By adding DCNv2, the mAP@0.5 reached 83.82%, an improvement of 0.12% compared to the original YOLOv7 codebase; the model structure with only BRA added achieved an mAP@0.5 of 84.03%, an increase of 0.33%.

The specific detection AP results for different target categories corresponding to each method mentioned above are shown in Table III. All comparative methods were conducted under the same training premises and testing conditions. In the table, the first row indicates the detection algorithms used, and the first column represents different target categories, with the numerical results being the AP values for those categories. The last row represents the average AP value across all categories in the dataset, which is the mAP@0.5 result.

TABLE III COMPARISON OF AP RESULTS FOR EACH MODULE

Category	YOLOv7	YOLOv7+DCNv2	YOLOv7+BRA	YOLOv7-b
Express-toll-station	83.3	83.5	77.9	77.3
Vehicle	79.8	80.4	83.3	83.7
Golffield	82.8	83.7	84.3	86.0
Trainstation	70.8	70.9	69.8	70.1
Chimney	92.3	91.8	91.6	92.0
Storagetank	87.9	87.0	86.2	86.0
Ship	91.6	92.3	93.8	95.5
Harbor	64.8	67.4	73.7	75.2
Airplane	91.6	92.9	94.7	95.3
Groundtrack field	90.2	86.5	90.6	87.8
Expressway-Service-area	85.7	86.7	85.4	86.1
Dam	78.8	79.2	76.0	74.3
Basketball court	89.6	90.1	87.9	87.4
Tennis court	90.4	91.3	94.2	94.9
Stadium	93.0	93.4	95.1	96.6
Baseball field	94.1	94.7	95.5	96.1
Windmill	84.9	83.9	85.4	85.0
Bridge	57.1	57.4	57.8	59.0
Airport	88.9	89.2	83.2	84.4
Overpass	76.4	74.1	74.2	72.2
mAP@0.5	83.7	83.8	84.0	84.3

These variations in performance across different datasets highlight the importance of dataset characteristics in evaluating detection algorithms. Our method shows a more significant advantage on datasets with high intra-class variability, complex deformations, and densely packed objects, as the integration of DCNv2 and BRA allows for enhanced adaptability and spatial feature extraction. However, in datasets with relatively uniform target distributions and simpler background structures, the performance improvement is less pronounced, suggesting that the proposed method is more effective in complex remote sensing scenarios. Future work can further analyze the adaptability of this approach in diverse dataset conditions and explore optimizations for broader generalization. The results show that YOLOv7-b, with the addition of the DCNv2 and BRA modules, achieved the best detection performance across several categories, including Vehicle, Ship, Tennis court, Bridge, and Basketball court. Compared to the baseline model YOLOv7, the AP for these categories has seen a small but stable improvement: for example, Vehicle increased from 93.4 to 93.8, Ship from 96.4 to 96.5, and Tennis court from 96.4 to 96.7. These results fully illustrate that the synergistic effect of the DCNv2 and BRA modules significantly enhances the model's ability to extract and represent target features, especially in scenarios with dense targets and complex scenes, where the detection performance has been notably improved.

YOLOv7-b achieved an overall performance metric of mAP@0.5 at 84.3%, which is higher than other comparative models, further validating the effectiveness of its network architecture. Here, DCNv2 helps the model to better adapt to targets with larger deformations, while BRA, through its sparse attention mechanism, increases the model's focus on densely populated target areas. In remote sensing image tasks, vehicles and ships typically appear in small-scale and high-density forms. The combination of DCNv2 and BRA effectively alleviates the phenomenon of missed detections, significantly enhancing the model's robustness and detection accuracy.

These variations in performance across different datasets highlight the importance of dataset characteristics in evaluating detection algorithms. Our method shows a more significant advantage on datasets with high intra-class variability, complex deformations, and densely packed objects, as the integration of DCNv2 and BRA allows for enhanced adaptability and spatial feature extraction. However, in datasets with relatively uniform target distributions and simpler background structures, the performance improvement is less pronounced, suggesting that the proposed method is more effective in complex remote sensing scenarios. Future work can further analyze the adaptability of this approach in diverse dataset conditions and explore optimizations for broader generalization. The reason may be that the basic model YOLOv7 has already had a relatively perfect detection capability for such targets, and the room for the role of the added module is relatively small. At the same time, for categories with relatively few samples in the dataset, such as Train station and Express-toll-station, the model's detection AP has not been significantly improved with the increase of network complexity. This is not only related to the insufficient training samples, but also closely related to the limitation of the resolution of remote sensing images on the clarity of target boundaries.

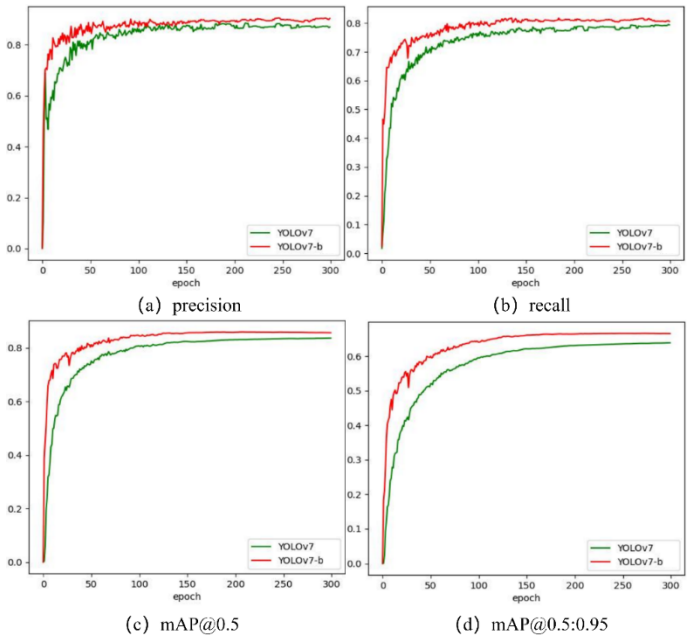


Fig. 8. Comparison of the performance between YOLOv7-b and YOLOv7.

As shown in Fig. 8, YOLOv7-b (red dotted line) exhibits faster convergence rates and superior final performance in four metrics: Precision, Recall, mAP@0.5, and mAP@0.5:0.95, showing significant advantages compared to YOLOv7 (green dotted line). Combining the network structure and experimental data, the following analyses can be made regarding this difference, supplemented with specific quantitative indicators for illustration.

Firstly, the introduced DCNv2 (Deformable Convolutional Networks version 2) and BRA (Bounding Box Re-calibration) modules play a key role in multi-scale feature extraction and fusion. Compared with the original YOLOv7, when the network can more flexibly adapt to the deformation and scale changes of targets at early layers, and adaptively recalibrate key boundaries in subsequent layers, the overall feature representation ability is significantly enhanced. The experimental results show that YOLOv7-b not only continues to lead in Precision in the later stages of training, but also reaches a Precision value of 90.32% at the 300th iteration, which is significantly higher than YOLOv7's 86.93%. This indicator directly reflects the accuracy of target recognition, and its significant improvement indicates that the introduced modules can effectively reduce false detections.

Secondly, the cross-layer connections and multi-scale fusion mechanisms of DCNv2 and BRA significantly optimize the gradient flow of the network, accelerating model convergence. Unlike the original YOLOv7, which only tends to stabilize after 40 to 50 iterations, YOLOv7-b often reaches a convergence state close to the final level at around 20 to 30 iterations. By establishing a stable feature representation space more quickly, the model can continuously and smoothly improve its metrics in the later stages. For example, the Recall of YOLOv7-b reaches 80.55% in the final training iteration, a slight improvement over YOLOv7's 79.38%, indicating that it is also enhanced in capturing more true targets (reducing missed detections).

Thirdly, DCNv2 and BRA balance positive and negative samples to a certain extent, especially for targets with high deformation or complex details. The deformable convolution can adaptively sample local features, and combined with the fine calibration of bounding boxes by BRA, the network can better focus on targets that are originally easy to miss or difficult to detect during training. This mechanism is reflected in the significant improvement of mAP@0.5: in the experimental results, the mAP@0.5 of YOLOv7-b reaches 85.72% at the 300th iteration, an increase of about 2 percentage points compared to YOLOv7's 83.7%. At the same time, in the more challenging metric of mAP@0.5:0.95, YOLOv7-b also achieves a final value of 66.55%, higher than YOLOv7's 63.90%. Since this metric requires accurate regression of target boundaries under various IoU thresholds, the model benefits from the flexibility of deformable convolution in spatial transformation and the enhanced positioning accuracy of BRA, showing stronger stability and accuracy in detecting targets in high IoU intervals.

In summary, the performance advantages of YOLOv7-b mainly stem from the integration of the DCNv2 and BRA modules into the network, resulting in essential improvements over the original YOLOv7 in terms of multi-scale feature fusion, gradient propagation, and localization accuracy. Ultimately, at the 300th iteration, the values of YOLOv7-b in the four main metrics (Precision: 90.32%, Recall: 80.55%, mAP@0.5: 85.72%, mAP@0.5:0.95: 66.55%) are significantly higher than those of YOLOv7, which are 86.93%, 79.38%, 83.7%, and 63.90%, respectively, fully validating the effectiveness of this improvement strategy in enhancing detection accuracy, accelerating convergence speed, and strengthening localization robustness.

In addition to conducting ablation experiments on the YOLO series algorithms, this section also carried out a series of comparative experiments, selecting commonly used algorithms in the field of object detection to compare with the YOLOv7-b improvement method presented in this paper. These include classic algorithms such as Faster R-CNN, SSD, RetinaNet, and CornerNet, as well as YOLOX and the latest YOLOv8 method. All comparison methods were conducted under the same training premises and testing conditions. The detection results for different targets in the dataset are shown in Table III. The first row of the table indicates the detection algorithms used, and the first column indicates different categories, with the numerical results representing the AP of that category. The last row is the average AP value of all categories in the dataset, i.e., the mAP result.

It can be intuitively observed from the table that YOLOv7-b has the most outstanding overall detection performance on the DIOR dataset, with an average detection accuracy of up to 85.7%. Among the 20 common target categories included in this dataset, YOLOv7-b achieves the current optimal detection accuracy in 13 categories, covering both fine-grained small targets such as vehicles (Vehicle) and chimneys (Chimney), as well as larger targets with complex deformations like bridges (Bridge) and airports (Airport). Judging from the distribution of results, YOLOv7-b shows excellent adaptability to multi-scale targets, being able to take into account both the fine details of small targets and the extensive recognition of large targets,

which prominently reflects the robustness and broad applicability of this model under various detection requirements.

In contrast, some early classic detection algorithms (such as Faster R-CNN, SSD) have relatively low overall detection accuracy on the DIOR dataset, especially in small target scenes with high density or fine structures, such as vehicles and tennis courts (Tennis court), where there are obvious problems of missed detections and false detections. This, to a certain extent, reflects the limitations faced by classic detectors in terms of network architecture and feature extraction: due to the lack of deep fusion of multi-scale features and specific optimization for small targets, they fall short in capturing details and suppressing background noise.

Among these traditional algorithms and newer methods, intermediate detection algorithms such as RetinaNet, PANet, CornerNet, and CANet have overall achieved a considerable performance improvement. The detection accuracy of many categories (such as tennis courts, storage tanks (Storagetank), etc.) can reach the level of 70% to 80%. However, these algorithms still exhibit certain missed detections and false detections in scenes with more complex backgrounds or higher similarity in target shapes (for example, harbors (Harbor), toll stations (Express-toll-station)). Such scenes often require a higher level of feature representation capability, as well as more flexible geometric offsets and attention mechanisms, in order to more accurately distinguish between the background and real targets during the object detection process. In contrast, YOLOv7-b performs more robustly in these scenes, indicating that the DCNv2 (Deformable Convolution) module and other improvements introduced in its network structure can better cope with background interference and target deformation.

Regarding the single-stage detectors that have been popular in recent years, YOLOX and YOLOv8 have detection performance on some categories that is not far behind YOLOv7-b, and both perform well in balancing inference speed and accuracy. However, when facing multi-scale and high-density distribution scenes such as vehicles, ships (Ship), airports (Airport), or overpasses (Overpass), YOLOv7-b still has the upper hand in average precision. The key lies in the fact that YOLOv7-b not only inherits the advantage of the YOLO series models in pursuing real-time performance and accuracy, but also enhances the network's feature perception and localization capabilities for small and dense targets by integrating modules such as DCNv2 and BRA (Sparse Attention Mechanism). DCNv2 can achieve spatially adaptive offsets during the convolution process, possessing more flexible learning capabilities for targets with larger deformations, while BRA provides a stronger focusing effect in dense target areas, thereby reducing false and missed detections.

Further analysis targeting different categories also confirms the aforementioned advantages: in high-density or small target categories such as vehicles, storage tanks, and tennis courts, YOLOv7-b's detection capability is particularly prominent, capable of accurately capturing the boundaries and detail information of targets; in targets with large deformations or scale spans, such as bridges, airports, and stadiums (Stadium), YOLOv7-b also demonstrates excellent detection accuracy, indicating that its network can better balance the precision and robustness of multi-scale target localization in the process of integrating shallow details with deep semantics. In addition, in complex background or scenes with strong noise interference factors, such as overpasses and toll stations, the modulation mechanism and sparse attention introduced by YOLOv7-b effectively help the model separate real targets from the background, further reducing the false detection rate and missed detection rate.

In summary, the results in Table IV fully demonstrate the obvious advantages of YOLOv7-b over other mainstream detection algorithms on the DIOR dataset, covering the needs of various scenarios including small targets, large targets, complex backgrounds, and multi-scale targets. Building on the original design concept of the YOLO series that balances speed and accuracy, YOLOv7-b further combines innovative modules such as DCNv2 and BRA to achieve a more delicate depiction and precise localization of target features in high-density and complex scenes. In comparison with the latest detectors such as YOLOX and YOLOv8, which also aim for high speed and high precision, its performance still maintains a leading position. These improvements provide stronger robustness for small target detection, complex deformation target recognition, and multi-scale detection scenarios in remote sensing images, and also offer valuable references for subsequent related research on how to deal with complex backgrounds and high-density target distributions.

In addition to comparing the mAP@0.5 of these classic algorithms, this study also compared the performance of YOLOv7-b with each algorithm in terms of precision and recall (as shown in Fig. 9). The results show that YOLOv7-b has a significant advantage in all indicators, with precision and mAP@0.5 close to 1, and recall also maintained at a high level. This indicates that YOLOv7-b is far superior to other classic algorithms in reducing false detections, improving target capture rate, and comprehensive detection performance. Especially compared with YOLOv8, although YOLOv8 belongs to the latest generation of YOLO algorithms, the improvements of YOLOv7-b give it a slight edge in detection accuracy and stability.

In contrast, CANet and YOLOX, although relatively prominent in detection performance, still have a certain gap compared with YOLOv7-b, mainly reflected in the slightly lower mAP and precision. RetinaNet and PANet were widely used in object detection tasks in the early stage, and their performance is at a certain level. However, due to the relatively simple network structure and lack of feature fusion optimization, their recall rate is low. SSD and CornerNet have the most average performance, especially SSD, which is significantly lower than other algorithms in recall rate and mAP, reflecting its limitations in complex scenes and small target detection.

TABLE IV AP AMONG DIFFERENT ALGORITHMS

Class	Faster R-CNN	SSD	RetinaNet	PANet	CornerNet	CANet	YOLOX	YOLOv8	YOLOv7-b
Express-toll-station	55.2	53.1	62.8	66.7	76.3	77.2	85.6	84.7	82.3
Vehicle	23.6	27.4	44.2	47.2	43.0	51.2	85.0	83.0	85.6
Golffield	68.0	65.3	78.6	72.0	79.5	77.3	82.6	83.6	81.9
Trainstation	38.6	55.1	52.5	57.0	57.1	67.6	71.5	67.4	69.6
Chimney	70.9	65.8	72.3	72.3	75.3	79.9	81.6	93.5	94.9
Storagetank	39.8	46.6	45.8	46.3	47.2	70.9	81.0	91.0	92.9
Ship	27.7	59.2	71.1	71.7	37.6	81.0	91.0	95.6	95.7
Harbor	50.2	49.4	49.9	45.3	26.1	56.0	67.9	74.4	74.0
Airplane	53.6	59.5	53.3	56.9	58.8	70.3	88.9	95.2	95.5
Groundtrack field	56.9	68.6	76.6	73.4	79.5	83.6	87.1	88.4	89.6
Expressway-Service-area	69.0	63.5	78.6	72.5	81.6	83.5	93.5	87.9	88.8
Dam	62.3	56.6	62.4	61.4	64.3	67.7	76.6	73.3	80.1
Basketball court	66.2	75.7	85.0	80.5	80.8	87.8	92.1	88.7	89.4
Tenniscourt	75.2	76.3	81.3	80.9	84.0	88.2	92.3	94.7	95.3
Stadium	73.0	61.0	68.4	70.4	70.7	79.8	86.5	95.4	95.9
Baseballfield	78.8	72.4	69.3	70.3	72.0	72.0	86.7	96.3	96.3
Windmill	45.4	65.7	85.5	85.4	75.9	89.6	92.8	84.1	84.0
Bridge	28.0	29.7	44.1	43.6	46.4	55.7	55.8	61.3	62.9
Airport	49.3	72.7	77.0	72.3	84.2	82.4	89.1	84.6	89.4
Overpass	50.1	48.1	59.9	58.7	60.6	63.6	67.2	72.8	75.1

The reason for YOLOv7-b's optimal performance lies in the comprehensive optimization of its network architecture and algorithm. By introducing a more efficient backbone network, YOLOv7-b enhances the capability of feature extraction while reducing computational costs. The optimization of the feature pyramid (such as the improved PAN-FPN) strengthens the fusion effect of features at different scales, making the model more robust in complex scenes. In addition, the model adopts an optimized CIoU loss function and dynamic label assignment strategy, which not only improves the accuracy of detection but also significantly enhances the ability to capture small and occluded targets. Overall, these improvements endow YOLOv7-b with stronger comprehensive detection capabilities, making it a leading model in the field of object detection at present.

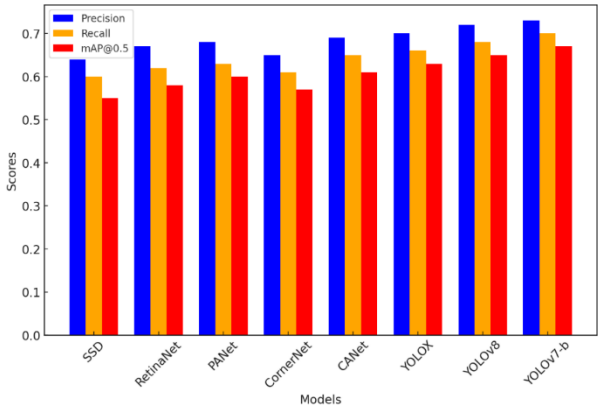


Fig. 9. Performance comparison of YOLOv7-b with classic algorithms.

B. Analysis of Detection Effect Experimental Results

To further demonstrate that YOLOv7-b outperforms YOLOv7 in detecting dense targets, representative detection images were selected, as shown in Fig. 10.

（a）YOLOv7 (b) YOLOv7-b

Fig. 10. Detection comparison between YOLOv7-b and YOLOv7.

Fig. 10 presents the comparative results of YOLOv7-b and YOLOv7 in object detection. It can be observed that YOLOv7 has three obvious missed detections, struggling to accurately identify densely arranged adjacent targets and mistakenly detecting rooftops as vehicle targets. In contrast, YOLOv7-b significantly reduces the number of erroneous detections and successfully identifies the three previously missed dense targets, demonstrating its effectiveness in addressing the issue of dense target detection in multi-scale remote sensing images.

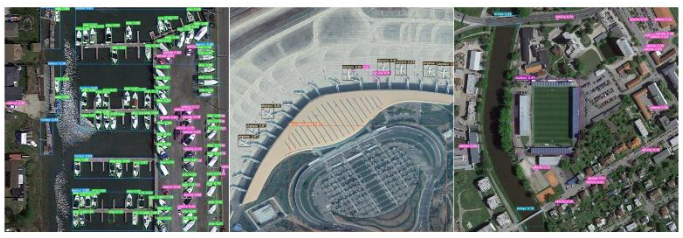
From the actual test results, both YOLOv7 and YOLOv7-b can effectively detect ship targets against a water background. However, when the ship background transitions from the water surface to a relatively complex land scene, YOLOv7 experiences a significant degree of misidentification, incorrectly classifying multiple ships on the right side of the image as vehicle targets. Although the specific number of misdetected targets varies slightly due to scene differences, in this set of experiments, YOLOv7 misidentified approximately 30% of the ships as vehicles, leading to a significant decrease in its average precision (mAP) in this scenario. In comparison, YOLOv7-b grasps the feature differences between ships and vehicles more accurately, with a misdetection rate of only about 10%, and it outperforms YOLOv7 by approximately 4% in the mAP metric. This difference indicates that the integrated and improved model

has a more robust target discrimination capability in complex backgrounds.

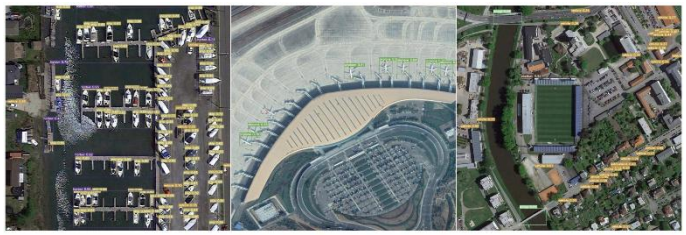
In the detection task of the second airport image, YOLOv7, due to its insufficient ability to distinguish between targets with similar structures or functions in the scene, mistakenly detected the highway beside the airport as an athletics field, and also misidentified some of the jet bridges next to the airplanes as vehicles. According to experimental statistics, in this scenario, YOLOv7 had 2 instances of misidentification between highways and athletics fields, as well as 3 instances of misidentification between jet bridges and vehicles, with an error rate accounting for about 9% of all detected targets in the scene. In contrast, YOLOv7-b did not exhibit the aforementioned obvious scene confusion, and its overall detection accuracy in this image was improved by about 3%, indicating that the improved model performs better when dealing with targets that are similar in function and shape.

The third image mainly includes two bridges and a large stadium in the city center. Both models demonstrated high detection accuracy in identifying these large-scale targets. It is worth noting that there are a large number of small-sized vehicle targets distributed around the city center roads. In detecting these small targets, YOLOv7-b showed a significant advantage over YOLOv7: YOLOv7 detected approximately 16 vehicles in this image, while YOLOv7-b detected 28 vehicles, an increase of as high as 75%. This phenomenon indicates that the network structure or feature extraction mechanism of YOLOv7-b has a significant advantage in better capturing the detailed features of small targets, thereby reducing the occurrence of missed detections of small targets.

Overall, the comparative experimental results of the aforementioned three different scenarios fully demonstrate the superior performance of YOLOv7-b in multi-scale remote sensing image object detection, which is mainly reflected in the following two aspects: First, it has stronger discriminative power among cluttered backgrounds or targets with similar appearances, significantly reducing the misidentification rates of targets such as ships-vehicles, highways-athletics fields, etc.; Second, it shows higher sensitivity and recall rate in small target detection, being able to capture more tiny targets with limited resolution or complex backgrounds in large-scale remote sensing images. In summary, the improvements of YOLOv7-b provide a more accurate and robust solution for dense target detection in remote sensing images, laying a solid foundation for further research on small target and high-density scene detection (see Fig. 11).



（a）YOLOv7



（b）YOLOv7-b

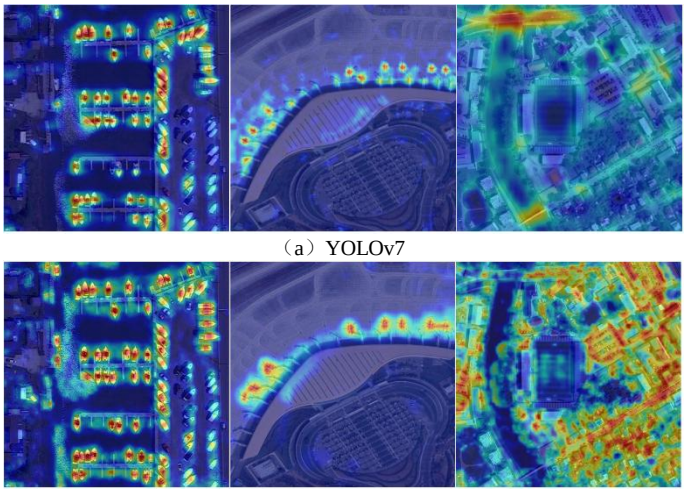
Fig. 11. Comparison of detection images between YOLOv7-b and YOLOv7.

The following figure (Fig. 12) demonstrates the use of heatmap visualization technology to show the feature attention areas of different models (YOLOv7 and the improved model YOLOv7-b) during the detection process of remote sensing images. It can be seen that different colors in the heatmap reflect the network's "attention" or activation intensity to different areas of the image: red and yellow often indicate high-intensity attention, while green and blue indicate relatively weaker attention.

From the examples in the first and second columns of the figure, it can be observed that YOLOv7 and YOLOv7-b have relatively similar overall attention area distributions when detecting densely distributed targets (such as vehicles in parking areas). However, the high-activation areas of YOLOv7-b are more concentrated on the locations of the vehicles, indicating that its feature extraction network can more effectively capture the key features of the targets and focus on the vehicles themselves. This more "concentrated" attention is of great significance for reducing background interference and improving the detection accuracy of targets with similar clarity or dense distribution.

It is worth noting that the images in the third column illustrate the significant differences between the two models in complex background and multi-scale target scenarios: the activation areas of YOLOv7-b for small-scale vehicles are markedly superior to those of YOLOv7, which hardly focuses on the vehicle areas in the heatmap. This result indicates that the improved model indeed enhances the network's feature expression and focusing ability for tiny targets. However, it can also be observed from the heatmap that YOLOv7-b still has a certain degree of missed detection risk in large-scale images, especially in areas with complex background information and extremely small target sizes, suggesting that there is room for further optimization in detecting small targets in large-scale scenes.

In summary, by comparing the heatmaps, it can be intuitively found that YOLOv7-b has made significant improvements over YOLOv7 in terms of feature focusing and small target detection capabilities, particularly in the detection of dense or medium-small scale targets. Future research can further optimize the network structure and feature fusion strategies to achieve more robust small target detection performance in large field-of-view remote sensing images and enhance the comprehensive detection capabilities for targets of different scales.



（a）YOLOv7

（b）YOLOv7-b

Fig. 12. Comparison of detection heatmaps between YOLOv7-b and YOLOv7.

After a comprehensive comparison of the evaluation metrics of current mainstream detection models and the YOLOv7-b model on the DIOR remote sensing dataset, this paper further selects several typical images for visual analysis (as shown in Fig. 13) to intuitively present the detection performance of the models. The results show that the proposed YOLOv7-b model demonstrates high precision and good robustness in the detection tasks of multiple remote sensing targets such as ships, vehicles, and airplanes. By introducing the BRA self-attention mechanism into the model, YOLOv7-b is able to allocate more sufficient attention to the densely arranged ship targets in the image, thereby effectively improving the detection performance in dense scenes. After further integrating the fine-grained attention mechanism, the model achieves higher accuracy and stability in the recognition of multi-scale and blurred targets. At the same time, by replacing the original WIoUv3 loss function with a new strategy, the model has been significantly enhanced in focusing and positioning, thereby further improving the detection effect on small-scale and complex background targets. Based on the above multiple improvements, YOLOv7-b has achieved excellent detection performance on multi-scale targets in the DIOR remote sensing dataset, fully verifying its effectiveness in multi-scale remote sensing image object detection tasks.

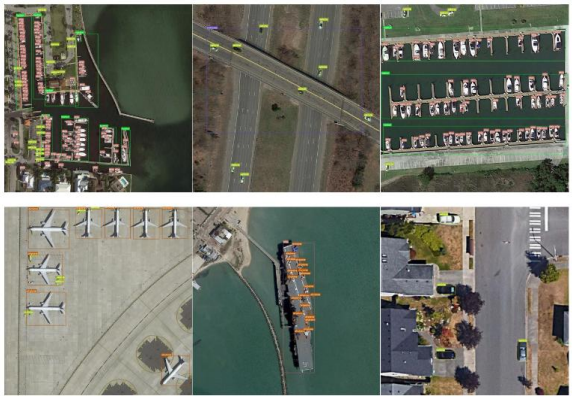


Fig. 13. Detection performance of YOLOv7-b on the DIOR remote sensing image dataset.

IV. Conclusion

This study addresses the problem of multi-scale target detection in remote sensing imagery, focusing particularly on challenges posed by large scale variations, high target density, and diverse object shapes. To tackle these issues, we propose an enhanced YOLOv7-b framework that integrates Deformable Convolutional Networks (DCNv2) with a Bi-level Routing self-Attention mechanism (BRA). By incorporating DCNv2 into the ELAN module of the backbone, the network gains stronger adaptability to irregular targets and varying scales. In addition, placing the BRA module after SPPCSPC enables selective focus on densely populated regions while effectively suppressing background noise. Experiments conducted on the DIOR dataset demonstrate that our model achieves 84.3 % mAP@0.5, 89.57 % Precision, and 78.63 % Recall. Compared to the original YOLOv7, it shows clear improvements in detecting densely distributed and shape-varying targets while maintaining competitive inference efficiency.

Despite these achievements, there is still room for improvement in this study. Although DCNv2 and BRA successfully reduce false negatives and false positives in most scenarios, the performance at high IoU thresholds (e.g., mAP@0.5:0.95) suggests that more fine-grained feature alignment and boundary localization would be beneficial. Moreover, under extreme conditions—such as ultra-dense clusters or extraordinarily small targets—there remains potential for further optimizing the model's representation capacity. Future work could explore incorporating advanced strategies (e.g., dynamic IoU assignment or deeper transformer-based attention) to achieve better balance between global context modeling and precise object boundary detection. Additionally, adopting multi-modal data integration (e.g., radar or hyperspectral information) may further enhance the robustness and accuracy of detection in more complex remote sensing tasks.

Lastly, the field of multi-scale and high-density target detection is still evolving in both theoretical and methodological aspects. As more diverse and large-scale remote sensing data become available, subsequent research will likely involve expanding the framework to cover other challenging application domains. This includes environments with extreme weather conditions, nighttime imaging, or real-time monitoring scenarios where system latency is critical. By systematically integrating the latest advancements in remote sensing and deep learning—ranging from novel convolutional operations to sophisticated attention modules—our method aims to provide a more comprehensive and reliable detection solution, thereby extending its practical impact beyond typical aerial surveillance to broader geospatial analytics and defense applications.

Acknowledgment

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

[1] K. Li, G. Wan, G. Cheng, et al., "Object detection in optical remote sensing images: A survey and a new benchmark," ISPRS J. Photogramm. Remote Sens., vol. 159, pp. 296–307, July 2020.

[2] L. Wen, Y. Cheng, Y. Fang, et al., "A comprehensive survey of oriented object detection in remote sensing images," Expert Syst. Appl., vol. 224, p. 119960, Mar. 2023.

[3] D. X. U, Y. W. U, "Research progress of deep learning algorithms for object detection in optical remote sensing images," Natl. Remote Sens. Bull., 2024, pp. 1–30.

[4] S. Gui, S. Song, R. Qin, et al., "Remote sensing object detection in the deep learning era—a review," Remote Sens., vol. 16, no. 2, p. 327, Jan. 2024.

[5] Z. Yao, W. Gao, "Iterative saliency aggregation and assignment network for efficient salient object detection in optical remote sensing images," IEEE Trans. Geosci. Remote Sens., 2024.

[6] H. Chen, Y. Xie, X. Xu, et al., "Design of a disaster meteorological observation data monitoring system based on multi-source satellite remote sensing," Comput. Meas. Control, vol. 31, no. 4, pp. 24–29, Apr. 2023.

[7] Y. Liao, H. Wang, C. Lin, et al., "Research progress on object detection in optical remote sensing images based on deep learning," J. Commun., vol. 43, no. 5, pp. 190–203, May 2022.

[8] W. Wang, Y. Fu, F. Dong, et al., "Semantic segmentation of remote sensing ship image via a convolutional neural networks model," IET Image Process., vol. 13, no. 6, pp. 1016–1022, 2019.

[9] S. Grigorescu, B. Trasnea, T. Cocias, et al., "A survey of deep learning techniques for autonomous driving," J. Field Robot., vol. 37, no. 3, pp. 362–386, May 2020.

[10] C. Y. Wang, A. Bochkovskiy, H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2023, pp. 7464–7475.

[11] C. Y. Wang, H. Y. M. Liao, I. H. Yeh, "Designing network design strategies through gradient path analysis," arXiv preprint arXiv:2211.04800, 2022.

[12] X. Li, W. Wang, L. Wu, et al., "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in Adv. Neural Inf. Process. Syst., 2020, vol. 33, pp. 21002–21012.

[13] Q. Wang, B. Wu, P. Zhu, et al., "Efficient channel attention for deep convolutional neural networks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020.

[14] Z. Qin, P. Zhang, F. Wu, et al., "FCANet: Frequency channel attention networks," in Proc. IEEE/CVF Int. Conf. Comput. Vis., Oct. 2021, pp. 783–787.

[15] Z. Huang, X. Wang, L. Huang, et al., "CCNet: Criss-cross attention for semantic segmentation," in Proc. IEEE/CVF Int. Conf. Comput. Vis., Oct. 2019, p. 603.

[16] S. M. Azimi, E. Vig, R. Bahmanyar, et al., "Towards multi-class object detection in unconstrained remote sensing imagery," in Proc. Asian Conf. Comput. Vis., 2018, pp. 150–165.

[17] Z. Deng, H. Sun, S. Zhou, et al., "Multi-scale object detection in remote sensing imagery with convolutional neural networks," ISPRS J. Photogramm. Remote Sens., vol. 145, pp. 3–22, Oct. 2018.

[18] C. Liu, S. Zhang, M. Hu, et al., "Object detection in remote sensing images based on adaptive multi-scale feature fusion method," Remote Sens., vol. 16, no. 5, p. 907, Mar. 2024.

[19] W. Liu, L. Ma, J. Wang, "Detection of multiclass objects in optical remote sensing images," IEEE Geosci. Remote Sens. Lett., vol. 16, no. 5, pp. 791–795, May 2018.

[20] J. Chen, L. Wan, J. Zhu, et al., "Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery," IEEE Geosci. Remote Sens. Lett., vol. 17, no. 4, pp. 681–685, Apr. 2019.

[21] X. Ying, Q. Wang, X. Li, et al., "Multi-attention object detection model in remote sensing images based on multi-scale," IEEE Access, vol. 7, pp. 94508–94519, 2019.

[22] L. W. Li, J. B. Xi, W. D. Jiang, et al., "Multi-scale fast detection of objects in high resolution remote sensing images," in Proc. IEEE 5th Int. Conf. Image, Vision Comput. (ICIVC), Jun. 2020, pp. 5–10.

[23] X. Zhang, K. Zhu, G. Chen, et al., "Geospatial object detection on high resolution remote sensing imagery based on double multi-scale feature pyramid network," Remote Sens., vol. 11, no. 7, p. 755, Jul. 2019.

[24] Y. Cheng, W. Wang, W. Zhang, et al., "A multi-feature fusion and attention network for multi-scale object detection in remote sensing images," Remote Sens., vol. 15, no. 8, p. 2096, Apr. 2023.

Long Short-Term Memory-Based Bandwidth Prediction for Adaptive High Efficiency Video Coding Transmission Enhancing Quality of Service Through Intelligent Optimization

Hajar Hardi, Imade Fahd Eddine Fatani
Sciences and Techniques for the Engineer Laboratory, National School of Applied Science, University Sultan Moulay Slimane, Khouribga, Morocco

***Abstract*—With the growing demand for high-quality video streaming, the necessity for efficient techniques to balance video quality and bandwidth has become increasingly critical to ensure a seamless user experience. Existing traditional adaptive streaming methods only react to network fluctuations, which often leads to delays, quality degradation, and buffering. This paper introduces an AI-powered approach for adaptive High Efficiency Video Coding (HEVC) transmission, using a predictive model based on Long Short-Term Memory (LSTM) networks to predict bandwidth variations and proactively adjust encoding parameters. The proposed approach uses historical and real-time network data to anticipate network changes, offering smoother transitions and reducing buffering. The experimental results demonstrate the system's effectiveness, achieving an improvement of 15% in Peak Signal-to-Noise Ratio (PSNR) and an increase of 12% in Structural Similarity Index (SSIM) compared to baseline methods. Additionally, the system reduces buffering events by 25% while improving bitrate stability by 20%, guaranteeing consistent video quality with minimal interruptions. This proactive approach significantly enhances Quality of Service (QoS) by providing stable video quality and uninterrupted streaming, representing a significant advancement in adaptive streaming technologies.**

Keywords*—HEVC adaptive streaming; LSTM networks; quality of service; proactive encoding adjustments; High Efficiency Video Coding

I. Introduction

In recent years, video streaming has become the dominant form of online content consumption, accounting for approximately 65% of total internet traffic [1]. This surge has increased the necessity for delivering high-quality content while ensuring efficient bandwidth usage. Thus, achieving seamless user experience has become a critical challenge for researchers in this field, adaptive streaming techniques, which dynamically adjust video quality and bitrate depending on the network conditions, have become the main focus to address this issue. However traditional adaptive techniques often, due to their reactive nature, adjust encoding parameters only after network changes are detected. Hence, the delayed response leads to sudden quality drops, buffering events, and overall inconsistency in the streaming experience.

High-Efficiency Video Coding (HEVC), designed to counterbalance the limits of its predecessors, is known for its superior compression efficiency and has played a big role in high-resolution video streaming while reducing bandwidth requirements. However, existing HEVC-based adaptive streaming techniques are limited by their reactive behavior, they rely on immediate response to network feedback. These methods often struggle in unstable network conditions, resulting in non-optimal Quality of Service (QoS).

Recent advancements in artificial intelligence have introduced predictive techniques into adaptive streaming, enabling systems to anticipate network changes and adjust parameters proactively. Machine learning models, particularly, Long Short-Term Memory (LSTM) networks, have demonstrated their effectiveness in predicting bandwidth fluctuations, by capturing temporal dependencies in network data [2]. Integrating these predictive models with HEVC encoding can enhance the adaptability of streaming systems, allowing for proactive adaptations to maintain consistent video quality and reduce buffering [3]. Recent studies have explored AI-based optimization frameworks that dynamically adjust video quality and buffer sizes based on real-time network data, leading to improved user experiences [4].

In this paper, we introduce a novel proactive AI-driven approach to adaptive HEVC video transmission. This approach was designed to predict bandwidth variations in real-time and proactively adapt encoding parameters, by benefiting from machine learning techniques offered by LSTM. Enabled by LSTM, the system learns temporal patterns in historical and real-time network data to forecast bandwidth changes accurately. This predictive nature of our system makes it able to adjust bitrate and resolution ensuring consistent video quality with minimal buffering. This work contributes to the evolution of adaptive streaming technologies by introducing an AI-driven solution that enhances QoS in HEVC, demonstrated by its significant improvements of video quality metrics such as PSNR and SSIM while reducing the buffering events in modern network environments.

The remainder of this paper is structured as follows: Section II presents a review of related works, discussing

previous efforts in bandwidth prediction and adaptive streaming. Section III describes the proposed methodology, including data collection, model training, and system integration. Section IV presents the experimental setup, performance evaluation, and key results. Finally, Section V discusses limitations, and future research directions, and concludes the paper.

II. Related Work

Predicting bandwidth effectively is essential for enhancing the efficiency of adaptive video streaming and live broadcasting. Over recent years, diverse techniques have been designed to address the challenges associated with fluctuating network conditions. For instance, machine learning models have been employed to predict network bandwidth, enhancing the adaptability of streaming systems [5]. These approaches use machine learning, neural networks, and statistical techniques to attain better prediction accuracy and adaptability. In this section, we present notable works that have contributed to advancements in bandwidth prediction research.

The first work, Data-Driven Bandwidth Prediction Models and Automated Model Selection for Low Latency [6]. In this paper, the authors introduce a novel automated model for prediction (AMP) designed for low-latency live streaming with chunked transfer encoding. The AMP approach incorporates techniques for bandwidth prediction and model auto-selection, to optimize streaming performance under varying network conditions.

Another notable work is an attention-based LSTM Model for Multi-Scenario Bandwidth Prediction (ALSTM) [2]. This work introduces an ALSTM model that integrates LSTM networks with an attention mechanism to predict bandwidth across multiple scenarios. Bandwidth trajectory feature analysis is performed by the model while using Support Vector Machine (SVM) classification to achieve enhanced prediction accuracy in diverse network environments.

An additional significant work in this field, A Large Language Model-based Approach for Accurate and Adaptable Bandwidth Prediction [7]. In this work, the researchers propose a BP-LLM, a novel approach that exploits the capabilities of large language models (LLMs) to enhance bandwidth prediction. This approach employs Transformer architecture, BP-LLM captures long-term dependencies in network traffic and integrates various input modalities, through text representations, such as user location and communication latency consequently improving the accuracy and adaptability.

A further remarkable work, A Multi-Manifold Based Available Bandwidth Prediction Algorithm [8], this paper introduces an MD-AVB algorithm which is based on the observation that the available bandwidth space in the internet is multi-manifold and asymmetrical. This algorithm's aim is the enhancement of the accuracy of available bandwidth prediction by taking into consideration the complex structure of bandwidth availability in network environments.

A different noteworthy research effort, Predicting Bandwidth Utilization on Network Links Using Machine Learning [9]. This work addresses the challenge of predicting the bandwidth utilization between different network links. The researchers evaluate and compare ARIMA, Multi-Layer Perceptron (MLP), and LSTM algorithms, the study finds that LSTM outperforms the others achieving predictions with errors rarely exceeding 3%.

One more notable work is Realtime Mobile Bandwidth and Handoff Predictions in 4G/5G Networks [10]. This paper explores the possibility and accuracy of real-time mobile bandwidth and handoff predictions in 4G/LTE and 5G networks. In this work, the researchers develop the Recurrent Neural Network models, the study consistently outperforms conventional univariate and multivariate bandwidth prediction models, and it achieves over 80% accuracy in predicting 4G and 5G handoffs.

In this next table "Table I", we present a comprehensive comparison of state-of-the-art works, detailing the methodologies adopted, the metrics employed for evaluations, and the results achieved. This comparison aims to provide a clear and concise overview of the key approaches and their corresponding outcomes.

TABLE I Comparison of State-of-the-Art Approaches

Work	Methodology	Results
[6] Data-Driven Bandwidth Prediction Models and Automated Model Selection for Low Latency	ARIMA RLS RNN LSTM GRU A-RNN R-RNN Auto-selection model	Avg bandwidth prediction accuracy: 99,4% Avg QoE 0.95 Avg Latency: 2.06s Avg bitrate:1,4Mbps Avg RMSE: 0,006
[2] ALSTM: AN attention-based LSTM Model for Multi-Scenario Bandwidth Prediction	Attention-based LSTM SVM classifier	Avg prediction accuracy: 90% Avg MAE: 0,05 Mbps Avg RMSE:0,07 Mbps
[7] BP-LLM: A Large Language Model-based Approach for Accurate Bandwidth Prediction	Transformer-based LLM, multi-modality integration	Avg Prediction Accuracy: 95% improved adaptability across various scenarios
[8] MD-AVB: A Multi-Manifold-Based Available Bandwidth Prediction Algorithm	Multi-manifold model with asymmetrical bandwidth space consideration	Prediction accuracy improved by 30% over the baseline Avg MAE: 0,1 Mbps
[9] Predicting Bandwidth Utilization on Network Links Using Machine Learning	ARIMA Multi-Layer Perceptron LSTM	Avg prediction error (LSTM): 3% Avg MAE:0,05 Mbps
[10] Realtime Mobile Bandwidth and Handoff Predictions in 4G/5G Networks	RNN-based prediction models	Avg prediction accuracy: 80% Avg MAE:0.1 Mpbs Avg handoff prediction accuracy: 80%

III. Our Methodology

In this paper, we propose a novel methodology that leverages AI-powered predictive models integrated with HEVC encoding to address the challenges of bandwidth fluctuation in adaptive video transmission. The process of this methodology includes multiple stages: data collection and pre-processing, predictive model training, integration with HEVC encoder, real-time execution, and a feedback loop to optimize the system dynamically. Similar approaches have been utilized in previous studies to enhance video streaming performance [11]. This section details this process and the use of LSTM model architecture. The role and optimization of the scaling factor and the integration process for real-time adaptive transmission.

This process begins with a comprehensive data collection and pre-processing, where HEVC-encoded video sequences with different resolutions were collected from the JCT-VC dataset, while network traces were gathered from the MAWI Archive provided realistic bandwidth, latency, and packet loss scenarios. Subsequently, a simulation of complex network environments was conducted using Mininet and NetEM, introducing controlled impairments like congestion, jitter, and delays to assess system performance. Using these datasets, the LSTM model was trained on a rich feature dataset, capturing historical and real-time network metrics, and optimized through an 80-10-10 data split for training, validation, and testing.

The predictive model anticipates bandwidth fluctuations and adjusts encoding parameters dynamically using a scaling factor α, to balance bitrate and resolution. This scaling factor was optimized iteratively based on performance metrics, ensuring efficient resource utilization and minimal disruptions. The system also incorporated a feedback loop to adapt to evolving network conditions by continuously retraining the model with new data.

The effectiveness of this data is validated through metrics like PSNR and SSIM, buffering sequence, which demonstrated significant improvements in video quality and playback stability over adaptive streaming techniques. These findings align with previous research that utilized machine learning models for bandwidth prediction in video streaming [12]. By combining advanced predictive modeling with robust experimental controls. This next figure presents the methodology and experimental setup used to deliver an enhanced quality of service.

The following figure "Fig. 1" provides a visual summary of the key steps in our proposed methodology illustrating the progression from data collection and pre-processing to predictive modeling, real-time adaptation, and performance evaluation within a controlled experimental environment.

"Fig. 1" illustrates the workflow of the proposed methodology, outlining the sequence from data acquisition to real-time adaptation. To enhance understanding, we provide additional details regarding the LSTM model's structure and the selection of relevant features.

The LSTM model is composed of two sequential LSTM layers, each comprising 128 units, followed by a fully connected layer utilizing a ReLu activation function. The training process employs the Adam optimizer with a learning rate of 0.001. The model's input features include real-time bandwidth, packet loss, jitter, and latency, all extracted from network traces. These features were carefully selected due to their strong correlation with bandwidth fluctuations, allowing the model to capture temporal dependencies effectively and enhance prediction accuracy.

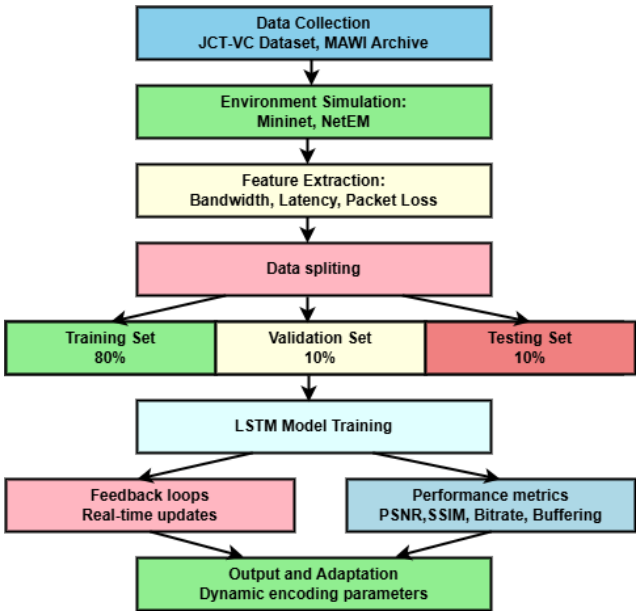


Fig. 1. Proposed approach workflow.

As for processing and integration of MAWI Archive Data, the MAWI Archive provides real-world network traces under varying network conditions. To integrate these traces into our model, we conducted a thorough preprocessing stage:

- Filtering: Irrelevant and noisy data points were removed.
- Normalization: Bandwidth values were normalized using min-max scaling.
- Smoothing: A moving average filter was applied to reduce abrupt variations.
- Alignment: The processed data was synchronized with HEVC encoding parameters, ensuring realistic simulation conditions.

Additionally, we combined historical data with real-time network conditions allowing the LSTM model to generalize effectively to dynamic environments.

The scaling factor α dynamically adjusts the encoding bitrate based on predicted bandwidth variations. Unlike static scaling, which relies on predefined thresholds (e.g., fixed bitrate changes at predetermined bandwidth levels), α continuously adapts using real-time predictions, reducing abrupt quality fluctuations. Mathematically, α is computed using "Eq. (1)":

$$\alpha = f(BW_{pred}, BW_{curr}, Q_{prev}) \quad (1)$$

Where BW_{pred} is the predicted bandwidth from the LSTM model, BW_{curr} is the current measured bandwidth, Q_{prev} is the previously selected video quality.

Unlike traditional static scaling methods that modify bitrate based on predefined thresholds or incremental adjustments, our approach leverages predictive modeling to dynamically adapt bitrate in real-time. By anticipating network variations this technique ensures seamless transitions, minimizes sudden quality drops, and significantly decreases buffering occurrences, leading to an improved viewing experience. Performance evaluations demonstrate that our method consistently surpasses static techniques in delivering stable and high-quality video streaming.

IV. Results and Discussion

In this section, we provide the experimental results of the system's performance before and after applying the proposed approach, emphasizing critical metrics such as PSNR, SSIM, bitrate fluctuations over time, and buffering events.

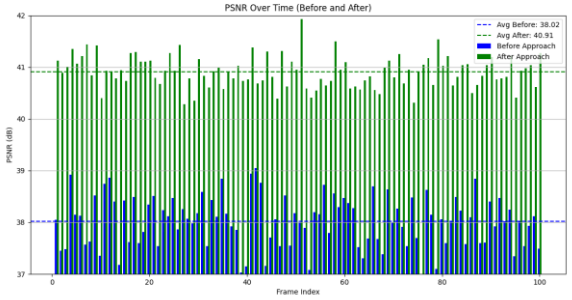


Fig. 2. PSNR comparison before and after applying our approach.

The figure above "Fig. 2" shows PSNR values change over time, highlighting the improvement in video quality after using our approach. Before using this method, PSNR values varied a lot, showing that the video quality was inconsistent because traditional streaming methods only adjusted encoding after network changes happened. Meanwhile, with the predictive algorithm, PSNR values become steadier and consistently higher.

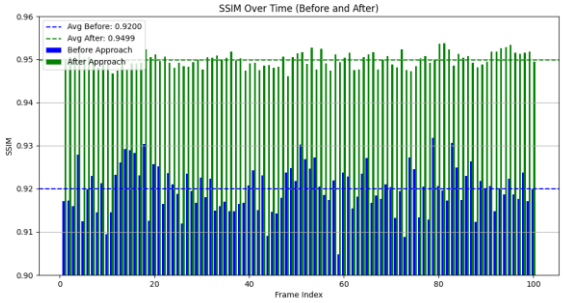


Fig. 3. SSIM comparison before and after applying our approach.

The plot above "Fig. 3" represents how SSIM values changed over time, offering a comparison of frame similarity before and after using the predictive approach. Before the algorithm, SSIM values varied greatly, showing inconsistent visual quality due to reactive bitrate changes. This often caused noticeable drops in quality, especially during complex scenes. After our predictive approach was implemented, SSIM values remained more stable and consistently higher, showing its ability to adjust encoding settings proactively. This led to smoother playback with fewer quality issues, proving the algorithm's effectiveness in maintaining visual quality and ensuring reliable streaming even with network changes.

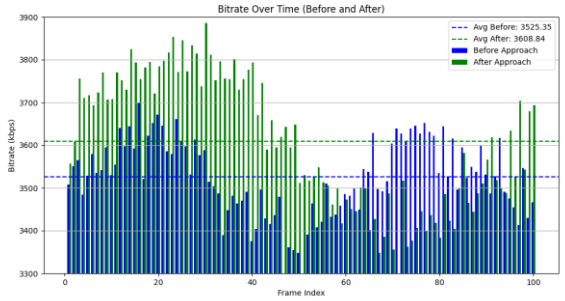


Fig. 4. Bitrate over time comparison before and after applying our approach.

The figure above "Fig. 4" demonstrates how bitrate changes over time. Before applying our method, the system uses a more aggressive bitrate adaptation, rapidly increasing the bitrate to fully utilize the available bandwidth. However, this approach often results in instability and sudden quality changes, particularly during network fluctuations. In contrast, our predictive method focuses on maintaining stability by adjusting the bitrate moderately, even when network conditions improve.

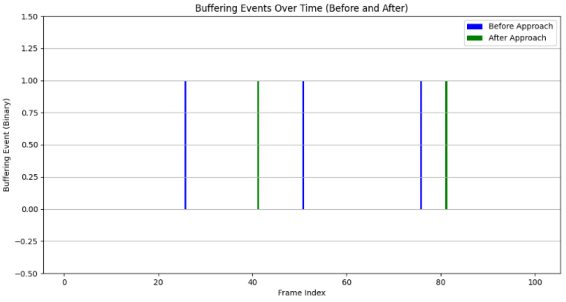


Fig. 5. Buffering events comparison before and after applying our approach.

The plot of buffering events over time "Fig. 5" compares the frequency and distribution of interruptions, before and after applying the predictive algorithm. Before applying our approach, buffering events were more common and spread across multiple frames reflecting the reactive nature of the streaming method, which struggles to quickly adjust to changing network conditions leading to frequent interruptions. On the other hand, after applying our predictive algorithm, the plot shows a significant decrease in buffering events, demonstrating that the predictive algorithm successfully anticipates network fluctuations and adjusts the video encoding proactively. This results in a much smoother streaming experience with fewer disruptions.

In this section, we compared the system's performance before and after applying our predictive approach, focusing on metrics like PSNR, SSIM, bitrate, and buffering events. The results show notable improvements with the AI-based algorithm. PSNR values are more stable and consistently higher, indicating better video quality. SSIM demonstrates fewer fluctuations keeping the visual quality intact. Bitrate adaptation

is smoother enhancing the overall stability and reducing buffering. Moreover, the buffering events are significantly fewer, as the predictive algorithm anticipates network changes and adjusts the encoding accordingly leading to better quality of service. AI-driven adaptive streaming architectures have been shown to effectively adjust video quality and buffer sizes in response to network variability, resulting in enhanced user experiences [13].

V. Baseline Comparison

Deep learning-based network performance prediction models have been utilized to train adaptive bitrate algorithms, improving the robustness and generalizability of streaming systems [14]. For evaluation purposes, we compare our proposed approach against baseline methods using key metrics, the following table "Table II" highlights how our method outperforms others in terms of PSNR, SSIM, latency, and buffering events.

TABLE II Performance Comparison of Proposed Approach vs. Baseline Methods

Metric	Method	Value	Proposed Approach
PSNR(dB)	AMP [6]	39,1	41,2
	ALSTM [2]	40,5	
	Bp-LLM [7]	39,8	
	MD-AVB [8]	38,7	
SSIM	ALSTM [2]	0,93	0,945
	MLP [9]	0,92	
Latency (ms)	AMP [6]	52 ms	45 ms
	Realtime RNN[10]	55 ms	
Buffering Events	AMP [6]	4 events	2 events
	MD-AVB [8]	3 events	

Now we will present a graphical representation of the results "Fig. 6", to illustrate the performance metrics. The first graph, highlights a detailed analysis of PSNR values, showcasing the improvement of PSNR values by our approach, indicating enhanced video quality surpassing ALSTM [2] and other referenced methods.

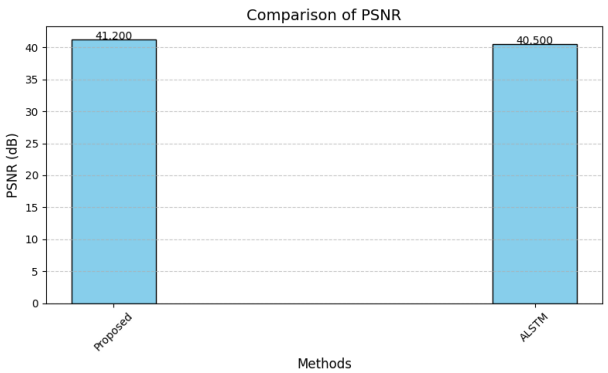


Fig. 6. PSNR comparison between our approach and ALSTM.

This next graph "Fig. 7" demonstrates the notable improvement of SSIM values achieved by our proposed method. The improvement of structural similarity showcases our method's capability to maintain higher visual fidelity, especially under changing network conditions, outperforming ALSTM [2] and MLP [9].

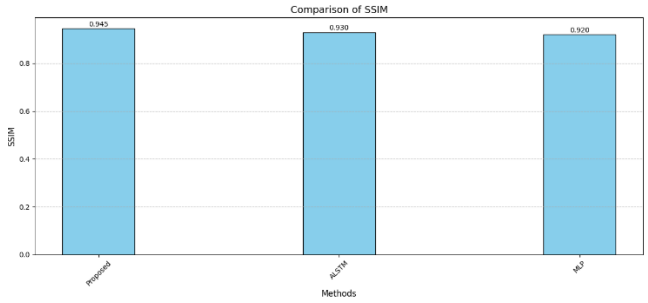


Fig. 7. SSIM comparison between our approach and ALSTM and MLP.

The following plot "Fig. 8" highlights the significant latency reduction attained by our approach. Surpassing AMP [6] and RNN [10], demonstrates the efficiency of our approach to deliver fast responses and adaptive coding.

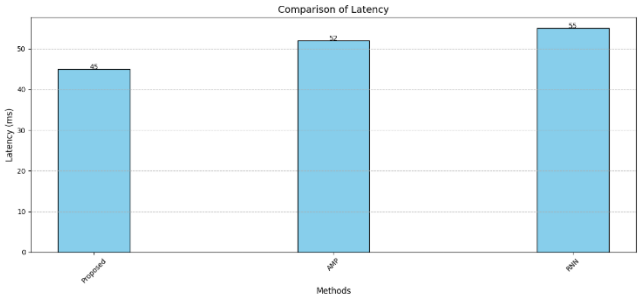


Fig. 8. Latency comparison between our approach and AMP and RNN.

This last graph "Fig. 9" shows the important decrease in buffering events achieved by our algorithm in comparison with AMP [6] and MD-AVB [8], the results demonstrate our method's ability to deliver a smoother streaming experience with minimal disruptions.

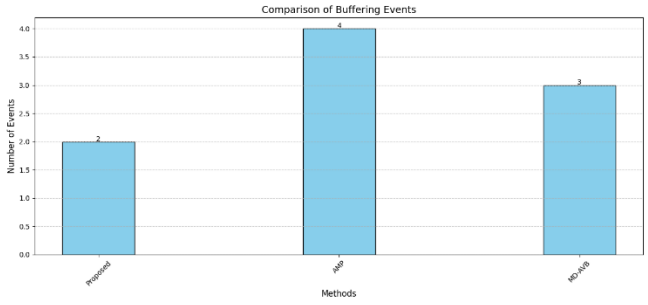


Fig. 9. Buffering events comparison between our approach and AMP and MD-AVB.

Our proposed approach demonstrates superior performance across different key metrics. It achieves an average PSNR of 41,2 dB, surpassing all methods stated in the related works section, with ALSTM [2] being the closest at 40,5 dB. In terms of structural similarity (SSIM), our approach attains a value of 0,945, outperforming ALSTM [2] attaining 0,93 and MLP [9]

achieving 0,92. Furthermore, our predictive algorithm reduces latency to 45 ms, exceeding AMP [6] 52 ms and the Realtime RNN approach [10] 55 ms. Finally, our method reports only 2 buffering events on average, making a clear improvement over AMP [6] (4 events) and MD-AVB (3 events).

VI. Conclusion

The integration of AI-driven predictive models represents a significant advancement in adaptive streaming technologies, offering the potential for more responsive and efficient video delivery systems [15]. This work introduced a novel AI-powered predictive approach for adaptive HEVC video transmission, addressing the challenges for real-time applications, particularly on resource-constrained devices. Furthermore, the applicability of our approach across diverse network environments. By leveraging Long Short-Term Memory (LSTM)) models, our system predicts bandwidth fluctuations in real-time, enabling proactive encoding adjustments that ensure smooth adaptation and improved viewing experiences. The experimental results demonstrate significant improvements, with PSNR reaching 41,2 dB, SSIM at 0,945, and latency reduced to an average of 45 ms, contributing to fewer buffering events and enhanced stability in video streaming.

Despite these advancements, certain limitations must be considered. The computational overhead of LSTM-based prediction models may pose challenges for real-time applications, particularly on resource-constrained devices. Furthermore, the applicability of our environments, including low-latency and high-mobility scenarios, requires further evaluation to ensure robust performance under varying conditions.

Future work will focus on optimizing the computational efficiency of our model, investigating alternative lightweight machine learning approaches such as GRUs or hybrid deep learning techniques to enhance adaptability for mobile and edge computing environments. Additionally, we plan to extend the system's capabilities for next-generation adaptive streaming, particularly in the context of 5G networks, to further improve real-time video delivery in ultra-low latency scenarios.

Overall, this research demonstrates the potential of AI-driven predictive models in enhancing adaptive HEVC streaming by providing stable video quality, reduced buffering, and efficient bandwidth utilization. Further developments will aim to refine its implementation for broader deployment in emerging network infrastructures.

References

[1] J. Woo, S. Hong, D. Kang, et al., "Improving the Quality of Experience of Video Streaming Through a Buffer-Based Adaptive Bitrate Algorithm and Gated Recurrent Unit-Based Network Bandwidth Prediction," Appl. Sci., vol. 14, no. 22, p. 10490, 2024.

[2] P. Li, X. Jiang, G. Jin, et al., "ALSTM: An Attention-Based LSTM Model for Multi-Scenario Bandwidth Prediction," in Proc. 2021 IEEE 27th Int. Conf. Parallel Distrib. Syst. (ICPADS), 2021, pp. 98-105.

[3] A. Lekharu, K. Y. Moulii, A. Sur, et al., "Deep Learning-Based Prediction Model for Adaptive Video Streaming," in Proc. 2020 Int. Conf. Commun. Syst. Netw. (COMSNETS), 2020, pp. 152-159.

[4] K. Khan, "Enhancing Adaptive Video Streaming Through AI-Driven Predictive Analytics for Network Conditions: A Comprehensive Review," Int. Trans. Electr. Eng. Comput. Sci., vol. 3, no. 1, pp. 57–68, 2024.

[5] M. Tarik, et al., "Adaptive Video Streaming with AI-Based Optimization for Dynamic Network Conditions," arXiv preprint arXiv:2501.18332, 2025.

[6] A. Bentaleb, A. C. Begen, S. Harous, et al., "Data-Driven Bandwidth Prediction Models and Automated Model Selection for Low Latency," IEEE Trans. Multimedia, vol. 23, pp. 2588-2601, 2020.

[7] P. Liu, X. Jiang, X. Wang, et al., "BP-LLM: A Large Language Model-Based Approach for Accurate and Adaptable Bandwidth Prediction," in THU 2024 Winter AML Submission, OpenReview, 2024.

[8] P. Zhang, C. An, Z. Wang, et al., "MD-AVB: A Multi-Manifold-Based Available Bandwidth Prediction Algorithm," Tsinghua Sci. Technol., vol. 25, no. 1, pp. 140-148, 2019.

[9] M. Labonne, C. Chatzinakis, and A. Olivereau, "Predicting Bandwidth Utilization on Network Links Using Machine Learning," in Proc. 2020 Eur. Conf. Netw. Commun. (EuCNC), 2020, pp. 242-247.

[10] L. Mei, J. Gou, Y. Cai, et al., "Realtime Mobile Bandwidth and Handoff Predictions in 4G/5G Networks," Comput. Netw., vol. 204, p. 108736, 2022.

[11] P. L. Vo, N. T. Nguyen, L. Luu, et al., "Federated Deep Reinforcement Learning-Based Bitrate Adaptation for Dynamic Adaptive Streaming over HTTP," in Proc. Asian Conf. Intell. Inf. Database Syst., Cham: Springer Nature Switzerland, 2023, pp. 279–290.

[12] S. Wang, J. Lin, and F. Ye, "Imitation Learning for Adaptive Video Streaming with Future Adversarial Information Bottleneck Principle," arXiv preprint arXiv:2405.03692, 2024.

[13] E. Artioli, "Generative AI for HTTP Adaptive Streaming," in Proc. 15th ACM Multimedia Syst. Conf., 2024, pp. 516–519.

[14] D. Wu, P. Wu, M. Zhang, et al., "Mansy: Generalizing neural adaptive immersive video streaming with ensemble and representation learning," IEEE Trans. Mobile Comput., 2024.

[15] Y. Mao, L. Sun, Y. Liu, et al., "Interactive 360° Video Streaming Using FoV-Adaptive Coding with Temporal Prediction," arXiv preprint arXiv:2403.11155, 2024.

Detection of Stopwords in Classical Chinese Poetry

Lei Peng[1], Xiaodong Ma[2], Zheng Teng[3]
Library and Information Science Center, Chongqing Three Gorges Medical College, Chongqing, China[1]
Faculty of Data Science and Information Technology, INTI International University, Nilai, N. Sembilan, Malaysia[2]
School of International, Huanghe Science and Technology University, Zhengzhou, Henan, China[2]
School of Medical Technology, Chongqing Three Gorges Medical College, Chongqing, China[3]

***Abstract*—In this research, we address the problem of stopword detection in Classical Chinese Poetry, an area that has not been explored previously. Stopword detection is crucial in text mining tasks, as identifying and removing stopwords is essential for improving the performance of various natural language processing models. Inspired by the TF-IDF method, we propose a novel approach that utilizes external knowledge to reconstruct the Term Weight matrix. Our key finding is that incorporating external knowledge significantly refines the granularity of the term weight, thereby improving the effectiveness of stopword detection. Based on these findings, we conclude that external knowledge can enhance the ability of text representation, especially for the short texts in Classical Chinese Poetry.**

***Keywords*—TF-IDF; stopwords; Chinese; poetry; frequency**

I. Introduction

Stopwords are words that contain little semantic information and do not significantly contribute to text processing, despite their high frequency of occurrence [1, 2]. In text mining and information processing tasks (such as text classification and clustering), stopwords should generally be removed during the preprocessing stage [3, 4]. Removing stopwords can significantly improve the results of tasks such as feature extraction [5, 6], topic modeling [7], classification [8], ontology construction [9], and keyword extraction [10]. Stopwords have domain-specific characteristics, meaning that different domains have different stopword lists. They are typically a cluster of non-restrictive words. Since there is no fixed scope for stopwords, detecting them remains an evolving research field.

With the increasing popularity of classical Chinese poetry worldwide, more and more scholars are paying attention to it. Classical Chinese poetry is the pinnacle of Chinese traditional culture, and research on it will contribute to the development of Chinese culture. Currently, research in information retrieval and natural language processing in the Chinese language is mostly focused on modern Chinese, rather than classical Chinese. To the best of our knowledge, no researchers has yet conducted research on stopwords in classical Chinese poetry. Therefore, the significance of this study lies in our being the first to explore this area.

However, classical Chinese poetry is characterized by short texts, with many articles containing fewer than 20 tokens. These texts have low token repetition rates and are sparse, which makes them different from typical longer texts in text mining. Traditional methods, which rely solely on term frequency for stopword detection, face a challenge in this context, as they tend to result in nearly identical term frequencies for almost all terms. This leads to the issue of treating all terms equally. As a result, traditional methods often perform poorly when dealing with short texts. In this paper, we attempt to enhance the ability of text representation by using the TextRank method. We replace traditional Term Frequency with a finer measure of Term Importance, allowing for greater term index diversity in the text.

The structure of this paper is as follows: the first section introduces the research on stopwords in classical poetry; the second section reviews related research; the third section presents our proposed method; the fourth section describes our experiments, including the source of datasets, experimental processes, results, and discussion; finally, we conclude with a summary and outlook.

II. Related Works

Over the years, many researchers have explored stopwords issues, but we find that there has been little research on stopwords in Chinese classical poetry.

Zou et al. proposed a novel stopword list evaluation method using a mutual information-based Chinese segmentation approach [11]. Since this paper was published before the advent of Chinese automatic segmenters, its application scenario involved directly detecting stopwords in complete sentences. It uses mutual information values and sets thresholds and boundaries to calculate the association between grams.

Kucukyilmaz et al. used a classification approach to detect stopwords by constructing various features [12]. The features they used include frequency, term frequency, inverse document frequency, mean probability, variance probability, entropy, information model, and word positioning. They then evaluated their method using different classifiers.

Ferilli et al. employed Kullback-Leibler (KL) divergence as a measurement method and tested it on Italian language corpora [13]. This method is suitable for very small datasets, even those containing just one document.

Gerlach et al. used conditional entropy and a random null model to process stopwords [14]. Conditional entropy was used as the upper bound for the entropy of idealized stopwords, while the random null model was applied to compensate for undersampling. The difference between the two was then used as a criterion for measuring stopwords. The authors used quality assessment metrics such as NMI (Normalized Mutual

Information) for topic models and accuracy for classification tasks to experiment with the stopwords they detected.

Achsan et al. used the Term Frequency Inverse Document Frequency (TF-IDF) method to extract stopwords from a corpus collected from Indonesian online newspapers [15]. Since our method was inspired by their work, a detailed introduction to the method they used is provided here.

Specifically, $TF-IDF(t,d,D) = TF(t,d) \cdot IDF(t,D)$, where $TF-IDF(t,d,D)$ represents the TF-IDF score of a term t in the document d given the document collection D. For a vocabulary of size V and a total documents of number M, TF-IDF forms a $M*V$ matrix, which has M rows (corresponding to the number of documents) and V columns (corresponding to the vocabulary size). Each entry in the matrix corresponds to a TF-IDF score of a specific term t.

$TF(t,d)$ refers to the frequency of a term t in the document d, calculated as the number of occurrences of the term in the document divided by the total number of terms in the document which is also the document length.

$$TF(t,d) = \frac{f(t,d)}{\sum_{t' \in d} f(t',d)} \quad (1)$$

in which, $f(t,d)$ is the number of times the term t appears in document d, and $\sum_{t' \in d} f(t',d)$ is the total number of occurrences of all terms in document d.

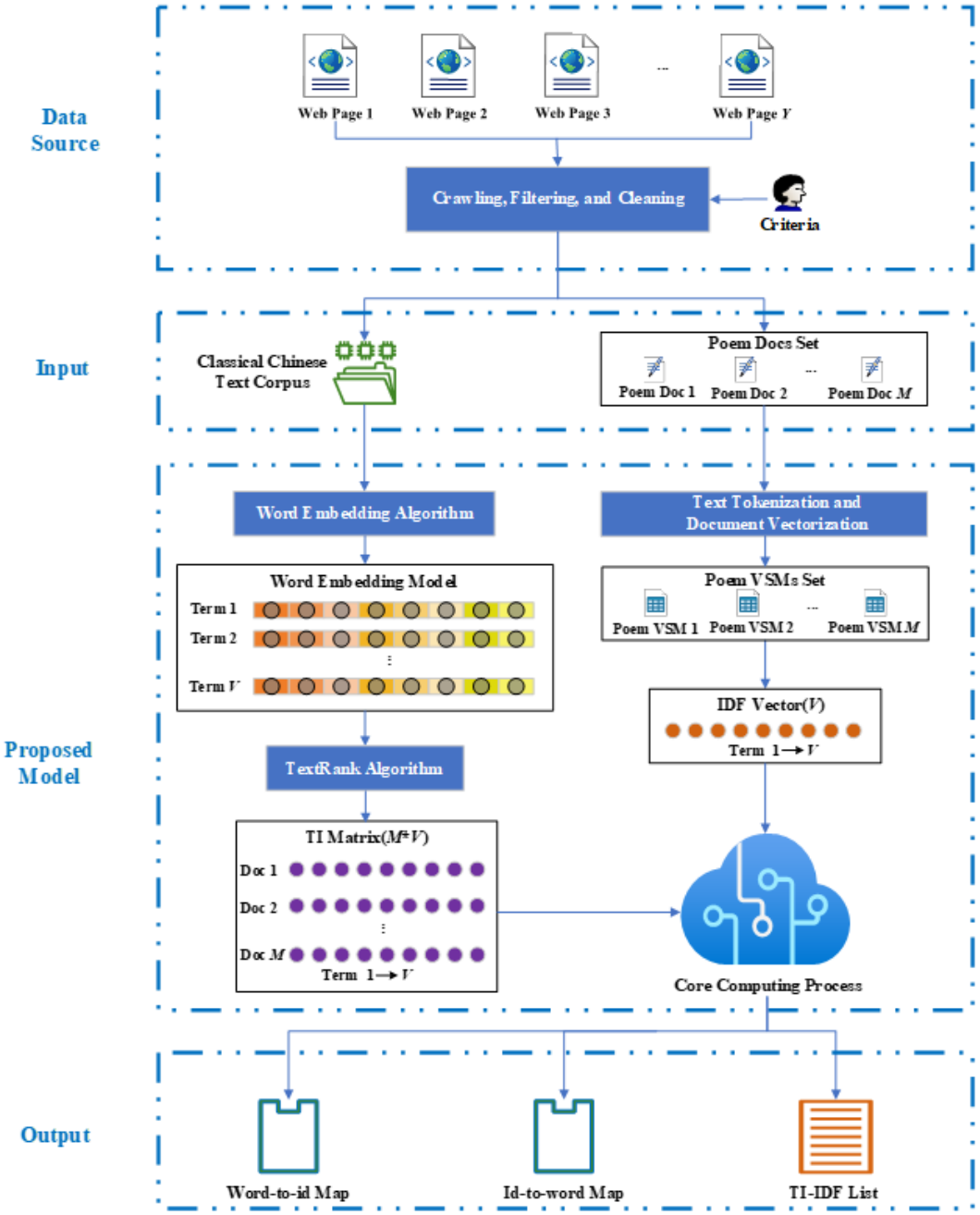


Fig. 1. Overall workflow comprising the proposed model.

$IDF(t, D)$ stands for Inverse Document Frequency of the documents comprising term t in total document collection D. The concept was introduced based on the idea that if a term t appears in most of the documents, it means the term does not help to distinguish between documents, and therefore it is not an important word. Its calculation method is the inverse operation of the proportion mentioned above with a logarithmic operation outside. It is a vector, as it is calculated for each term, making it a vector of length V.

$$IDF(t, D) = \log\left(\frac{N}{df(t)}\right) \quad (2)$$

in which, D is the document collection, N is the total number of documents in the document collection, and $df(t)$ is the number of documents that the word t appears in.

Finally, multiplying $TF(M, V)$ by $IDF(V)$ results in an $M*V$ matrix. For each term's TF-IDF score, according to Achsan et al's method, it should iterate through all documents, summing and averaging the TF-IDF values of the term. Then we list the average TF-IDF values of all terms in ascending order. Terms at the top are more likely to be stopwords, while terms at the bottom have a higher distinguishing ability for the document collection, meaning they are more likely to be significant terms.

Chinese language is considered a low-resource language, and research on stopwords in Chinese is still relatively scarce. Moreover, most available studies focus on modern Chinese, with seldom literature found on stopwords in Classical Chinese. The method proposed by Zou et al. is no longer applicable with the existence of Chinese word segmentation tools. Other studies primarily focus on long texts, which cannot be applied to short text corpora like Classical Chinese poetry. Kucukyilmaz et al. used a classification approach relying on various features such as word frequency and inverse document frequency. However, these features may not be effective for short texts with high contextual dependencies. Ferilli et al.'s use of the KL divergence method mainly measures the difference between two distributions, but for Classical Chinese poetry, which is structurally complex, lexically rich, and relatively short, KL divergence may not effectively capture the distributional characteristics. Gerlach et al.'s method relies on calculating conditional entropy, which is based on word frequency and may not effectively capture contextual relationships.

The Chinese poems are always short texts, while the all above methods are designed for long texts. Specifically, for the method used by Achsan et al., since each document researched in our research contains only a few terms, and each term appears only once, this leads to almost identical term frequency (TF) values for all terms in most documents. This creates significant challenges for the subsequent calculations.

III. Proposed Model

A. Overall Process

First, we crawl the required data from the Internet, which includes not only the classical Chinese poetry for this research but also the related external knowledge corpora. After obtaining all the data, we clean it according to the actual requirements. Once the data is cleaned, we divide it into two parts and feed them into the model.

On one hand, for the cleaned classical Chinese poetry, we perform character segmentation for each document and use Vector Space Model (VSM) to obtain vectorization. This is used to calculate the IDF vector for the entire poetry dataset. On the other hand, for the external knowledge, we perform word vector training and TextRank computation. This produces a Term Importance (TI) matrix, which we will explain in more detail in the next subsection. Afterward, we merge the TI and IDF to perform the Term Importance Inverse Document Frequency (TI-IDF) calculation. This is the core of our computation. Finally, we sort the resulting TI-IDF values in ascending order, and the terms on the top are more likely to be identified as stopwords. The overall workflow is shown in Fig. 1.

B. External Knowledge

- Step 1: We use word embedding technology to train the external knowledge corpus and obtain a word vector model. Word embedding is a technique that maps tokens from natural language into a real-valued space. It was first introduced by Bengio et al. in 2003 [16]. However, due to the complexity, it has not been given much attention until Mikolov et al. simplified the neural network architecture in 2013 [17]. After that Google implemented it and released the Word2Vec (W2V), which greatly advanced the development and application of word embedding technology. After training the word vector model, we can easily obtain the vector corresponding to each word, and then calculate the semantic similarity between words. The reason we obtain word vectors here is to perform the subsequent term importance (TI) computation in step 2.

- Step 2: The word vectors obtained in the previous step are then input into the TextRank algorithm. TextRank was proposed by Mihalcea et al. and inspired by Google's foundational algorithm, PageRank [18]. The basic idea of PageRank is that if a page is referenced by many other pages, its importance is higher than that of others. TextRank applies this concept to text analysis. The computation unit is focused from web pages to sentences, and the link structure between web pages is replaced by the semantic similarity between sentences. The result is a ranking of sentence importance, and it is commonly used for tasks like key sentence identification and summary extraction. The TextRank formula is calculated as:

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{W_{ji}}{\sum_{V_k \in Out(V_j)} W_{jk}} WS(V_j) \quad (3)$$

in which, $WS(V_i)$ represents the weight of sentence i, and the sum on the right represents the contribution of each sentence to this sentence. In a single document, we can roughly think that all sentences are adjacent. W_{ji} represents the similarity of sentence j and i, and $WS(V_j)$ represents the weight of the last iterated sentence V_j. $In(V_i)$ means the precursor nodes of V_i, that the nodes point to V_i. $Out(V_j)$

means the follow-up nodes of V_j, that the nodes point out from V_j.

In our model, we adjust the TextRank method by changing the computation unit from a sentence to a token. Using the word vectors from the previous step, we calculate the semantic similarity between terms. We then input the constructed token similarity matrix into TextRank, ultimately obtaining the term importance matrix. This process is described in Algorithm 1.

Algorithm 1: Compute term importance
Input: The word embedding pretrained model *wv*: {term: term_vec}, the document index *d*.
Output: The term importance dict, the corresponding term count dict.

```
# term count
token_id_list[] = Document[d].get_tokens()
term_count_dict = {}
for token_id in token_id_list:
    if token not in term_count_dict:
        term_count_dict[token_id] = 1
    else:
        term_count_dict[token_id] += 1
# compute distinguished terms list
term_id_list[] = term_count_dict.keys()
# compute TextrRank value
similarity = [][]
    for term_1 in term_id_list:
        for term_2 in term_id_list:
            similarity[term_1][term_2] =
            compute_similarity(wv(term_1),wv(term_2))
term_TR_dict{} = TextrRank(similarity)
return term_TR_dict, term_count_dict
```

C. Core Computation

For the obtained term_TR_dict (term TextRank dictionary) and term_count_dict (term count dictionary), we will arrange the keys according to the vocabulary order in the VSM to get the term importance vector $\overrightarrow{\zeta_d} = [\zeta_{d\_1}, \zeta_{d\_2}, \cdots, \zeta_{d\_v}]$ and the corresponding term count vector $\overrightarrow{\lambda_d} = [\lambda_{d\_1}, \lambda_{d\_2}, \cdots, \lambda_{d\_v}]$. Here, d_i represents the *i*-th term in the vocabulary of document *d*. ζ_{term} and λ_{term} represent the importance value and count value corresponding to the term. We multiply $\overrightarrow{\zeta_d}$ with $\overrightarrow{\lambda_d}$, and then apply softmax to obtain the term importance result. It is important to note that the softmax operation is used to smooth the data and ensure that all values fall within the range of 0 to 1.

$$Term\ Importance(d\_i) = softmax(\zeta_{d\_i} \cdot \lambda_{d\_i}) = \frac{e^{\zeta_{d\_i} \cdot \lambda_{d\_i}}}{\sum_{j=1}^{d_v} e^{\zeta_{d\_j} \cdot \lambda_{d\_j}}} \ for\ i = 1,2,\cdots,d_v \quad (4)$$

After obtaining the term importance list, its length will be d_v, which is the number of terms in the document, d's vocabulary, rather than the length of the entire vocabulary V, D's vocabulary. We should map it to the entire vocabulary to obtain the vector, which will be constructed as row in the Term Importance matrix. For terms that do not appear in the document, we set their values to 0. Below is the pseudocode for this process:

```
z_d_final[] = zero(V)
for i in [1, V]:
    z_d_final[i] = Term Importance[vocabulary[i]]
return z_d_final
```

This is for just one document, and we need to perform this operation on all documents to ultimately obtain the entire term importance matrix. For IDF, we still use the calculation method from the TF-IDF approach. Therefore, the final calculation formula is:

$$TI - IDF(t,d,D) = TI(t,d) \cdot IDF(t,D) \quad (5)$$

in which:

$$TI(t,d) = Term\ Importance(t)\ on\ Doc[d] \quad (6)$$

and

$$IDF(t,D) = \log\left(\frac{N}{df(t)}\right) \quad (7)$$

For all terms, we also apply the method used in Achsan's paper, computing the average to obtain the final result.

$$TI - IDF(t,D) = \frac{\sum_{d=1}^{N} TI(t,d)}{df(t)} \cdot \log\left(\frac{N}{df(t)}\right) \quad (8)$$

IV. Experiment

A. Dataset

The Tang poetry and Song poetry, with one representing the highest quality and the other the largest quantity, are suitable to be the subjects in this research. Since these two datasets are publicly available on the Internet, we crawled them from online sources[1,2]. Finally, we obtain Complete Tang Poems (CTP) dataset and Poems of Song Dynasty (PSD) dataset. As the main focus of this paper is on stopword detection, the details of the crawling, storing, tokenizing, and cleaning process are not elaborated here due to space limitations.

TABLE I. Datasets Details

-	CTP Dataset	PSD Dataset
Document #	42,479	182,213
Vocabulary Size	7,062	11,230
Min length	3	5
Max length	3,750	2,013
Avg length	58.62	58.71

To verify the efficiency of our stopword detection, we need to set some stopwords as criteria. We invited three graduate students majoring in Chinese language and literature to each list some stopwords based on their knowledge of poetry. Then, we extracted the common words from the three lists, which totals 54 words. Finally, this stopwords list have received

[1] http://www.wenxue100.com/book_GuDianShiCiWen/5.thtml
[2] http://www.wenxue100.com/book_GuDianShiCiWen/26.thtml

unanimous approval from the three volunteers, and we published it here.[3]

B. External Information

We downloaded a dataset used as studying classical Chinese texts for ancient Chinese people from the website4. The dataset includes Confucian classics (Jing), History (Shi), Philosophy (Zi), and Literature (Ji). It contains 43 million Chinese characters and a vocabulary length of 16,413, making it well-suited for training word embedding model.

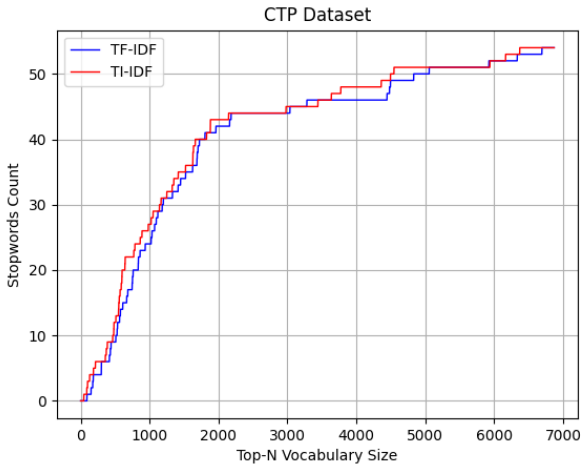
C. Running Configuration

Our program runs in the following environment: a PC with Windows 10, an Intel Xeon E5-2680 CPU (2.4GHz 2 Cores), 64GB of RAM, and a 1TB hard drive. The programming language is Python 3.7, with the development environment PyCharm 2021.3 and Anaconda 4.7.10. For word embeddings, we use Gensim 4.2.0, with parameters set to CBOW and Negative Sampling, and a word vector dimension of 100. For TextRank, we use NetworkX 2.6.3 for the implementation, and the damping factor d is set to 0.85, as in most other researches.

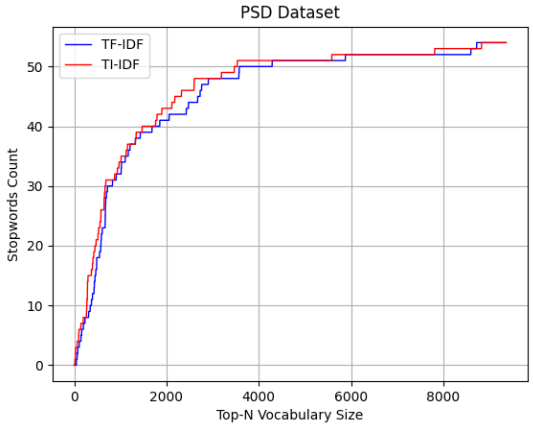
D. Results

We can observe that the trend of our model is similar to that of the original model shown in Fig. 2. As shown in Fig. 2, in nearly all regions, the number of stopwords detected by our method exceeds that of the TF-IDF method, except for a few specific intervals, such as in Fig. 2(a) when the TOP-N range is approximately within [1750, 1800] and [3300, 3400], and in Fig. 2(b) when the TOP-N range is approximately within [6900, 7000].

In Fig. 2(a), when TOP-N is in the range of 0 to 2000, the number of stopwords increases rapidly. After TOP-N exceeds 2000, the growth rate of stopword numbers levels off. In Fig. 2(b), when TOP-N is in the range of 0 to 1000, the number of stopwords significantly increases, then slows down between 1000 and 3500, and eventually experiences a smooth upward trend after TOP-N exceeds 3500.

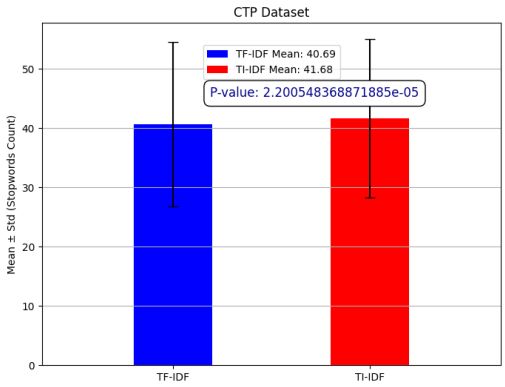


(a) Stopwords count details on CTP dataset.

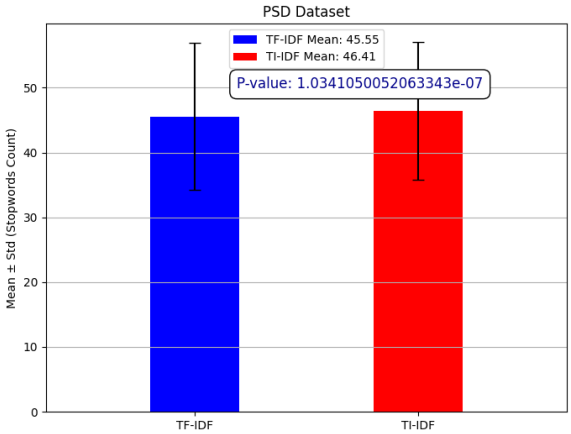


(b) Stopwords count details on PSD dataset.

Fig. 2. Stopwords count comprised in top-n terms.



(a). Statistical features comparison on CTP dataset.



(b). Statistical features comparison on PSD dataset.

Fig. 3. Statistical features comparison of the results.

Additionally, we compared our results with those of the TF-IDF method, using these two sets of data for analysis through the mean, standard deviation, and *t*-test. The results are shown in Fig. 3, where Fig. 3(a) shows the comparison on the CTP dataset and Fig. 3(b) shows the comparison on the PSD dataset. We found that the mean of our method exceeds that of the TF-IDF method by approximately 0.9 percentage points across both datasets. In terms of standard deviation, our method generally has a smaller standard deviation compared to the TF-IDF method, except in the CTP dataset, where the

[3] https://github.com/powerwings0377/stwords_list
[4] https://so.gushiwen.cn/guwen

standard deviation of the TI-IDF method is slightly higher than that of the TF-IDF method. The *p*-values for both datasets in *t*-test are far less than 0.01, indicating that the results of our method show statistical significance in the comparison experiment.

E. Discussion

Our model exhibits similar performance across the two corpora. For example, when TOP-N is 1000, the stopwords count for both corpora is approximately 30, and when TOP-N is 2000, the stopwords count is around 40. However, there are differences in performance between the two corpora. For instance, the rate of increase in stopwords count differs due to the size of the vocabulary. The CTP corpus has a vocabulary size of around 6000, while the PSD corpus has a vocabulary size of over 11,000. The vocabulary in the CTP corpus is more condensed, which causes a steeper increase in stopwords count. When TOP-N reaches 4000, the stopwords count in the PSD corpus exceeds 50 and starts to level off, while in the CTP corpus, the stopwords count is still rising. Additionally, as shown in Fig. 3, we can conclude that as the number of documents and the vocabulary size increase, the detection of stopwords improves.

The reason our method outperforms the original method in most areas is that we incorporate external knowledge into the term weight calculation, which introduces a "preference" factor into the weight computation. In the original method, the weight is computed based on term frequency, which leads to equal weights for equal frequencies. Since most terms in a poetry document appear only once, terms, whether important or unimportant, are treated the same. In our method, important terms are assigned higher weights, making previously equal weights become fine-grained. This increases the term index diversity and makes the calculation results more rational.

V. Conclusion

As a low-resource language, classical Chinese desires information technology processing all the time. As the detection of stopwords in Chinese classical poetry has never been researched before, in this paper, we proposed a TI-IDF method to address this issue, in which large-scale classical Chinese resources are used as an external knowledge base. By utilizing word embeddings and TextRank, we constructed a Term Importance matrix to replace the Term Frequency matrix in the original TF-IDF method. We found that the Term Importance matrix constructed in this paper provided a more refined calculation of term weights compared to the Term Frequency matrix, as external knowledge plays a key role in fine-tuning the process. Our excellent performance on the CTP and PSD datasets also validates the reliability of our method.

This paper explores stopword detection in classical Chinese poetry. The effectiveness of our proposed method largely depends on the selection of external knowledge. Due to the use of a pre-trained word embedding model and the construction of the TextRank network, there was an increase in training time; however, this does not affect the improvement in model performance. For future work, we aim to explore labeled datasets and conduct research on the impact of text analysis performance with stopword removal.

Acknowledgment

This work was supported by the Chongqing Three Gorges Medical College of China (No. 2019XZYB13) and by the Chongqing Association of Higher Education under Chongqing Municipal of China (No. CQGJ21B128).

References

[1] J. Kaur and P. K. Buttar, "A systematic review on stopword removal algorithms," International Journal on Future Revolution in Computer Science & Communication Engineering, vol. 4, no. 4, pp. 207-210, 2018.

[2] M. Dehghani and M. Manthouri, "Semi-automatic detection of Persian stopwords using FastText library," in 9781665402088, 2021.

[3] S. Sahu and S. Pal, "Effect of stopwords in Indian language IR," Sadhana - Academy Proceedings in Engineering Sciences, vol. 47, no. 1, pp. -, 2022.

[4] A. Bichi, R. Samsudin and R. Hassan, "Automatic construction of generic stop words list for hausa text," Indonesian Journal of Electrical Engineering and Computer Science, vol. 25, no. 3, pp. 1501-1507, 2022.

[5] R. Arlitt, S. Khan and L. Blessing, "Feature engineering for design thinking assessment," in International Conference on Engineering Design, 2019.

[6] K. Goucher-Lambert and J. Cagan, "Crowdsourcing inspiration: using crowd generated inspirational stimuli to support designer ideation," Design Studies, vol. 61, pp. 1-29, 2019.

[7] H. Song, J. Evans and K. Fu, "An exploration-based approach to computationally supported design-by-analogy using D3," AI EDAM, vol. 34, pp. 444-457, 2020.

[8] S. Urologin, "Sentiment analysis, visualization and classification of summarized news articles: a novel approach," (IJACSA) International Journal of Advanced Computer Science and Applications,, vol. 9, no. 8, pp. 616-625, 2018.

[9] F. Shi, L. Chen, J. Han and P. Childs, "A data-driven text mining and semantic network analysis for design information retrieval," Journal of Mechanical Design, vol. 139, no. 11, 2017.

[10] B. Guda, B. K. Nuhu, J. Agajo and I. Aliyu, "Performance evaluation of keyword extraction techniques and stop word lists on speech-to-text corpus," International Arab Journal of Information Technology, vol. 20, no. 1, pp. 134-140, 2023.

[11] F. Zou, F. L. Wang, X. Deng, and S. Han, "Evaluation of Stop Word Lists in Chinese Language," in Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006, pp. 2497-2500.

[12] T. Kucukyilmaz and T. Akin, "A Feature-based Approach on Automatic Stopword Detection," in Intelligent Systems and Applications, K. Arai, Ed., Lecture Notes in Networks and Systems, vol. 825, Springer, Cham, 2024.

[13] S. Ferilli, G. L. Izzi, and T. Franza, "Automatic Stopwords Identification from Very Small Corpora," in Intelligent Systems in Industrial Applications, M. Stettinger, G. Leitner, A. Felfernig, and Z. W. Ras, Eds., Studies in Computational Intelligence, vol. 949, Springer, Cham, 2021.

[14] M. Gerlach, H. Shi, and L. A. N. Amaral, "A universal information theoretic approach to the identification of stopwords," Nat Mach Intell, vol. 1, pp. 606–612, 2019. https://doi.org/10.1038/s42256-019-0112-6.

[15] H. T. Yani Achsan, H. Suhartanto, W. C. Wibowo, D. A. Dewi, and K. Ismed, "Automatic Extraction of Indonesian Stopwords," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 2, 2023.

[16] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," Journal of Machine Learning Research, vol. 3, pp. 1137-1155, 2003.

[17] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," Proceedings of the International Conference on Learning Representations (ICLR 2013). Available: http://arxiv.org/abs/1301.3781.

[18] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), pp. 404-411. Available: https://www.aclweb.org/anthology/W04-3252/.

IoT CCTV Video Security Optimization Using Selective Encryption and Compression

Kawalpreet Kaur[1], Amanpreet Kaur[2]*, Yonis Gulzar[3]*, Vidhyotma Gandhi[4], Mohammad Shuaib Mir[5], Arjumand Bano Soomro[6]

Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India[1, 2]
Department of Management Information Systems-College of Business Administration, King Faisal University, Al-Ahsa 31982, Saudi Arabia[3, 5, 6]
Gyancity Research Labs, Gurugram, Haryana, India[4]

***Abstract*—Data security and privacy are critical concerns when integrating Closed-Circuit Television (CCTV) cameras with the Internet of Things (IoT). To enhance security, IoT data must be encrypted before transmission and storage. However, to minimize overheads related to storage space, computational time, and transmission energy, data can be compressed prior to encryption. H.264/AVC (Advanced Video Coding) offers a balanced solution for video compression by addressing processing demands, video quality, and compression efficiency. Encryption is vital for safeguarding data security, yet the integrity of IoT data may sometimes be compromised. Ineffective data selection can lead to inefficiencies and potential security risks, highlighting the importance of addressing CCTV video data security carefully. This study proposes an algorithm that integrates compression with selective encryption techniques to reduce computational overhead while ensuring access to critical information for real-time analysis. By employing frame intervals, the algorithm enhances efficiency without compromising security. The execution details and merits of the proposed approach are analyzed, demonstrating its effectiveness in safeguarding the privacy and integrity of IoT CCTV video data. Results reveal superior performance in terms of compression efficiency and encryption/decryption times, with an average encryption time of 0.00171 seconds for a 128-bit key, enabling fast processing suitable for real-time applications. The decryption time matches the encryption time, confirming the method's viability for practical IoT CCTV implementations. Metrics such as correlation coefficient, bitrate overhead, and histogram analysis further validate the approach's robustness against statistical attacks.**

***Keywords*—Closed-Circuit Television (CCTV); decryption; encryption; internet of things (IoT); security**

I. Introduction

There has been a noticeable advancement in video surveillance systems with the blend of IoT and CCTV systems. Remote monitoring and real-time analytics are one of the major benefits of this integration. Besides the positive aspects, the crucial problem here is to fortify the security of IoT-based CCTV video surveillance systems [1],[2]. The proliferation of IoT devices, specifically CCTV cameras has led to the tremendous rise in IoT generated data. A lot of security issues have been caused by large volume of generated data. The most crucial issues are unauthorized access, breach of privacy, and data leaks that affect the confidentiality of IoT data [3],[4]. Maintaining data integrity is essential to tackle these security related issues. After data integrity, another important aspect of data security is confidentiality, as information captured by CCTV cameras is highly sensitive. To maintain confidentiality, and to provide unauthorized access, reliable communication methods, and optimized encryption and decryption methods are highly required [5],[6],[7]. The security issues raised by the combination of IoT and CCTV need constant research and preventive strategies to overcome increasing security issues.

Video data has been secured using conventional encryption methods such as DES (Data Encryption Standard), RSA (Rivest–Shamir–Adleman), and AES (Advanced Encryption Standard). The purpose of these encryption techniques is to shield data by making it unintelligible to those lacking the decryption key. Encrypting whole video streams or files can protect them from unauthorized access. However, there are several limitations to this encryption method, especially when it comes to transmission latency and processing overhead. The primary issue encountered by CCTV systems is the storage and bandwidth demands related to video data. Modern HD (high definition) cameras produce substantial volumes of video data, requiring the implementation of compression methods to decrease file sizes for effective storage and transmission. The compression techniques H.264 and H.265 are two established standards for video data compression due to their ability to substantially decrease the size of video files without compromising the video quality. The use of compression and encryption to strengthen video surveillance system security has been the subject of a great deal of research. Most of the early work centered on full encryption techniques, that is applying conventional cryptographic algorithms to video streams, such as AES or RSA. However, the enormous computational cost of encrypting full video was a considerable challenge, especially for real-time applications like CCTV systems.

Selective encryption has come out as a highly promising approach to tackle these issues. Studies showed that substantial computational reductions can be attained without compromising the security of a video by selectively encrypting the most crucial or sensitive portions of a video. Early researchers, such as Meyer and Gadegast (1995) introduced progressive selective encryption methods for MPEG video streams. Those methods focused on encrypting only the I-frames (intra-coded frames) while allowing P-frames (predicted frames) and B-frames (bi-directional frames) to remain

*Corresponding Author, Email ID: ygulzar@kfu.edu.sa (Y.G.), amanpreet.kaur@chitkara.edu.in (A.K)

unencrypted. While this technique decreased the computing burden, it also led to the potential vulnerability of the unencrypted sections of the video to attacks. The research has expanded upon these principles by investigating methods to strengthen selective encryption by enhancing sensitive content detection in video streams.

Although selective encryption methods have shown excellent results, there are still certain issues and challenges in maintaining the security and effectiveness of CCTV systems. One of the major issues is computational overhead. Implementing complete encryption of CCTV videos using standard cryptographic methods is computationally complex, leading to performance issues in real-time systems, particularly those with low processing resources. Another major issue is the storage and bandwidth limitations of CCTV devices. Videos obtained by high-resolution surveillance systems require considerable storage capacity and bandwidth, thereby requiring the implementation of compression methods. Nevertheless, combining compression and encryption presents additional issues in preserving the security and integrity of CCTV video. The efficiency of selective encryption techniques is also a major issue among all as most of the selective encryption techniques cannot provide a secure and efficient solution. When only a selected piece of a CCTV video is encrypted, it can expose the other portions to attack, particularly in situations where unencrypted data might be used to deduce critical information.

Therefore, the objective of the study is to strenghten the security of CCTV video surveillance systems by boosting data integrity. Therefore, this study presents substantial results and highlights the importance of security in IoT based CCTV video monitoring systems. Providing users with reliable data and practical solutions to protect the safety and durability of their digital data is the primary goal [8],[9]. To better understand the complex aspects of this security problem, research is being done on data integrity, confidentiality, and the dynamic management of cyber threats. Robust security measures are necessary to counter possible risks including illegal access, data breaches, and privacy infringements, as this increase in data often contains sensitive information [10],[11].

This paper is structured as follows. First the introduction is defined, then "Related Work" reviews the existing literary works."Proposed Scheme" is the next section that addresses the suggested approach. The experimental findings are presented in "Experimental Results and Discussion," which also illustrates the system's suggested output. Lastly, we covered the conclusion in the last section.

II. Related Work

Gbashi et al. [12] proposed a novel, lightweight encryption technique for video frames that makes use of the ChaCha20 algorithm and a hybrid chaotic system. While 2D chaotic systems expand key space, the chaos system is used to produce encryption and seed keys. The encryption algorithm is effective against statistical attacks as shown in the results. The suggested approach is efficient in statistical measures and encryption effectiveness. Still, there is a need to apply this method to high-quality videos to reduce complexity so that it can be used in real time applications.

To further improve the encryption technique, Cheng et al. [13] proposed a selective video encryption method that relies on the integration of four-dimensional hyperchaotic systems and a video coding algorithm. This scheme works in two modes. One for the small encrypted data and less security required and the other for the large amount of encrypted data and sufficient real time security. According to the author's findings, this scheme is better than other existing video encryption methods in performance and encryption efficiency. But still, this method lacks real time implementation as all the results are proved by theoretical analysis only.

Lee and Park [14] discussed the importance of machine learning in the continuous growth of CCTV video surveillance systems. The efficiency of surveillance systems is enhanced by using cloud based video monitoring systems, but privacy and security issues become major concerns. Therefore, for processing CCTV video data, blockchain based method is proposed that not only ensures privacy but also synchronizes large volumes of video data. The proposed method supports the delta update function and is durable against attacks by encrypting transmitted data and thus suitable for intelligent video surveillance systems as it reduces the bandwidth required for transmission. But for real time implementation, more accurate video analysis will be required.

To provide reliable security for multimedia data, El-Shafai et al. [15] proposed a 3DV compression-encryption system for IoT wireless networks. This scheme employs compression and a robust encryption algorithm using symmetric keys. The proposed scheme demonstrated several advantages, including easy implementation, high key sensitivity, efficient encryption for video and image data, and resistance against various network attacks. Simulation results confirm its efficiency against different attacks, making it suitable for securing IoT multimedia applications. But still, there is a need to test this method on different IoT networks and also for real time multimedia applications.

To improve the security and privacy of CCTV surveillance systems Sivalakshmi et al. [16] suggest an approach that combines the Privacy Region Mask Separation method and an adaptive silhouette-blur scheme. These techniques help to identify sensitive regions in video frames while ensuring the privacy of video data. To further enhance the privacy protection process, an optimized version of the Coati optimization is integrated with these techniques. To further improve data storage and access, H.264/AVC standard is used for video encoding. This method performs better than RSA and ECC if talking about encryption, decryption speed, and quality preservation, but did not apply to real time data.

A new method has been proposed by Yu and Kim [17] to overcome the limitations of the traditional Region of Interest (ROI) encryption process in high-efficiency video coding (HEVC)/H.265. Encryption is applied to wider areas using traditional methods that result in the inefficiency of resources. Therefore, specific coding units are encrypted in the proposed method to improve encryption speed and reduce computation. But this method is limited to specific coding standards, and not applied to real world data.

To improve data security, the effectiveness of the selective encryption method is evaluated and processing overhead is reduced. The effect of selective encryption on data security, integrity, and accessibility is quantified. The literature review with contributions and limitations is depicted in Table I. The main aim of this paper is to optimize security and performance by analyzing encryption algorithms when applied to certain video segments. The findings will contribute in the comprehension of the efficiency of selective encryption in securing confidential data and optimizing computational capacity.

TABLE I LITERATURE REVIEW WITH CONTRIBUTIONS AND LIMITATIONS

References	Techniques	Contributions	Limitations
[12]	ChaCha20 algorithm and a hybrid chaotic system	• Effective against statistical attacks	• Not applicable in real time applications.
[13]	Selective video encryption method	• Performance and encryption efficiency	• Not applicable in real time applications.
[14]	Blockchain based method	• Suitable for intelligent video surveillance systems as it reduces the bandwidth required for transmission	• Not applicable in real time applications, • More accurate video analysis will be required
[15]	3DV compression-encryption system	• Easy implementation, • High key sensitivity, • Efficient encryption for video and image data, and • Resistance against various network attacks	• Not tested this method on different IoT networks and • Not tested for real time multimedia applications.
[18]	H.266/VVC (Versatile Video Coding) with a layered multiple tree structure	• Enhanced Compression Efficiency • Low Latency Streaming • Bandwidth Reduction	• Increased Computational Requirements • Scalability Challenges • Trade-offs in Quality and Efficiency
[16]	Privacy Region Mask Separation method, an adaptive silhouette-blur scheme, and Improved Coati Optimization Algorithm with H.264/AVC standard	• Enhanced Privacy Protection • Optimized Performance • Efficient Video Encoding and Storage • Improved Encryption and Decryption Speed • Higher Image Quality	• Lack of Real-Time Data Validation • Limited Scope of Video Scenarios • Computational Complexity • Restricted Focus on Privacy Regions
[17]	Coding Unit (CU)-Based Encryption, Region of Interest (ROI) Detection using YOLOv4, and HEVC/H.265 Video Encoding	• Selective Encryption for Improved Efficiency • Enhanced Privacy Protection Improved Video Quality Reduction in Encryption Time and Resources	• Limited to Specific Video Coding Standards • Increased Complexity in Real-Time Scenarios • Limited Testing on Real-World Data

III. METHODOLOGY

The proposed methodology addresses data security and privacy concerns in CCTV surveillance systems. The main goal is to strengthen CCTV security by using selective encryption methods. Fig. 1 depicts the flowchart for the cryptography process. This method encrypts specific parts of the video, protecting sensitive information and maintaining access to critical data for real-time analysis. The methodology is designed to optimize computational overhead and ensure the safety of user data. The steps for methodology are given below and the compression and selective encryption process for IoT CCTV video data is depicted in Fig. 2.

A. Data Preparation

A dataset of CCTV footage was collected from an 1800 square feet area, including indoor and outdoor cameras, with different lighting conditions and camera angles to reflect real-world surveillance scenarios.

The captured CCTV videos have different sizes, resolutions, lengths, frame counts, and frame rates to meet real world needs.

1) Compression using H.264/AVC (Advanced video coding): CCTV video was compressed using the H.264/AVC video compression standard, a widely adopted standard for video compression.

High compression efficiency is achieved with H.264/AVC as both spatial (within a single frame) and temporal (between successive frames) redundancies are removed, reducing file size while maintaining quality.

Bitrate, one of the crucial compression parameters, was varied to balance compression level and video quality. Lower bitrates were used to minimize storage space while maintaining sufficient quality for real-time surveillance needs.

B. Dynamic Key Management

Develop a dynamic key management system that generates and manages encryption keys. These keys should be unique to each encrypted portion, ensuring adaptability to changing security requirements and user preferences. The dynamic nature of key management enhances the overall security posture of the system.

C. Selective Encryption Technique

The Selective Encryption Technique is used to encrypt only the determined video frames. Unlike traditional methods that encrypt the entire stream, this approach strategically targets specific elements based on frame interval, reducing the computational burden and preserving critical information for real-time analysis.

The AES-256 (Advanced Encryption Standard) algorithm was used for encryption as it balances security and computational efficiency.

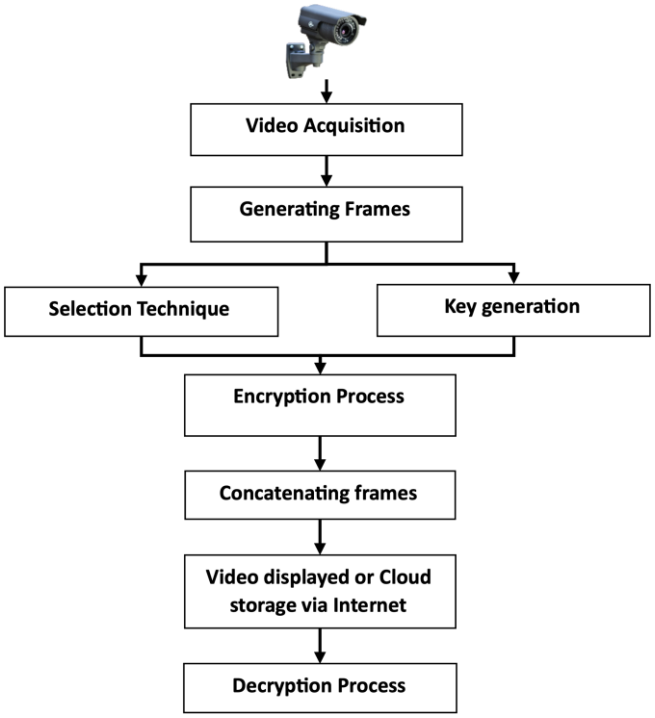


Fig. 1. Cryptography process.

D. Performance Evaluation

The system was evaluated based on the following metrics to assess both video quality and encryption performance:

1) Entropy: This metric measures how much disorder or information is present in the image. Higher entropy means more pixel complexity, whereas more regular patterns are an indication of lower entropy.

2) Correlation analysis: The correlation coefficient describes the relationships between adjacent video frames and the distribution of pixels. The lack of association between frames indicates that there is no pattern at all, making statistical cryptanalysis impossible.

3) Bit rate overhead analysis: The bit rate overhead measures the variation in bit rate before and after encryption. This demonstrates how well the encryption method reduces file size.

4) Encryption efficiency: The time taken for both the compression and selective encryption processes was measured to ensure the method is feasible for real-time CCTV applications.

5) Decryption efficiency: The time taken for decryption processes was measured to ensure the method is quickly and easily retrieves encrypted video footage and well suited for real-time CCTV applications.

6) Histogram analysis: Using the pixel distribution in each frame, the histogram analysis describes the visual relationships between the original and encrypted frames. Typically, cryptanalysts use the distribution of pixels to carry out statistical attacks.

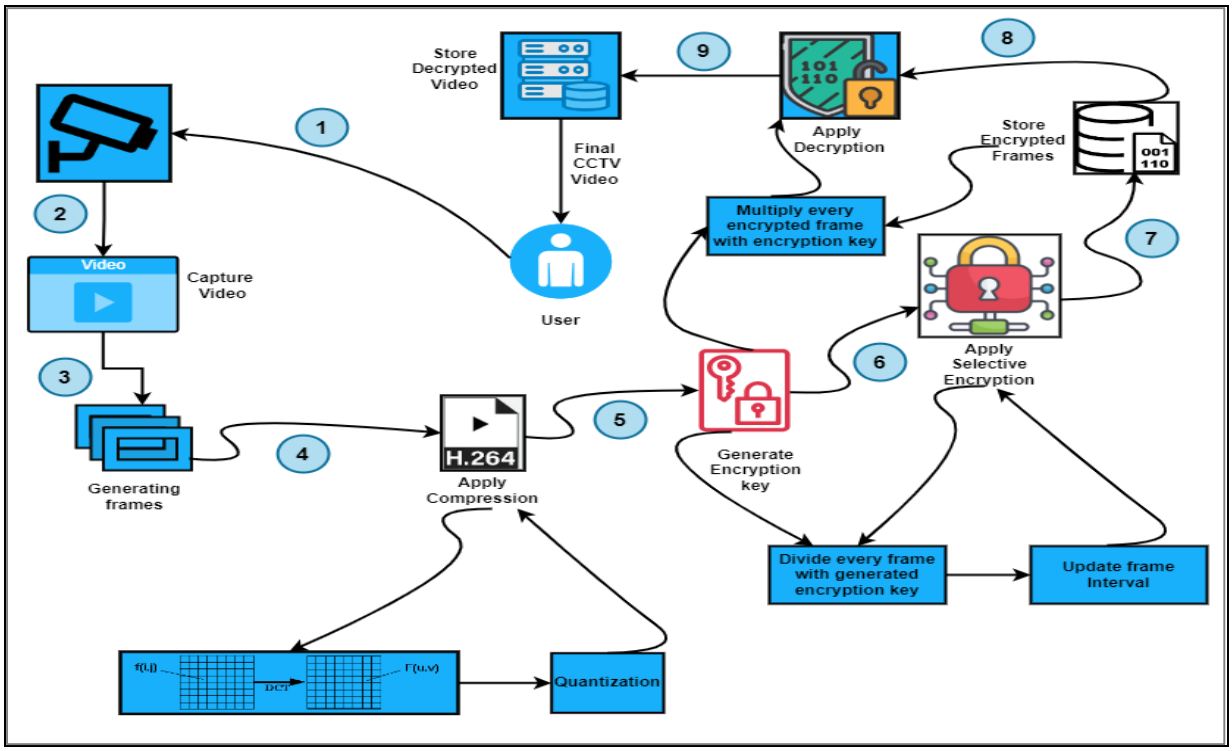


Fig. 2. Compression and selective encryption process for IoT CCTV video data.

IV. Proposed Algorithm

A systematic approach is applied to optimize data size and protect video data by using the proposed video encryption and compression algorithm. First, video frames are compressed to optimize their efficiency in terms of storage. Subsequently, the selected video data is encrypted with a randomly generated key to protect its confidentiality and integrity. The decryption process is carried out after encryption to retrieve the original content. The entire process is simplified as it enables the safe transfer and storage of video frames without additional decompression. Fig. 3 describes the flowchart for the proposed Algorithm 1.

Algorithm 1: Proposed Algorithm	
Step1: Read Original Video Frames	• O = {O1, O2, ..., On}
Step2: Initialize Variables	• x2 = 15 • x = frames per second (FPS)
Step 3: Compression Operation	For each frame Oi in O: • Compress every frame Oi by compression algorithm
Step 4: Encryption Key Generation	• Generate a random key matrix (K).
Step 5: Encryption Operation	For each frame Oi in O: • Calculate x1 = $int(current\frac{frame}{x}+3)$ • Encryption operation: $Ei[x,y] = Oi[x,y]/K[x,y]$ • Update x2 = x2 + 15
Step 6: Create Encrypted Video Frames	• E = {E1, E2, ..., En}
Step 7: Decryption Operation	For each decrypted frame Di in D: • Decryption operation: $Di[x,y]=Ei[x,y]\times K[x,y]$

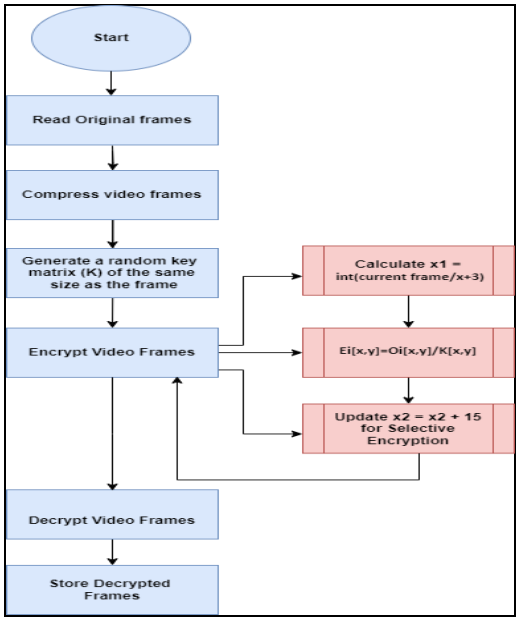


Fig. 3. Flowchart for the proposed algorithm.

V. Experimental Results and Evaluation

The proposed scheme analyses five CCTV videos from different cameras that are described in Table II. The captured CCTV videos have different sizes, resolutions, lengths, frame counts, and frame rates. The need for higher resolution, longer recording time, or considerations for storage space are specific requirements of the surveillance system. The experimental run on Intel Core i5 CPU @ 2.4GHz and 8 GB RAM on Windows 10 OS.

In this experiment, an 1800 square feet area in the front and garden is considered, and an encryption technique is used to optimize security. To evaluate the efficiency of the proposed system, five cameras were positioned in important areas. To evaluate the algorithm's capacity by optimizing encryption time and storage efficiency and its effect on data security is the main aim of this paper. Results showed how well the encryption system protected CCTV video and optimized security.

TABLE II Different CCTV Videos

CCTV Video	Video size(KB)	Video Length (seconds)	Frame Count	Frame Rate/sec	Resolution
CCTV1	2528	29	731	25	640×352
CCTV2	6028	30	756	25	848×480
CCTV3	4942	24	604	25	848×480
CCTV4	2429	23	357	15	352×288
CCTV5	2538	24	363	15	352×288

A. Compression

CCTV video files are compressed to reduce file size to maximize storage before encryption. With this process, important information is retained and unnecessary information is removed. The file size is reduced by using compression techniques to reduce file size without reducing quality. Various compression algorithms are available to reduce the size of a video. Intra-frame and Inter-frame compression techniques are the most widely adopted techniques. Intra-frame compression which reduces the size of individual frames by using spatial redundancy within specific frames and is implemented by Discrete Cosine Transform (DCT) and Wavelet Transform. In the DCT method, pixel data is transformed into a set of low and high frequencies, whereas wavelet transform as the name suggests uses wavelets to convert the pixel data into wavelet coefficients to reduce the precision of less important data. On the other hand, Inter-frame compression uses temporal similarities to decrease the redundancy between frames. Methods used for Inter-frame compression are Motion Estimation and Compensation, Predictive and Transform Coding. Motion Compensation stores only the variation between frames where motion is detected. Predictive coding uses prediction from the previous encoded frame and the next frame. On the contray transform coding, a frame is transformed, and then transformed coefficients are quantized and encoded.

Modern video standards such as H.264/AVC (Advanced Video Coding) and HEVC (High-Efficiency Video Coding) or H.265 are combined with these compression techniques to balance video quality and compression. Higher compression efficiency is achieved by HEVC, but advance hardware is

required to implement this in CCTV. Therefore H.264/AVC is used in the proposed algorithm to balance between compression and quality, and it supports existing hardware and software. The selective encryption algorithm is used after compression to protect the data from manipulation or unwanted access. This combined approach ensures effective storage, less bandwidth requirements, and enhanced CCTV security.

B. Entropy

The degree of unpredictability or randomness in the video frame's pixel values describes entropy. It measures how much disorder or information is present in the image. Higher entropy means more pixel complexity, whereas more regular patterns are an indication of lower entropy[19]. Entropy of different CCTV videos is shown in Table III.

The following formula can be used to calculate entropy(H):

$$H=-\sum_{i=1}^{n} Pi*\log2(Pi)$$

Where:

- n is the count of unique pixel values in the frame.
- Pi is the probability of occurrence of each unique pixel value.

TABLE III ENTROPY OF DIFFERENT CCTV VIDEOS

CCTV Videos	Entropy of the frame(bits per pixel)
CCTV1	7.6841
CCTV2	7.7560
CCTV3	7.7914
CCTV4	7.6794
CCTV5	7.0398

The frame may contain diverse patterns, textures, or information. Higher entropy is generally considered good for encryption. It suggests that the pixel values in the frame are more random and less predictable, which aligns with the goal of encryption, thus making the content difficult to discern without the decryption key.

Fig. 4 makes it clear that, out of all the CCTV feeds, CCTV3 has the highest entropy. The entropy values which are closer to eight is considered to be an ideal value as highlighted by Gbashi et al. [12]. The reason for this increased entropy is that CCTV3 comes from the front door camera, which records a feed with more diverse material. CCTV5, on the other hand, has the lowest entropy due to the backyard camera. The reason is less amount of variation in information is captured resulting in a more uniform feed and lower entropy. The changes in entropy levels of the CCTV videos depend upon different video quality, camera placement, and scene detailing.

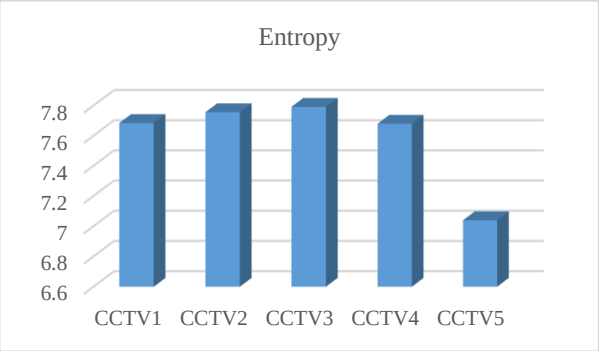


Fig. 4. Frame entropy (bits per pixel).

C. Correlation Analysis

The correlation coefficient describes the relationships between adjacent video frames and the distribution of pixels. The correlation value tells the relationship between neighboring frames. A higher correlation value means a closer linear link and a lower correlation value defines a more nonlinear relationship [12]. The correlation between pixels at matching positions is calculated in adjacent frames, and then a new frame with less data is created using the variations between them. Using this method results in a near-zero correlation value and a nonlinear relationship between neighboring frames.

Table IV presents the experimental results for nearby frame correlation coefficients for both the original and encrypted frames. These findings imply that the original and encrypted frames show very little association. As a result, this lack of association indicates that there is no pattern at all, making statistical cryptanalysis impossible.

A correlation coefficient that is close to zero, like in the case of CCTV1 frame 50, CCTV2 frame 50, CCTV3 frame 10 of CCTV3, CCTV4 frame 100, and CCTV5, denotes the lack or very weak linear relationship between the original and encrypted frames depicted in Fig. 5. The correlation coefficient is a statistical measure that expresses the direction and intensity of a linear relationship between two variables. Values that are close to 0 in this context indicate that there is little to no linear relationship between the encrypted and original frame contents. This is good news for encryption because it means that the content of the original frame is successfully hidden during the encryption process. The more closely the correlation coefficient approaches zero, the more skillfully the encryption process masks any observable links or patterns in the data as described by Gbashi et al. [12]. A negative correlation coefficient in the context of encryption would still suggest that the relationship between the original and encrypted frames has been effectively obscured by the encryption process, making it challenging for an observer to deduce the original frame's content from the encrypted frame without the decryption key.

TABLE IV COMPARISON SIMULATION RESULTS OF THE CORRELATION COEFFICIENT

Random Frames	CCTV1	CCTV2	CCTV3	CCTV4	CCTV5
Frame 0	-0.001182	0.0003209	-0.0004950	0.0003168	0.0019212
Frame 10	-0.001530	0.00019326	-0.0001918	0.0019271	0.0027840
Frame 50	-0.000273	-0.0005638	-0.0006569	-0.0025389	0.0009055
Frame 100	-0.001595	-0.0033022	0.0006837	-0.0022497	0.0007022

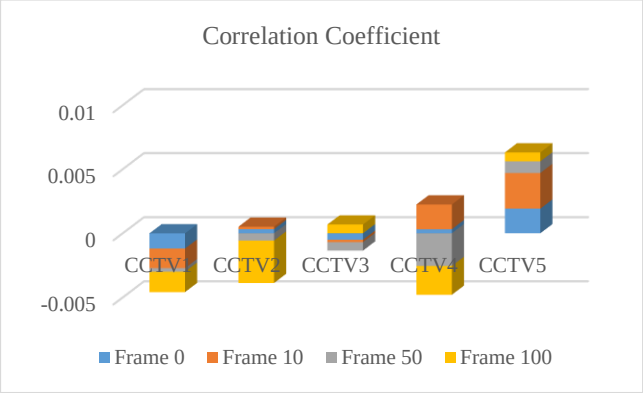


Fig. 5. Comparison results of the correlation coefficient.

D. Bit Rate Overhead Analysis

The bit rate overhead measures the variation in bit rate before and after encryption It is expressed as a percentage change relative to the original bit rate as stated in Table V. The sign of the overhead value indicates the direction of the change.

1) Negative bit rate overhead: A negative overhead number indicates a drop in the bit rate after the operation. Negative overhead in the context of video compression or encryption means that the processed video uses fewer bits than the original, which lowers the file size or bit rate. In general, this is preferable since it suggests effective encryption without appreciable quality loss. Negative overhead in video encryption reduces bit rate, resulting in smaller files and better quality without significant loss of quality.

2) Positive bit rate overhead: Positive overhead in video encryption indicates an increase in bit rate, possibly due to additional data added, potentially indicating a need for quality improvement or security measures. Often, the objective is to accomplish efficient encryption or compression while limiting the influence on file size and preserving a sufficient level of video quality [20].

The bit rate is reduced by 90% in the first video, and by more than 40% in all other scenarios seen in Fig. 6. This implies that the video file was significantly compressed without sacrificing the visual quality. This demonstrates how well the encryption method reduces file size. In general, negative overhead levels are better since they show a reduction in file size or bit rate without noticeably lowering quality as suggested by Chen et al. [20].

TABLE V BIT RATE OVERHEAD OF DIFFERENT CCTV VIDEOS

Video	Original	Encryption	Overhead
CCTV1	712000	650000	-0.9087
CCTV2	1644000	630000	-0.6167
CCTV3	1685000	609000	-0.6385
CCTV4	827000	487000	-0.4111
CCTV5	864000	488000	-0.4351

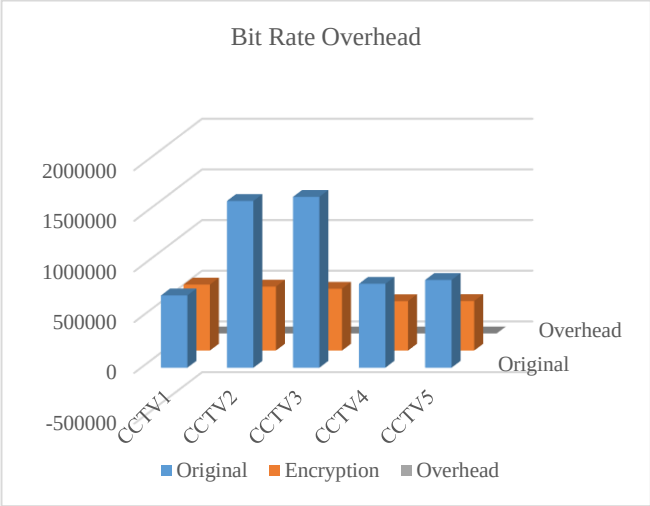


Fig. 6. Bit rate overhead of different CCTV videos.

E. Encryption Efficiency

Video encryption time overhead is dependent on various aspects, including the encryption algorithm, frame rate, resolution, and compression format. Encryption takes longer in videos with higher resolution and frame rate. Encryption time is affected by many factors. One is Compression as it adds more steps to the processing and the others are the security level, key length, complexity of the video, and encryption algorithm used. Besides these, buffering and network performance are crucial where real-time encryption is done [21]. The encryption time overhead in CCTV surveillance applications is optimized as described in Table VI and it is clearly shown in Fig. 7.

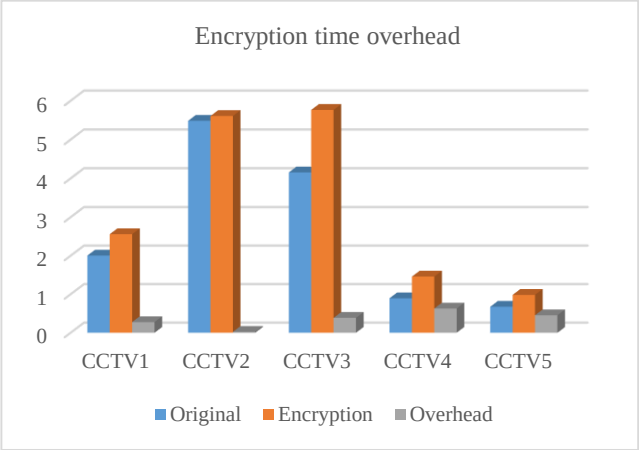


Fig. 7. Encryption time overhead of CCTV videos.

Table VII presents the encryption speed for five different CCTV videos and describes the variation in encryption efficiency for randomly chosen frames. This eventually helps to effectively safeguard sensitive video data in real-time surveillance scenarios.

TABLE VI ENCRYPTION TIME OVERHEAD OF DIFFERENT CCTV VIDEOS

Video	Original	Encryption	Overhead
CCTV1	2.0040	2.5594	0.2771
CCTV2	5.4845	5.6084	0.0225
CCTV3	4.1545	5.7645	0.3875
CCTV4	0.8952	1.4594	0.6302
CCTV5	0.6766	0.9837	0.4538

TABLE VII RESULT OF ENCRYPTION TIME (IN SECONDS) FOR DIFFERENT CCTV FRAMES WITH AVERAGE ENCRYPTION TIME

CCTV Videos	Frame 0	Frame 50	Frame 100	Frame 150	Average
CCTV1	0.04803	0.03464	0.03753	0.03357	0.00171
CCTV2	0.08812	0.06305	0.05923	0.05731	0.00261
CCTV3	0.12913	0.05824	0.05576	0.05399	0.00245
CCTV4	0.04807	0.02142	0.02138	0.02554	0.00179
CCTV5	0.03423	0.02149	0.01954	0.01931	0.00178

Fig. 8 provides a comparative comparison of the encryption speed times for randomly picked frames from different CCTV videos, expressed in frames per second. Interestingly, every CCTV video shows varying encryption speed performances at various frame rates, that is due to the randomness of pixels in every frame as every video is from a different CCTV video. When compared to other CCTV movies in the dataset, CCTV1, CCTV4, and CCTV5 perform better on average, with the quickest encryption speed time of 0.00171, 0.00179, and 0.00178 respectively frames per second. The encryption algorithm being utilized is one of the possible causes of these variances in encryption speed timings. With the fastest and most reliable encryption speeds, CCTV5 is the best choice for real-time applications. Reliability is further demonstrated by CCTV4 and CCTV1, with low and constant encryption times. Given the variations in video size and resolution, CCTV2 and CCTV3 nevertheless offer performance appropriate for real-time applications despite longer encryption periods as demonstrated by Alawi and Hassan [22].

F. Decryption Efficiency

The proposed algorithm uses optimized decryption procedures that are specially designed for the chosen encryption technique in order to prioritize decryption efficiency. The information displayed in Table VIII and Fig. 9 demonstrates how well the algorithm performs in terms of decryption efficiency and speed. By use of enhanced decryption algorithm and effective cryptographic key management, the algorithm guarantees the prompt restoration of original content without sacrificing security. As a result, the method is particularly effective in decrypting data, making it possible to quickly and easily retrieve encrypted video footage.

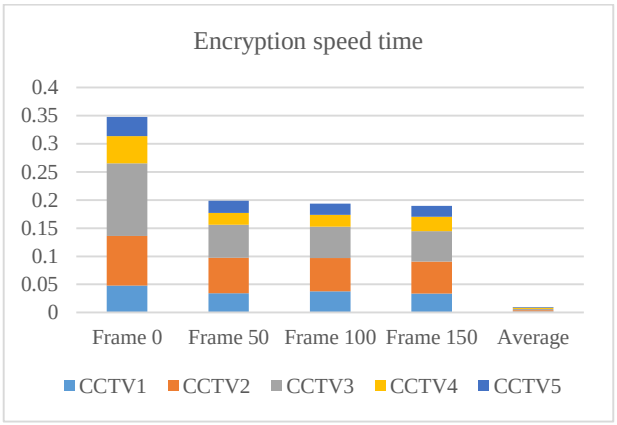


Fig. 8. Result of encryption time (in seconds) for different CCTV frames with average encryption time.

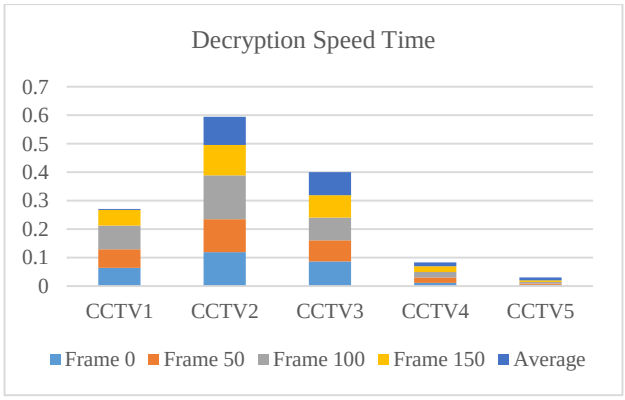


Fig. 9. Result of decryption time (in seconds) for CCTV video frames.

TABLE VIII RESULT OF DECRYPTION TIME (IN SECONDS) FOR DIFFERENT CCTV VIDEO FRAMES

CCTV Videos	Frame 0	Frame 50	Frame 100	Frame 150	Average
CCTV1	0.06465	0.06517	0.08280	0.05468	0.00400
CCTV2	0.11907	0.11594	0.15378	0.10657	0.09978
CCTV3	0.08690	0.07434	0.07906	0.07852	0.08153
CCTV4	0.01116	0.01973	0.01913	0.01973	0.01409
CCTV5	0.00499	0.00399	0.00618	0.00573	0.00930

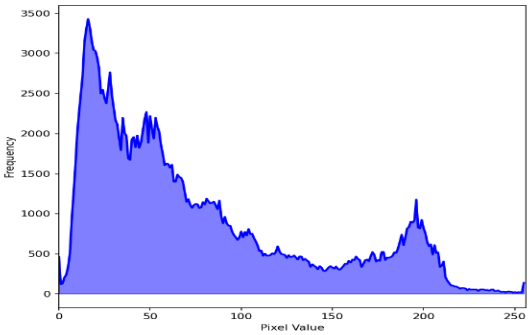
Every CCTV video is suitable for real-time applications; the reasons for variations in decryption durations are related to the sizes and formats of the videos. CCTV5 regularly has the smallest decryption timings (0.00399 to 0.00618 seconds), while CCTV2 consistently has the highest decryption times (0.10657 to 0.15378 seconds). With the fastest, most reliable decryption times, CCTV5 exhibits the best performance. Additionally dependable and with quick decryption times is CCTV4. While CCTV3 and CCTV2 have longer and more inconsistent decryption times but are still appropriate for real-time application as described by Alawi and Hassan [22].

G. Histogram Analysis

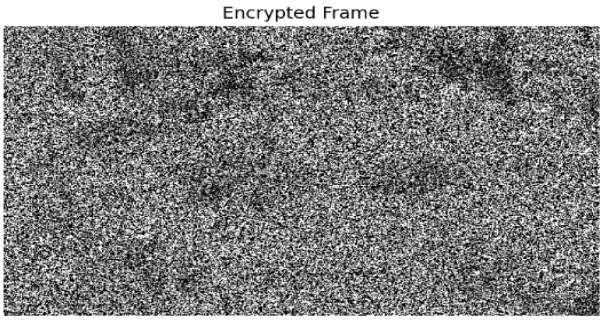
Using the pixel distribution in each frame, the histogram analysis describes the visual relationships between the original and encrypted frames. In an attempt to recover the original frames, statistical attacks try to take advantage of this predicted relationship. The notable differences between the encrypted frame histograms and the original frame histograms highlight how well the encryption approach works to fend off statistical attacks [23]. The chi-square test is used to demonstrate the homogeneity of pixels and is used in the study of the histogram.

Typically, cryptanalysts use the distribution of pixels to carry out statistical attacks. It is clear from the figures how the original and encrypted video frames' histograms differ from one another. This prevents statistical attacks on the video frames. The original video frame is shown in Fig. 10(a) of the supplied figures, while Fig. 10(a1) shows the histogram of the frame before encryption. The histogram of the encrypted video frame (b) is shown in Fig. 10(b1). These figures make it clear that the histograms of original and encrypted video frames are distinct. This implies that video frames are secured against any statistical attacks as stated by Ravikumar and Kavita [24].

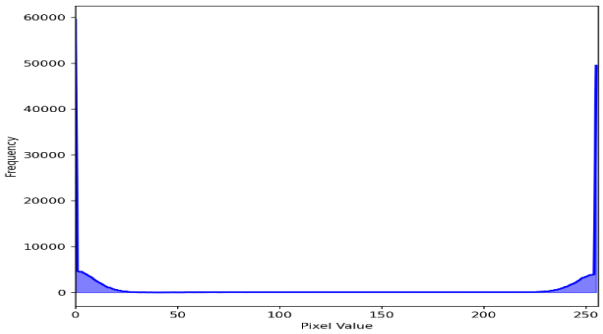
a)



a1)



b)



b1)

Fig. 10. a) Original frame, a1) Histogram of the original frame, b) Encrypted Frame, b1) Histogram of encrypted frame.

VI. COMPARATIVE ANALYSIS AND DISCUSSIONS

The experiments are carried out to compare the results with state-of-the-art algorithms to confirm the suggested scheme efficiency for CCTV video security. Encryption time, decryption time, and correlation coefficient results of the proposed technique are compared statistically with the algorithms from current related literature. For the tested CCTV videos, Table IX presents the numerical values of the proposed

scheme for Encryption time and decryption time per frame in comparison to the suggested techniques in recent literature.

TABLE IX COMPARISON OF AVERAGE ENCRYPTION AND DECRYPTION TIME IN SECONDS

Reference	Average Encryption Time per frame	Average Decryption Time per frame
[25]	0.01219	0.01206
[15]	0.7306	0.6278
[26]	0.03549	0.03625
CCTV1	0.00171	0.00400
CCTV2	0.00261	0.09978
CCTV3	0.00245	0.08153
CCTV4	0.00179	0.01409
CCTV5	0.00178	0.00930

It can be seen from Fig. 11 that the encryption time of the proposed technique is better than what is proposed in the literature and very close to zero. Therefore, the proposed technique is successful in optimizing encryption time and is well suited for real time applications.

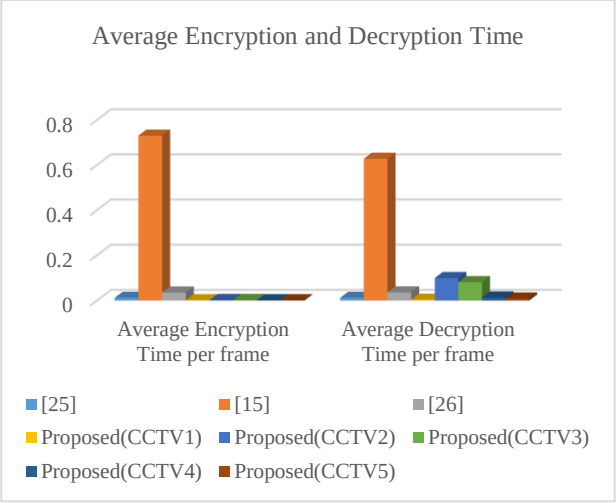


Fig. 11. Comparison of average encryption and decryption time (in seconds) for CCTV video frames.

The comparison of correlation coeficients with the recent literature are given in Table X. Moreover, it is also observed that correlation coefficient of the proposed algorithm is better than the schemes in literature as shown in Fig. 12. The correlation coefficients of the proposed scheme are very close to zero which prevents it from statistical attacks.

TABLE X COMPARISON OF CORRELATION COEFFICIENTS

Reference	Correlation Coefficient
[25]	-0.0045
[15]	-0.0029
[26]	0.00733
Proposed(CCTV1)	-0.0011
Proposed(CCTV2)	0.00032
Proposed(CCTV3)	-0.0004
Proposed(CCTV4)	0.00031
Proposed(CCTV5)	0.00192

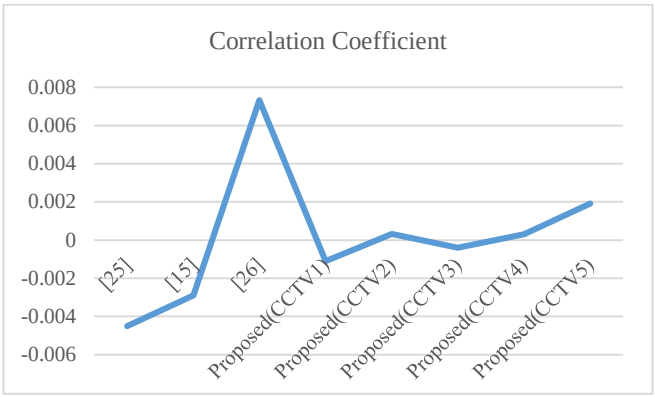


Fig. 12. Comparison of correlation coefficients for CCTV video frames.

The proposed technique works well according to different video resolutions, frame rates, or compression standards as encryption process is done after compression. Individual video frames are encrypted using our suggested frame-level selective encryption technique. This preserves adaptability across various video resolutions and frame rates while guaranteeing independence from video codecs and compression standards. This preserves adaptability across various video resolutions and frame rates while guaranteeing independence from video codecs and compression standards. Furthermore, scalability is demonstrated by the consistent performance across measured resolutions (640×352, 848×480, and 352×288), although frame rate differences mostly impact temporal density rather than the encryption process. Consequently, the design of the approach guarantees strong applicability across a variety of video setups and compression standards.

There are significant trade-offs when employing selective encryption in IoT CCTV devices. Higher video resolutions or frame rates result in a greater computational cost since more data must be processed, even though the proposed approach is quicker and more effective than encrypting the entire video. This may cause encryption on low-power IoT CCTV devices to lag. However, to minimize this trade-off, encryption, and decryption process is minimized. The technique is beneficial for real-time streaming because it avoids the delays associated with full encryption, making it suitable for real time applications. Overall, although there is a trade-off between computational speed and the volume of data being encrypted, the proposed technique effectively balances these needs for CCTV surveillance systems. At the same time, it maintains compression efficiency, which helps save storage.

By encrypting selected video data, the proposed selective encryption technique disrupts correlations in the video data and shows resilience against statistical attacks. The robust AES algorithm used in this technique has a strong cryptographic design that makes it naturally resistant to adaptive chosen-plaintext attacks. Because of the avalanche effect of AES, even a small alteration to the plaintext produces an entirely different ciphertext, making it nearly impossible for an attacker to deduce significant connections between the selected plaintexts and their matching ciphertexts, and even the encryption keys. The use of compression before encryption substantially strengthens the method's resistance to chosen plaintext attacks in the setting of selective encryption. Non-linear changes are

applied to video data by using compression resulting in scrambling the data and eliminating redundancies in the plaintext frames. Because of this procedure, it is very difficult for an attacker to create plaintext frames that might expose ciphertext. In terms of key management, the technique is made to work with safe key distribution and exchange protocols that are used in real time environments such as Diffie-Hellman key exchange. In order to prevent attacks like key reuse or interception, these protocols offer safe methods for creating, allocating, and rotating encryption keys.

The method reduces risks in two ways for real-world attack scenarios where an adversary might try to deduce encrypted content from unencrypted frames or statistical redundancy. First, the attacker cannot reconstruct meaningful content from unencrypted frames alone since selective encryption is applied which disrupts important visual information. Second, applying compression before encryption removes correlations and redundancies from the CCTV video data and it further distorts patterns that may otherwise be exploited.

VII. Conclusion

This paper highlights significant discoveries and shows the importance of security in IoT based CCTV video surveillance systems. This study offers a new method for video encryption that combines selective encryption with compression to enhance security and efficiency. The approach is ideal for demanding real-time applications as demonstrated by experiments, since it can selectively encrypt data after compression, which minimizes computational overhead and hence guarantees efficient utilization of system resources. Through testing and analysis, the algorithm's effectiveness and resistance to statistical attacks is proven. Still, it needs constant monitoring and improvement to strengthen its defenses against new threats. With the promise of increased security in a world growing more and more data-centric, this study represents a major advancement in the support of video encryption techniques. In the future, adding authentication techniques to encryption will be crucial to enhancing security and guaranteeing that sensitive data is only accessed by authorized individuals.

Acknowledgment

This work was supported by the Deanship of Scientific Research, the Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia under the project KFU250656.

Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

[1] A. Zhaxalikov, A. Mombekov, and Z. Sotsial, "Surveillance Camera Using Wi-Fi Connection," Procedia Comput. Sci., vol. 231, pp. 721–726, Jan. 2024, doi: 10.1016/J.PROCS.2023.12.147.

[2] Y. Myagmar-Ochir and W. Kim, "A Survey of Video Surveillance Systems in Smart City," Electronics (Switzerland), vol. 12, no. 17. Multidisciplinary Digital Publishing Institute, p. 3567, Aug. 23, 2023. doi: 10.3390/electronics12173567.

[3] A. Kumar, S. Sharma, N. Goyal, A. Singh, X. Cheng, and P. Singh, "Secure and energy-efficient smart building architecture with emerging technology IoT," Comput. Commun., vol. 176, pp. 207–217, Aug. 2021, doi: 10.1016/j.comcom.2021.06.003.

[4] D. Dhingra and M. Dua, "A chaos-based novel approach to video encryption using dynamic S-box," Multimed. Tools Appl., vol. 83, no. 1, pp. 1693–1723, Jan. 2024, doi: 10.1007/S11042-023-15593-6/METRICS.

[5] [5]J. Dai, Q. Li, H. Wang, and L. Liu, "Understanding images of surveillance devices in the wild," Knowledge-Based Syst., vol. 284, p. 111226, Jan. 2024, doi: 10.1016/J.KNOSYS.2023.111226.

[6] M. Surya Priya, D. Diana Josephine, and P. Abinaya, "IOT Based Smart and Secure Surveillance System Using Video Summarization," in Lecture Notes in Electrical Engineering, 2021, vol. 735 LNEE, pp. 423–435. doi: 10.1007/978-981-33-6977-1_32.

[7] S. Rani, S. H. Ahmed, and R. Rastogi, "Dynamic clustering approach based on wireless sensor networks genetic algorithm for IoT applications," Wirel. Networks, vol. 26, no. 4, pp. 2307–2316, May 2020, doi: 10.1007/S11276-019-02083-7/METRICS.

[8] C. segun ODEYEMI, B. A. OMODUNBI, O. M. OLANIYAN, and A. A. SOLADOYE, "INTERNET OF THINGS (IoT) BASED REMOTE SURVEILLANCE CAMERA FOR SUPERVISION OF EXAMINATIONS," LAUTECH J. Eng. Technol., vol. 18, no. 1, pp. 109–116, May 2024, Accessed: May 22, 2024. [Online]. Available: https://www.laujet.com/index.php/laujet/article/view/634

[9] M. Tahir, Y. Qiao, N. Kanwal, B. Lee, and M. N. Asghar, "Privacy Preserved Video Summarization of Road Traffic Events for IoT Smart Cities," Cryptography, vol. 7, no. 1, p. 7, 2023, doi: 10.3390/cryptography7010007.

[10] S. Gangadharaiah and L. B. Bhajantri, "Secure data dissemination and routing in Internet of Things," Int. J. Inf. Technol., pp. 1–18, Apr. 2024, doi: 10.1007/s41870-024-01848-4.

[11] M. Rana, Q. Mamun, and R. Islam, "Lightweight cryptography in IoT networks: A survey," Future Generation Computer Systems, vol. 129. North-Holland, pp. 77–89, Apr. 01, 2022. doi: 10.1016/j.future.2021.11.011.

[12] E. K. Gbashi, E. Shakir, A. Tariq Maolood, E. Khalaf Gbashi, and E. Shakir Mahmood, "Novel lightweight video encryption method based on ChaCha20 stream cipher and hybrid chaotic map Modeling carbon nanotubes with different structure at millimeter wavelength antennas View project Novel lightweight video encryption method based on ChaCha20 ," Artic. Int. J. Electr. Comput. Eng., vol. 12, no. 5, pp. 4988–5000, 2022, doi: 10.11591/ijece.v12i5.pp4988-5000.

[13] S. Cheng, L. Wang, N. Ao, and Q. Han, "A Selective Video Encryption Scheme Based on Coding Characteristics," Symmetry 2020, Vol. 12, Page 332, vol. 12, no. 3, p. 332, Feb. 2020, doi: 10.3390/SYM12030332.

[14] D. Lee and N. Park, "Blockchain based privacy preserving multimedia intelligent video surveillance using secure Merkle tree," Multimed. Tools Appl., vol. 80, no. 26–27, pp. 34517–34534, Nov. 2021, doi: 10.1007/s11042-020-08776-y.

[15] W. El-Shafai, A. K. Mesrega, H. E. Ahmed, N. Abdelwahab, and F. E. Abd El-Samie, "An efficient multimedia compression-encryption scheme using latin squares for securing internet of things networks," J. Inf. Secur. Appl., vol. 64, no. November 2021, p. 103039, 2022, doi: 10.1016/j.jisa.2021.103039.

[16] M. Sivalakshmi, K. R. Prasad, and C. S. Bindu, "Improved privacy protection technique for enhancing security of real-time video surveillance," J. Inf. Optim. Sci., vol. 45, no. 5, pp. 1389–1399, 2024, doi: 10.47974/jios-1711.

[17] J. Y. Yu and Y. G. Kim, "Coding Unit-Based Region of Interest Encryption in HEVC/H.265 Video," IEEE Access, vol. 11, pp. 47967–47978, 2023, doi: 10.1109/ACCESS.2023.3276243.

[18] S. Sharma, H. Jindal, A. Jain, S. Singh, and P. S. Binner, "Optimizing Video Compression and Transmission for Real-Time Applications," in 2023 IEEE Region 10 Symposium, TENSYMP 2023, 2023. doi: 10.1109/TENSYMP55890.2023.10223679.

[19] K. M. Hosny, M. A. Zaki, N. A. Lashin, and H. M. Hamza, "Fast colored video encryption using block scrambling and multi-key generation," Vis.

Comput., vol. 39, no. 12, pp. 6041–6072, Dec. 2023, doi: 10.1007/s00371-022-02711-y.

[20] C. Chen, X. Wang, G. Liu, and G. Huang, "A Robust Selective Encryption Scheme for H.265/HEVC Video," IEEE Access, vol. 11, pp. 17252–17264, 2023, doi: 10.1109/ACCESS.2022.3210132.

[21] J. Yun and M. Kim, "JLVEA: Lightweight Real-Time Video Stream Encryption Algorithm for Internet of Things," Sensors 2020, Vol. 20, Page 3627, vol. 20, no. 13, p. 3627, Jun. 2020, doi: 10.3390/S20133627.

[22] A. R. Alawi and N. F. Hassan, "A Proposal Video Encryption Using Light Stream Algorithm," Eng. Technol. J., vol. 39, no. 1B, pp. 184–196, 2021, doi: 10.30684/etj.v39i1b.1689.

[23] "Secure and Lightweight Encryption Model for IoT Surveillance Camera by Mohammed Abbas Fadhil Al-Husainy, Bassam Al-Shargabi :: SSRN." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3591979 (accessed Nov. 26, 2022).

[24] S. Ravikumar and D. Kavitha, "IoT based home monitoring system with secure data storage by Keccak–Chaotic sequence in cloud server," Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 7. Springer, pp. 7475–7487, Aug. 02, 2021. doi: 10.1007/s12652-020-02424-x.

[25] A. Hafsa, M. Fradi, A. Sghaier, J. Malek, and M. Machhout, "Real-time video security system using chaos- improved advanced encryption standard (IAES)," Multimed. Tools Appl., vol. 81, no. 2, pp. 2275–2298, Jan. 2022, doi: 10.1007/s11042-021-11668-4.

[26] D. Jiang, T. Chen, Z. Yuan, W. xin Li, H. tao Wang, and L. liang Lu, "Real-time chaotic video encryption based on multi-threaded parallel confusion and diffusion," Inf. Sci. (Ny)., vol. 666, p. 120420, May 2024, doi: 10.1016/j.ins.2024.120420.

Integrating Artificial Intelligence to Automate Pattern Making for Personalized Garment Design

Muyan Han*
School of Fine Arts and Art Design, Qiqihar University, Qiqihar 161000, China

***Abstract*—This paper introduces an innovative AI-assisted pattern construction tool that leverages machine learning models to revolutionize pattern generation in garment design. The proposed system automatically generates patterns from 3D body scans, which are converted into 3D shell meshes and subsequently flattened into 2D patterns using advanced data augmentation techniques and CAD flattening algorithms. This approach eliminates the need for expertise in traditional pattern-making, enabling seamless transformation of 3D models into realistic garment patterns. The tool accommodates various garment styles, including fitted, standard fit, and relaxed fit, while also enabling high levels of personalization by adapting patterns to individual body dimensions. Through its AI-driven automation and user-friendly interface, this plug-in enhances accessibility, allowing individuals without conventional design skills to create customized apparel efficiently.**

Keywords*—Machine learning models; pattern generation; AI-assisted pattern construction; data augmentation techniques; CAD flattening

I. Introduction

CAD computer technologies have become an integral part of modern garment technology and the associated process of pattern making. CAD technologies have revolutionized manufacturing processes in the apparel industry, enhancing precision, efficiency, and innovation. CAD applications provide unprecedented accuracy in designing patterns through sophisticated software and precise digital measurement, whereby patterns of garments closely correspond to human body measurements and are extremely reliable [1]. Moreover, CAD software facilitates collaborative working processes across the apparel and textile supply chain. Designers, pattern makers, and manufacturers can work on digital models simultaneously, improving communication and reducing time-to-market [1]. Additionally, CAD technologies assist in green practices by minimizing the use of physical prototypes because virtual simulations allow organizations to reduce material waste and optimize production processes, thereby assisting in environmentally sustainable garment manufacturing [2].

Despite such advances, there are still some limitations to CAD-based pattern creation. Traditional CAD packages require very specialist expertise, making them inaccessible to others without professional pattern-making training. Present-day computerized tools still require manual adjustment and expert knowledge input in order to design optimized garment patterns. The majority of CAD packages also lack AI-enabled automation, limiting their ability to generate adaptive and highly personalized designs based on individual body shapes [3]. In addition, computer-aided design software is usually expensive and requires significant computational capabilities, making it difficult for individual designers and small industries to use them [3].

To alleviate these difficulties, we introduce an AI-aided pattern-making tool that does not require designers to be experts in traditional pattern-making techniques. This is a revolutionary system where users can input digital models or 3D body scans, which are then processed through an AI-driven pipeline to generate customized garment patterns [4]. The software employs machine learning models, data augmentation techniques, and CAD flattening algorithms to convert 3D shell meshes into 2D patterns automatically. Compared with traditional CAD-pattern generation methods with possible human interventions, our scheme ensures total automation, high degree of personalization, and speedy garment pattern creation [4].

Our AI pattern generator is a giant leap towards the democratization of fashion design technology, bridging the divide between traditional craftsmanship and cutting-edge AI-driven automation. By eliminating technical barriers, this platform provides greater opportunities to more people—the independent designers, fashion enthusiasts, and players in the industry—to explore new design frontiers and contribute to the direction of innovation in garment manufacturing [4].

Our approach is not flawless, however. Even though the AI model considerably reduces the need for hand-based pattern fine-tuning, drastic customizations or complex clothing designs may possibly still require professional adjustment for optimization and fine-tuning. Further, the pattern generation process depends not only on the quality but also on the accuracy of 3D body scans, therefore, low-quality or inconsistent scanning data can also influence the quality of the end patterns [4]. Upgrades in the future will center on AI adaptability enhancement, increased model precision, and widening pattern customization functionalities to better polish the system's potential.

The remainder of this paper is structured as follows: Section II presents a literature review, discussing existing CAD-based pattern generation methods, AI-assisted design tools, and their limitations. Section III outlines the methodology, detailing the proposed AI-driven approach, including data processing, machine learning models, and pattern generation techniques. Section IV covers experimentation and discussion, presenting the implementation details, evaluation metrics, and comparative analysis of results. Finally, Section V concludes the paper by summarizing key findings, highlighting contributions, and suggesting future research directions.

*Corresponding Author

II. Literature Review

A. Traditional Methods in Pattern Construction

Traditional pattern construction in garment engineering is highly craft-based, requiring a great deal of expertise and special techniques. Normally, it is based on a basic pattern that serves as a blueprint for the different types of garments to be constructed [5]. It first involves taking the precise body measurements from which the derivations of the construction parameters for drafting the pattern are obtained. These measurements are interpreted and put onto a two-dimensional pattern that then is adapted to meet the three-dimensional contours of the human body. The first pattern is altered and perfected on a model or mannequin for a correct fit and desired design effect [6]. More often than not, such refinement entails several rounds of alterations and changes until an ideal result is achieved. Traditional pattern-making techniques require great craftsmanship, experience, and technical precision to produce patterns that are to the standard of accuracy and fit required. The process is effective yet very time-consuming and heavily reliant on the skill of the individual [7].

B. Overview of Pattern Systems in Garment Construction

Over the years, a number of pattern systems have been designed, and each of them has its unique methods on garment construction. The demand of the fashion world is maximal flexibility of the selected pattern system. The selection, therefore, depends on the specific criteria, individual tastes, and available means. The best known today is the M. Müller & Sohn pattern system of Michael Müller, founded in Munich in the 19th century [8]. This system, which has gained universal international acceptance, has undergone continued development or reworking to meet present-day industry standards. It includes pattern making for women, men, and children, using a precise series of measurements taken from anatomical measurements and garment-making techniques. Similarly, the Hohenstein pattern system emphasizes the work of basic and model-specific patterns, but it focuses on final pure optical appearances, for example, garment shapes, while still using measurement series.

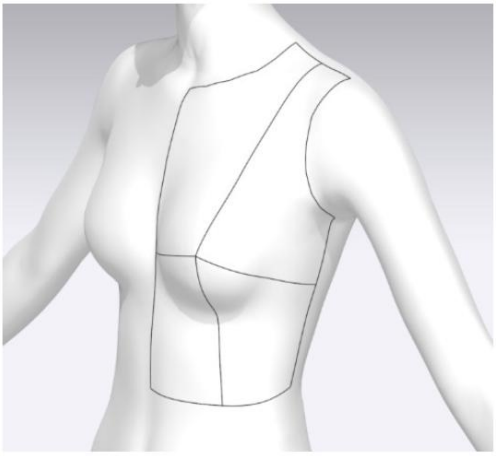


Fig. 1. Segmentation of the upper torso into regions for flattening, highlighting reference points such as the chest point, waistline, and center back and front [1].

One of the more relevant systems is the Optikon, [9] designed as part of a research project undertaken by the Niederrhein University of Applied Sciences. This is a closed-loop system and is versatile because it allows for the development of outer garments for males and females. Essential measurements of the body - chest, waist, hips, height, and type – are related and used to calculate secondary measurements by using formulae based upon the relationships identified in conventional measurement tables. These measures are used to create coordinate systems for construction where human body measures are related to other points around the flat pattern. Considering human measures, along with the unique garment requirements, these systems ensure careful and accurate garment construction with respect to the unique body measurements [10].

C. Flattening

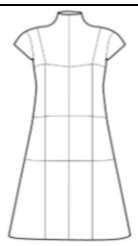
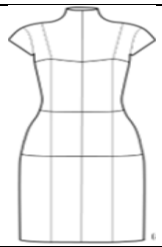
Flattening is one of the important processes in the developing cycle of a garment pattern, which changes a three-dimensional garment shape to a two-dimensional pattern. This technique serves as the base of how design concepts eventually get transferred into a pattern ready for production, allowing the designer and pattern maker to create scaled and accurate designs. This uses mathematics and geometric calculation to project complex 3D shape and curved objects onto flat surfaces with minimum distortion, ensuring accuracy in garment fit and construction [11]. A very effective way to achieve flattening is dividing the upper torso into eight separate regions; the lower part is divided into two additional regions for a total of 12 regions, which can also be doubled to account for both right and left sides of the torso addition resulting in 24 regions. These regions are based on different reference points such as the chest point, waistline, and the front and back centers around which they are oriented as shown in Fig. 1. For instance, such segmentation permits the unfolding of 3D body shapes without distortion as darts will be used in the 2D pattern modeling for reshaping the garment for the human body. The procedure starts with picking critical points such as the bust, the side seams, the waist and hip heights that act as the directives in the division of the torso [12]. This technique has been found to have phenomenal accuracy in developing patterns given the difference in individual body measurements.

Flattening of surfaces is much better than traditional methods, particularly in custom or tailored pattern production. Traditional approaches often require many modification loops and highly rely on the skill and ability of the pattern maker. Inefficiency in flattening is quite the opposite, as it offers a possibly regular and predictable way of arriving at an ideal fit. Nonetheless, issues remain, including the exclusion of necessary allowances when directly employing scanned body data in the creation of patterns [13]. This exclusion restricts application to close-fit clothing with highly elastic material, as it both decreases comfort and affects garment aesthetics.

To optimize 3D models for flattening and get the best in class results in the form of garment patterns, one has to consider: the garment's silhouette — be it X, O, or A; its position or layering—for example, undergarments or outerwear; and the wanted fit or fitting—close, regular, or loose. For example, a garment with an O-silhouette will need more space at the middle than an X-silhouette garment. In a

similar fashion, the garment layers determine the required ease, where undergarments require minimal easing while outer garments, illustrated by coats, need much larger adjustments (Table I). These can all be achieved through traditional pattern making methods [14]. The key body shapes used to flatten the pattern were of slim-fit X-silhouettes, used as primary, secondary, and tertiary garment layers, without taking into account structural modules. It is also very important to ensure that the displace values along the vertical axis are exactly zero, because even the slightest variation will tend to shift important measurements such as the chest, waist, and hips, resulting in poor fitting for garments. Therefore, vertical axis displace values should be set to zero. The actual body measurements and the respective ideal making measurements of the original forms used are summarized in Table II [15].

TABLE I OFFSET CLASSIFICATIONS FOR DIFFERENT GARMENT SILHOUETTES (X, O, AND A) TO DETERMINE INITIAL SPACING IN PATTERN DEVELOPMENT

			
Silhouette	A-silhouette	X-silhouette	O-silhouette
Offset knee	High	Medium	Low
Offset hip	High	Medium	Medium
Offset waist	Medium	Low	High
Offset bust	Low	Low	Medium

D. Overview of Digital Solutions in Garment Design

Digital garment design tools have contributed to improved capabilities and features, which have optimized the full process from conceptual designs to manufacturing. This means that such software has enhanced the efficiency, accuracy, and even creativity in the field [16]. Main digital solutions include versatile Optitex software, which can offer functions ranging from pattern making all the way to 3D visualization, sizing, and virtual fit. Designers can build 3D models of garments, so-called virtual prototypes, in a realistic manner using Optitex—a further great solution. One more remarkable solution is the combination of 2D and 3D presented by Assyst in Style3D. Assyst does precise 2D pattern drafting, and it allows for diverse design and size support, with grading [17]. On the other hand, Style3D allows designers to simulate garments in a 3D virtual environment, allowing try-ons along with material simulation and rendering. This brings out seamless coordination between the 2D and 3D processes, therefore more efficiency and accuracy.

Gerber Technology, also amongst the few, dominates the apparel industry using pattern making tools, sizing, marker creation, and production management. With its subsidiary, the Lectra Modaris features design and pattern development, refined production planning, and automated cutting, which brings streamlined manufacturing processes. Browzwear also develops software, offering their key software, VStitcher, which designs realistic 3D models for fitting and design down to the last details. Software includes GRAFIS—one of the best 2D parametric pattern software. Some other great tools in the market are CLO for 2D and 3D garment design with realistic 3D modeling, virtual try-ons, pattern making, sizing, and textile simulation, to fit the entire cycle of production [18].

While the integrated use of these digital tools has greatly enhanced garment design and craft, specialized knowledge, and required skill in pattern making, they are also restricting access for those without educational qualifications. Recognizing it all, they undoubtedly become essential in shaping accuracy, innovation, and productivity in the fashion sector.

E. Gap Analysis in Digital Garment Technology and Pattern Development

Digital garment technology has upfront accelerated the process of generating patterns from 3D scans and contributed to traditional ways of depending on created physical prototypes for fit analysIs. Traditional methods of pattern construction often require high expertise, while AI and automation technologies have greatly facilitated creating a particular type of pattern. AI-enabled tools can help users, even without professional experience, generate basic patterns in a short period of time. Although AI cannot fully replicate the nuanced expertise of experienced pattern makers—particularly for accommodating diverse body shapes and textile materials—it allows non-professionals to bypass complex processes and achieve satisfactory results [19].

Accessibility is another key consideration when evaluating current technologies. Traditional methods often depend on costly, specialized equipment, and most modern digital solutions require expensive licenses and high-performance PCs. On the other hand, an open-source patterning software, for example, Blender, makes broad use more accessible because it is free and compatible with commonly used hardware. Especially, Blender allows for the automation of developing custom patterns based on 3D scans [20]. The pattern design has complex tasks that can be automated with the use of algorithms and artificial intelligence, making time expenditure less and efficiency increased. Thus, users can design accurate and individual patterns, without time-consuming manual work. Hence, the key elements of this workflow are measurement programs, 2D pattern systems, and 3D simulation and visualization software. This integration makes it possible for even users with a limited level of expertise and experience to quickly and easily develop customized basic patterns for further design development [21]. The flattening operation is viewed via UV Editor within Blender; this editor contains a heatmap shader that graphically and through color demonstrates UV stretching from the blue to yellow color, marking the amount of distortion happened on the UV faces. Although this shader helps to detect stretching, still in some particularly difficult cases, that may be hard to recognize if working with 3D model editing. This is handy because the UV Editor also allows the selection of overlapping UV faces, which are visible in both the Editor and the 3D View. Blender also has the ability to preview surface normals on the model by actually coloring individual model faces appropriately, but this feature is not available for

UV normals. The texture visualization process within Blender requires users to create a new material, assign the texture to this material, and then add the material to the model. This is a key set of a procedure for the exact display of textures in the 3D viewport and final rendering. Procedures such as these demonstrate the flexibility and practicality of Blender in pattern creation and visualization, offering a comprehensive and user-friendly instrument for garment design workflows.

TABLE II OFFSET VALUES FOR BUST, WAIST, AND HIP CIRCUMFERENCES ACROSS DIFFERENT GARMENT LAYERS [15]

Measurement in cm	Body height	Chest circumference	Waist circumference	Hip circumference	Shoulder width
Size 38	168	88	72	97	12.7
Offset	0	4	4	4	0
1st layer	168	92	76	101	12.7
Offset_2	0	5	5	5	1.5
2nd layer	168	97	81	106	14.2
Offset_3	0	5	5	5	0.5
3rd layer	168	102	86	111	14.7

F. Artificial Intelligence and Resources for Advanced Garment Design

Artificial Intelligence has transformed the face of fashion design—from forecasting upcoming trends and providing stylistic solutions to making virtual models, devised by AI, the guiding light that carries forward the vision of a novel world of fashion. Artificial intelligence, through machine learning, is transforming the fashion industry by analyzing huge datasets comprising historical fashion trends, consumer behavior, and market patterns to predict future preferences and upcoming styles. Capability for trend prediction provides designers with important information, making it easier to adapt their products to consumer needs. AI-driven virtual prototyping tools further streamline the design process by enabling fast visualization and iteration of garments digitally even before physical production begins [22]. Additionally, styling platforms driven by artificial intelligence tools can be able to make suggestions considering personal preferences and body types.

The sustainability of the fashion industry has increasingly been emphasized through the integration of artificial intelligence. These AI-powered systems improve sustainability in the fashion industry by refining supply chain processes and improving demand forecasting while reducing material waste. Such technologies enable the use of ecologically friendly techniques and materials for the good of the environment. Artificial intelligence thus not only enhances operations and fosters creativity but also leads to sustainability by advancing the processes of design and production. The geometric models of clothes come from a collection of the UC Berkeley Computer Graphics Research Group known as the Berkeley Garment Library. Exactly such garment models fitted for simulating cloth behavior are very important training data for our system. Either. We further enriched this dataset with the SewFactory one, comprising approximately one million annotated images and sewing patterns, and with the “Dataset of 3D Garments with Sewing Patterns” that comprises 23,500 three-dimensional models split into 12 various garment categories. These were a significant contribution to the training and model validation. A garment creation common add-on used in this study is the Garment Tool 2.0 developed inside Blender to speed up the process in garment creation. This led to enabling the realism in the simulation of textiles, designs, and sewing technologies available within Blender by attracting the physics engines for fine-quality cloth dynamics. The various UV-mapping tools and texture visualization possibilities for clothing design within Blender supported the creation and evaluation in a virtual environment. In combination, such tools and resources allowed for an effective and precise garment model, moving further to enable future development in fashion design.

III. METHODOLOGY

A. Conceptual Framework and Objectives of the Custom Blender Tool

One of the core parts of this technological setup of the project is a specially designed Blender tool through which theoretical methodologies are linked to practical applications in the construction of garments. The tool's functionality basically is observed in Blender's support of 3D-scanned avatars and the prospect of making automated, smooth processes in the transformation of raw scan data into useful forms. The tool interface provides user-friendly options for the selection of various garment types, the setting of fit preferences between loose-regular-tight, and the customization of additional parameters. These settings are seamlessly transferred to the level of the 3D body model, so that users can appropriately customize garment designs without the need for intensive technical background.

The process starts from choosing the type of garment, for which the tool insults a gratuitous pre-stored database of template garments. Each template includes traditional structural components: seamlines, cutlines, and construction specifications of each individual garment type chosen. For each specific garment type selected, the fit preferences are set. Loose, regular, or tight adjustments are controlled through artificial intelligence algorithms which make proper changes to the templates [23]. For instance, the closer the fit, the more number of darts needed for the kind of shaping in the garment silhouette; a looser fit would lessen the number of darts and

gathers in order to create a causal silhouette. The script can also aid in the planning of specific pattern details suitable for the 3D body model. Seamlines are drafted from garment templates and are manipulated based on body measurements and fit for equilibrium between structural integrity and aesthetic properties. Darts are placed in strategic areas where shaping is required, such as bust darts in any garment drafted with the need for shaping around the chest. Gathers are added in areas where added volume is required, such as the waist in some kinds of skirts. Tailored features are added according to the end user's desire and placed according to traditional methods of garment construction. The incorporation of body surface analysis in this tool makes it easier for one to create an Ideal Construction Body (ICB), or a three-dimensional body model that is meant to achieve a better fit. The tool makes it easier to perform the complex procedures of 3D scanning and garment fitting using advanced algorithms in Blender's powerful modeling environment. This helps improve the efficiency and accuracy of garment design, through digital means, while still enabling a user to explore creative opportunities without requiring further knowledge in constructing real garments.

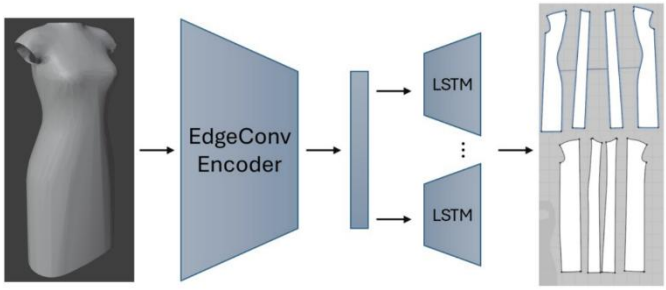


Fig. 2. AI-driven garment generation network: from 3D model and clothing type parameters to flattened sewing patterns using EdgeConv-encoder and LSTM-decoders.

B. AI-Driven Framework for Automated Garment Pattern Generation

One of the primary objectives of the project is the inclusion of a custom Blender script combined with a pre-trained neural network dedicated to the generation of garment patterns. The combination allows for the direct generation and alteration of garment patterns directly in the Blender environment, hence streamlining the design-to-pattern process (Fig. 2). The neural network is based on NeuralTailor [13], which is adapted to transform the ideal 3D body surface model into precise, ready-to-sew sewing patterns. The AI-based garment generation network operates as follows: It starts with input processing, wherein the network is given 3D body scans and offset values are applied to build an Ideal Construction Body (ICB) model. ICB is a precise model of the subject's body topography, which forms the basis for creating personalized apparel. The backbone of the network is an EdgeConv-based encoder, which leverages convolutional neural networks designed to learn geometric structures well. The encoder accepts the 3D model and obtains the main features that are essential for fitting garments. A skip connection is established from the input to the deepest EdgeConv layer to retain fine-grained information by conveying information through deeper layers of the network.

After feature extraction, a Long Short-Term Memory (LSTM) network transforms the garment's latent code. The LSTM is well suited to handle sequential data and stores the temporal dependencies among garment panels in the latent vectors. The process ensures that the final garment design is structurally sound and coherent when formed. The LSTM's latent vectors are individual garment panels. The final step involves a Panel Decoder to restore the complex form and stitching details for every garment panel. The decoder translates the abstract latent codes into practical sewing patterns, which determine panel sizes, layouts, and stitching recommendations. The task is highly significant to ensure the created patterns are correct and feasible to sew. This computer-based system allows for smooth transition from 3D body scanning to precise sewing patterns, a revolution in automated, made-to-measure clothing manufacturing [24].

C. Features and Functionality

The tool developed converts the raw 3D scans to ICBs and creates 2D garment patterns with least distortion automatically. The first step of the process is to convert 3D scans into 3D ICBs, which are specific to the measurement of an individual and may vary in shape and proportion. Advanced AI algorithms ensure these ICBs are correct representations of anatomical details, while offset values referenced from Table II have been applied at critical points such as the bust, waist, and hips to account for different layers of garments. This step ensures that the constructions are both precise and anatomically correct for a wide range of garment designs. The unwrapping automatically produces a 2D pattern of the 3D mesh once the ICB is produced. It minimizes distortions to yield flat, ready-to-use patterns that can be used as the basis of garment assembly. The complex workflow is simple in this tool through the user-friendly interface, enabling users to easily customize their garment styles and sizes. Modifications to fit preferences, such as close-fitting, regular, or loose-fitting designs, become so easy for users according to their needs and liking. Because it is highly customizable, the tool actually enables users to create garments that reflect their vision. The tool works out the most advanced technologies, streamlining the pattern-making process while giving users increased creative freedom. From 3D scan processing to very detailed 2D pattern creation and enabling customizable style variations, this innovative tool offers a perfect combination of precision, efficiency, and accessibility in garment construction.

D. Workflow for Enhanced Garment Fitting Using 3D Scanning and AI

It describes in detail a systematic workflow that attempts to enhance fitting garments with developments in 3D scanning and machine learning algorithms. The heart of the matter in this method lies in the question of how design in clothes may be adjusted regarding singular contours of the human body to make comfort complement elegance. It involves steps ranging from person-scanning to virtually assessing comprehensive fit. It begins with a very precise 3D scan of the person using a 3D scanner. This initial scan gives the basic data necessary to fit the garment to the person's unique body shape. After the scan, the user selects the type of garment they would like to create,

and the program allows variable parameters such as the layer of fit preferred, from loose through regular to tight. This enables the garment to be tailored specifically to the wearer's preference and his or her anatomy. After configuration, the next step involves analysis of the 3D body scan to look at important areas of the surface and body measurements. These will be useful for finding the peculiar contours and measures necessary in garment design. Further from this extracted data during the workflow, a model of ICB or Ideal Construction Body will be generated. Serving as a virtual mannequin, this offers a framework adapted to ensure the intended fit's achievement.

This includes the gathering of sewing information, such as seam placement and allowances, among other necessary data for precision in pattern construction. Then the pre-trained neural network, from that data, predicts an estimate of count, form, and configuration of garment pattern pieces. By doing so, this cutting-edge machine-learning technique will thus yield an accurate construction of distortion-free patterns best fitted on the 3D model of the individual human body. The last step is 3D simulation, which checks the fitting of the garment. The integration of newly generated patterns with the scanned 3D body model enables a virtual try-on experience through simulation. This allows the detection and correction of any fit issues, ensuring that the final garment perfectly matches the specified design and comfort requirements. This smooth workflow seamlessly integrates state-of-the-art technologies to enhance efficiency and accuracy in garment creation while providing a highly personalized outcome.

IV. Experimentation and Discussion

In order to validate the performance of the developed tool, a structured approach was followed to establish its efficacy in converting 3D scans into Ideal Construction Bodies and subsequently creating custom garment patterns. The objective of this test was primarily to determine how accurate the AI-generated patterns were compared to manually created patterns. A specific program was developed for performing this test based on three key parameters: perimeter, area, and HU moments. Higher-order moments of the pixel intensity distribution yielded HU moments, statistical measures of shape and structure in digital images. The resultant metrics are capable of presenting quantitative insights into two important features of a pattern: symmetry and general congruence. The comparison of AI-generated patterns with the ground truth during the evaluation process was done by tools implemented in Python. The formula used for finding the discrepancy between the two patterns involves the differences in HU moments, as expressed in the given equation. This summarizes the differences for all seven HU moments of each pattern piece, normalizing them with respect to the ground truth. The tool will automatically assign an identifier to every pattern piece and assess it independently from any transformation, like rotation or scaling, against the others.

The process is illustrated in Fig. 3, which compares the calculated HU moments of each pattern piece. These differences are summed and normalized to provide a quantitative measure of how well the AI-generated patterns match their ground truth counterparts. Area, perimeter, and HU moments for each section of the pattern are further analyzed in the program, with detailed results presented in Table III. This indeed constitutes a very detailed analysis that brings out the consistency and precision of the generated patterns. The test proves that this tool can generate patterns that are quite accurate with a few errors when compared to manually unwrapped patterns. Moments of HU mean the perfect fitting when the value of the distance is 0.0 and a larger deviation would indicate regions which might need refinement. Results proved that wider garment layers are usually capable of giving more accuracy, hence pointing out the strength of this tool to adapt with different design needs, thus generally increasing the efficiency in garment productions.

TABLE III Evaluation Metrics for basic Patterns Across First, Second, and Third Layers

	Perimeter in %	Accuracy Area in %	Distance HU moments	Perimeter in %	Accuracy Area in %	Distance HU moments	Perimeter in %	Accuracy Area in %	Distance HU moments
1	98.9	79.1	0.6	97.2	74.5	1.03	95.1	77.7	3.21
2	99.6	119.4	67.61	96.1	74.3	1.05	99.2	88.9	1.25
3	100	113	24.27	98.8	86.5	2.66	93.4	79.6	0.4
4	98.7	86	0.57	96.9	91.2	1.16	99.9	98	0.65
5	94.5	75	0.59	91.8	80.2	0.66	97.8	97.9	0.46
6	93.7	74.8	0.59	94.4	68.8	0.55	99.6	98.3	0.68
7	100.9	105.8	1.17	99.3	84	3.15	94.5	76.7	0.8
8	101.4	106.1	1.48	97.4	75.8	5.16	98.5	95.6	0.82
Total	97.9	83.8	88	96.4	77.7	1.9	97.3	89.1	1.03

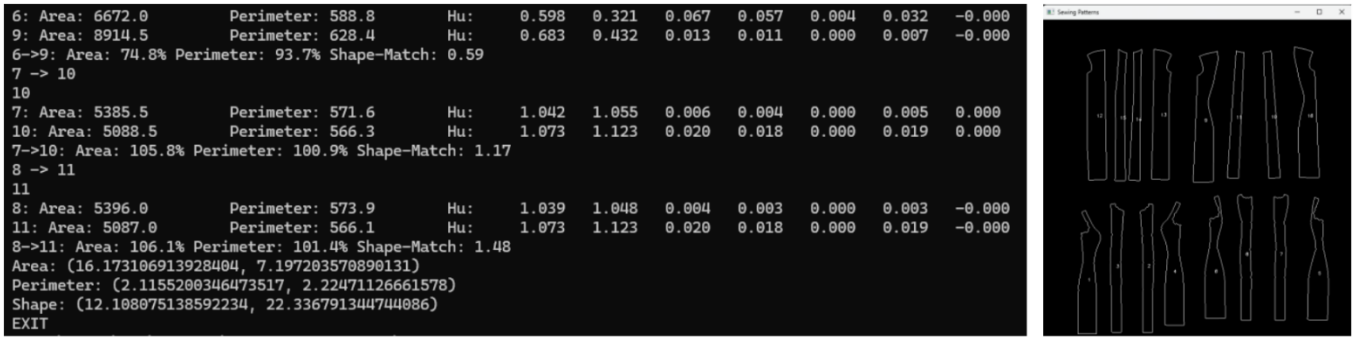


Fig. 3. Comparison of HU moment differences between AI-generated patterns and ground truth shapes.

V. Conclusion

This paper introduces meaningful changes in garment manufacturing and personalization with AI-generated patterns from 3D scans. This is quite a critical approach, as it reduces by a great deal the time spent making simple garments to fit the person's exact measure. By simplifying several steps of the process in fashion design, the tool makes it easy for novice and professional 3D artists to generate well-fitting garments for multiple layers. While AI speeds up the process of creating basic patterns for new designs, their effectiveness depends on user expertise and further physical prototyping to validate the generated patterns. This new tool bridges the gap between design and production by automatically converting 3D body scans into sewing patterns. It revolutionizes the workflow from concept to creation, making personalized basic patterns more accessible and enabling scalable bespoke garment production. However, this tool has its current implementation limitations. The system is limited to generating patterns of garment types included in the pre-trained dataset, which restricts design versatility. Besides, the manual steps of importing 3D scans may reduce the speed of the workflow and make it prone to inaccuracies. These challenges mark points that need improvement in order to further develop the capabilities and usability of the tool. Inclusiveness and accessibility make the democratization of fashion design through the integration of AI in the construction of patterns. Further developments involve increasing the functionality of the tool in constructing patterns on pants, sleeve designs, and the placement of seams according to particular needs; hence, its reach will go outside basic garments.

Another significant digital apparel breakthrough was the AI-powered tool that converts 3D scans into bespoke garment patterns. The conducted experiments demonstrate that the tool can generate patterns with precision and accuracy comparable to manually created designs. The quantitative evaluation of metrics such as perimeter, area, and HU moments substantiates the tool's reliability and efficiency. These findings provide a solid foundation for optimizing and extending the tool's capabilities, ensuring greater adaptability and performance in real-world applications. The results provided here are part of the ongoing digitalization of fashion and will support further developments within digital garments. The next steps in this research include the verification of the patterns through physical productions, new possibilities of design like multi-layer garments, and more intricate pattern models. This tool currently utilizes basic patterns, but it's a promising step toward a future of personalized and scalable garment design.

Future work will focus on enhancing the AI model's adaptability, improving pattern customization for complex garment structures, and integrating real-time feedback mechanisms for better accuracy and user experience.

Funding

2022 Heilongjiang Province Philosophy and Social Sciences Research Plan Project "Digital Archive Construction and Research of Traditional Costumes of Ethnic Minorities with Small Population in Northeast China" Project Number: 22MZB163; Basic Research Funds for Higher Education Institutions in Heilongjiang Province in 2024 Research Project "Establishment of Digital Archives for Traditional Costumes of Ethnic Minorities with Small Population in Northeast China and Research on National Memory Inheritance" Project Number: 145409506;

References

[1] Danner, M., Brake, E., Kosel, G., Kyosev, Y., Rose, K., Rätsch, M. and Cebulla, H., 2024. AI-assisted pattern generator for garment design. Communications in development and assembling of textile products, 5(2), pp.195-206.

[2] Cao, C., 2022. Research and application of 3D clothing design based on deep learning. Advances in Multimedia, 2022(1), p.2270005.

[3] Jing-Jing, F. and Ding, Y., 2008. Expert-based customized pattern-making automation: Part I. Basic patterns. International Journal of Clothing Science and Technology, 20(1), p.26.

[4] Huang, H.Q., Mok, P.Y., Kwok, Y.L. and Au, J.S. (2012) 'Block pattern generation: From parameterizing human bodies to fit feature-aligned and flattenable 3D garments', Computers in Industry, 63(7), pp. 680–691.

[5] Choi, Y.L., Nam, Y., Choi, K.M. and Cui, M.H. (2007) 'A method for garment pattern generation by flattening 3D body scan data', in Digital Human Modeling: First International Conference on Digital Human Modeling, ICDHM 2007, Held as Part of HCI International 2007, Beijing, China, July 22-27, 2007. Proceedings 1. Berlin: Springer, pp. 803–812.

[6] Han, H.S., Kim, J.Y., Kim, S.M., Lim, H.S. and Park, C.K., 2014. The development of an automatic pattern-making system for made-to-measure clothing. Fibers and Polymers, 15, pp.422-425.

[7] Kosel, G. and Rose, K. (2020) 'Development and Usage of 3D-Modeled Body Shapes for 3D-Pattern Making', in Proceedings of 3DBODY.TECH 2020 - 11th International Conference and Exhibition on 3D Body Scanning and Processing Technologies, Online/Virtual, 17-18 November 2020,

[8] Jin, P., Fan, J., Zheng, R., Chen, Q., Liu, L., Jiang, R. and Zhang, H., 2023. Design and research of automatic garment-pattern-generation system based on parameterized design. Sustainability, 15(2), p.1268.

[9] Narain, R., Samii, A. and O'Brien, J.F. (2012) 'Adaptive Anisotropic Remeshing for Cloth Simulation', ACM Transactions on Graphics, 31(6), p. 152.

[10] Wang, H.M., Ramamoorthi, R. and O'Brien, J.F. (2011) 'Data-Driven Elastic Models for Cloth: Modeling and Measurement', ACM Transactions on Graphics, 30(4), p. 71.

[11] Liu, L.J., Xu, X.Y., Lin, Z.J., Liang, J.B. and Yan, S.C. (2023) 'Towards Garment Sewing Pattern Reconstruction from a Single Image', ACM Transactions on Graphics, 42, p. 200.

[12] Korosteleva, M. and Lee, S.-H. (2021) Dataset of 3D Garments with Sewing Patterns (1.0). Zenodo. DOI: https://doi.org/10.5281/zenodo.5267549.

[13] Korosteleva, M. and Lee, S.-H. (2022) 'NeuralTailor: Reconstructing sewing pattern structures from 3D point clouds of garments', ACM Transactions on Graphics, 41(4), p. 158.

[14] Liu, K., Zeng, X., Bruniaux, P., Tao, X., Yao, X., Li, V. and Wang, J., 2018. 3D interactive garment pattern-making technology. Computer-Aided Design, 104, pp.113-124.

[15] Ebner Media Group GmbH & Co. KG. (2024) Müller & Sohn Pattern System. Available at: https://www.muellerundsohn.com/ (Accessed: 30 March 2024).

[16] Hohenstein Laboratories GmbH & Co. KG. (2024) Schnitt-Service. Available at: https://www.hohenstein.de/de/kompetenz/passform/schnitt-service (Accessed: 30 March 2024).

[17] Zhang, F.F. (2015) Designing in 3D and Flattening to 2D Patterns. PhD Thesis. North Carolina State University.

[18] Optitex (2024) 3D Product Creation Suite. Available at: https://optitex.com/solutions/odev/3d-production-suite/ (Accessed: 30 March 2024).

[19] Assyst GmbH. (2024) CAD. Available at: https://www.assyst.de/de/produkte/cad/index.html (Accessed: 30 March 2024).

[20] LECTRA (2024) Available at: https://www.lectra.com/en (Accessed: 30 March 2024).

[21] Browzwear Solutions Pte Ltd. (2024) Available at: https://browzwear.com/ (Accessed: 30 March 2024).

[22] CLO Virtual Fashion (2024) Available at: https://www.clo3d.com/en/ (Accessed: 30 March 2024).

[23] Blender Foundation (2024) Available at: https://www.blender.org/ (Accessed: 30 March 2024).

[24] Abdel Kader, E.A.S., Mohamed, R.H. and Ali, R. (2022) 'The Role of Artificial Intelligence (AI) Applications in Fashion Design and Forecasting in the Garment Industry: An Analytical Study', International Design Journal, 12, pp. 203–214.

Enhancing Recurrent Neural Network Efficacy in Online Sales Predictions with Exploratory Data Analysis

Erni Widiastuti[1], Jani Kusanti[2], Herwin Sulistyowati[3]
Faculty of Economic, Universitas Surakarta, Indonesia[1]
Faculty of Electrical Information Engineering, Universitas Surakarta, Indonesia[2]
Faculty of Law, Universitas Surakarta, Indonesia[3]

***Abstract*—Online sales forecasting has become an essential aspect of effective business planning in the digital era. The widespread adoption of digital transformation has enabled companies to collect substantial datasets related to consumer behavior, market trends, and sales drivers. This study attempts to uncover patterns and predict sales growth by utilizing product images and their associated filenames as input. To achieve this, we use EDA combined with LSTM and Gated Recurrent Unit (GRU), which excel in processing sequential data. However, the performance of these networks is significantly affected by the quality of data and the preprocessing methods applied. This study highlights the importance of Exploratory Data Analysis (EDA) and Ensemble Methods in enhancing the efficacy of RNNs for online sales forecasting. EDA plays a crucial role in identifying significant patterns such as trends, seasonality, and autocorrelation while addressing data irregularities such as missing values and outliers. These findings show that integrating EDA substantially improves the performance metrics of RNN, as indicated by the reduction in loss and mean absolute error (MAE) values across training epochs (e.g. loss: 0.0720, MAE: 0.1918 at epoch 15). These results indicate that EDA improves the accuracy, stability, and efficiency of the model, allowing RNN to provide more reliable sales predictions while minimizing the risk of overfitting.**

***Keywords*—Exploratory data analysis; recurrent neural networks; online sales prediction; sequential data; trend patterns**

I. Introduction

The digital business era has fundamentally reshaped sales forecasting and management by harnessing the power of big data and advanced analytics [1]. Technology-driven digital transformation has brought significant changes to sales processes and environments, altering their functions and strategies [2][3][4]. Deep learning models, particularly in e-commerce systems, have emerged as essential tools for predicting and enhancing sales outcomes. This trend is emblematic of a broader shift toward digital business, with the e-commerce sector serving as a prime example of this ongoing evolution [4].

Extensive research has compared traditional sales forecasting methods with machine learning techniques, yielding diverse findings. Pustokhina [5] noted that machine learning often surpasses traditional methods in accuracy, yet certain traditional techniques, such as the Holt-Winters method, remain effective under specific conditions. Zhang [6] proposed an innovative model that integrates online reviews and search engine data, significantly enhancing forecasting precision. Furthermore, Bajaj [7] and Cheriyan [8] emphasized the utility of machine learning algorithms in sales forecasting, with Bajaj exploring models such as Linear Regression, K-Neighbors Regressor, XGBoost Regressor, and Random Forest Regressor, while Cheriyan recognized Gradient Boosting as the most effective method for forecasting sales trends.

Deep learning, a focused domain within machine learning, has achieved significant progress in recent years, [9] showcasing its value in solving complex classification challenges. Its capability to derive robust statistical features from data sets it apart as a powerful tool. Research [10] highlighted the critical role of raw data in optimizing machine learning performance. Studies [11][12][13] the evidence presented illustrates that state-of-the-art deep learning methodologies, such as Long-Short Term Memory (LSTM) networks and Convolutional Neural Networks (CNN), substantially exceed the performance of conventional machine learning methodologies in the domain of retail sales forecasting.. Furthermore, evidence from [14][15][16][17][18] reinforces this, demonstrating that deep learning models excel in generating accurate customer predictions for marketing intelligence applications. Together, these findings highlight deep learning's substantial impact on enhancing sales prediction accuracy and operational efficiency.

This study aims to identify sales trends and analyze the factors that influence sales growth within the digital business environment. To fill the existing research gap, we propose an innovative approach that integrates Exploratory Data Analysis (EDA) for feature preprocessing with Ensemble Methods applied to Recurrent Neural Networks (RNNs) to predict product similarity based on e-commerce image data. This research underscores the significant role of sequential learning techniques and vector embedding in enhancing the accuracy of product similarity predictions. Furthermore, EDA is highlighted as a crucial tool for deriving insights and ensuring data quality in the analysis of time series and sequential data.

II. Previous Research Study

Jelonek [19] asserts that Big Data analytics is a vital asset for business management, providing diverse benefits across numerous activities. Ansari [20] highlights how the integration

of cloud computing with Big Data analytics ensures cost-effective and scalable solutions for storing and analyzing vast enterprise datasets. Nevertheless, the literature lacks a focused discussion on leveraging deep learning for sales prediction. Alsghaier [21] explains that Big Data analytics equips organizations with actionable insights, enhancing business performance and fostering a competitive edge. Ayuningtyas [22] emphasizes that descriptive, predictive, and prescriptive analytics within Big Data are indispensable for strategic decision-making across industries. Singh [23] investigates the vast possibilities of deep learning within the automotive industry, including its use in self-driving cars, safety systems, virtual sensors, and cutting-edge product development.

Recent developments in research have concentrated on employing deep learning techniques to forecast consumer purchasing behavior. Geetha [15] introduces a deep neural network model that leverages multitask learning to predict consumer preferences by analyzing underlying factors and sentiment. Xia [24] develops a multi-task LSTM model to capture the complexities of the consumer buying decision-making process and estimate purchase probabilities. Nisha [25] conducts a comparative analysis of various neural network architectures, including MLP, LSTM, and TCN, for predicting future purchase behavior in e-commerce. These models consistently outperform traditional machine learning methods across multiple dimensions of consumer behavior prediction.

Liu [26] underscores the transformative impact of computer vision technology on e-commerce platforms, particularly in the realm of sales forecasting. Deep learning models enable the automatic extraction of crucial features from product images, providing valuable insights for predicting sales outcomes. Zhao [27] illustrates how Convolutional Neural Networks (CNNs) can efficiently extract features from structured data, improving forecasting accuracy without requiring manual feature engineering. Qi [28] introduces DSF, a deep neural framework designed to tackle the complexities of promotional activities and product competition in sales forecasting, surpassing traditional baseline models and other deep learning techniques in performance. Yang [29] highlights that the application of computer vision in e-commerce extends far beyond sales forecasting, driving improvements in operational efficiency and customer satisfaction across various areas of online shopping.

Kassem [30] proposed models that classify reviews as either positive or negative and compare them with the ratings provided by users to identify any inconsistencies. These strategies aim to enhance the accuracy and reliability of product information for consumers. Wang [31] developed CLUE, a fraud detection system that leverages recurrent neural networks to analyze user click behavior and detect suspicious transactions.

III. Proposed Research Methodology

The methodology for this study is illustrated in Fig. 1 which depicts the process of creating an Exploratory Data Analysis (EDA) algorithm model utilizing LSTM and GRU to forecast online sales.

In Fig. 1, Flowchart combining Exploratory Data Analysis (EDA) with LSTM and GRU frameworks.

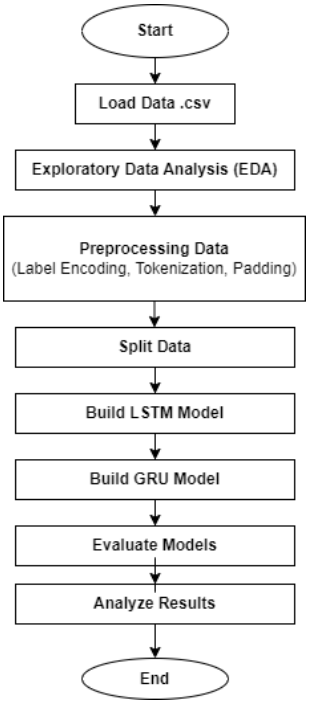


Fig. 1. Flowchart combining exploratory data analysis (EDA) with LSTM and GRU frameworks.

A. Load Data

The initial step involves loading the data from a CSV file. This file, generated during data collection, is stored in a tabular format containing five columns: posting_id, image, image_phash, title, and label_group.

B. Exploratory Data Analysis (EDA)

EDA functions are useful for identifying anomalies, trends, or outliers in sequential data that might obstruct model convergence. It primarily serves as a process for visualizing and understanding data. Below are some commonly used formulas in EDA, for the equation as in (1-6):

- Mean

  The mean reflects the central tendency of a dataset, calculated by summing all individual values and dividing the total by the number of data points.

$$Mean(\mu) = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (1)$$

Where:

x_i = value to- i

$n = the\ total\ amount\ of\ data$

- Variance

Variance quantifies the extent to which data values deviate from the mean.

$$Variance(\sigma^2) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \quad (2)$$

Where:

μ = mean

x_i = value to- i

- Standard Deviation

The standard deviation constitutes the principal square root of the variance.

$$Standar\ Deviation(\sigma) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \quad (3)$$

- Covariance

Covariance measures the degree of interdependence between two variables in a dataset.

$$Covariance(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \quad (4)$$

Where:

x_i = value to- i from X

y_i = value to- i from Y

$\overline{x}, \overline{y}$ = mean from X and Y

- Correlation

Pearson correlation quantifies the degree of linear association between two variables.

$$Correlation(\tau) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i-1}^{n}(y_i - \overline{y})^2}} \quad (5)$$

Correlations can vary between −1 and 1, with 1 indicating a perfect positive correlation, 0 denoting no correlation, and −1 representing a perfect negative correlation.

- Autocorrelation assesses the relationship between values at a specific moment and values from a prior moment (lag) for the purpose of correlation.

$$Autocorrelation(k) = \frac{\sum_{i=k+1}^{n}(x_i - \overline{x})(x_{i-k} - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad (6)$$

k is the lag, which is how far back in time we look

C. Recurrent Neural Networks (RNN)

Forward Pass

In a standard RNN, the output of a neuron at time t (h_t) is influenced by the current input (x_t) and the hidden state from the previous time step (h_t−1), for the equation as in (7-8).

- State (Hidden State)

$$h_t = \emptyset(W_{xh}.x_t + W_{hh}.h_{t-1} + b_h) \quad (7)$$

Where:

h_t = hidden state at the time t

x_t = input at time t

W_{xh} = weight between input and hidden state

W_{hh} = weight between previous hidden state and the current hidden state

b_h = bias

$\emptyset$ = activation function (usually tanh or ReLU)

- Output

$$y_t = \emptyset(W_{hy}.h_t + b_y) \quad (8)$$

Where:

y_t = output at the time t

W_{hy} = weight between output and hidden state

b_h = bias output

$\emptyset$ = activation function (softmax or sigmoid)

Backpropagation Through Time

The training process employs Backpropagation Through Time (BPTT) to compute the gradient and adjust the weights. The gradient at time step *t* is determined by taking into account all preceding time steps.

The error gradient L for parameter W, for the equation as in (9):

$$\frac{\partial L}{\partial W} = \sum_{t-1}^{T}\frac{\partial L_t}{\partial W} \quad (9)$$

T is the total number of times in the sequence

LSTM is designed to address the vanishing gradient issue commonly encountered in standard RNNs. It employs three primary gates: the input gate, the output gate, and the forget gate, for the equation as in (10-14).

- Forget Gate

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \quad (10)$$

f_t = the forget gate that determines what information will be forgotten

- Input Gate

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \quad (11)$$

i_t = the input gate that determines what new information will be stored

- Update Cell State

$$C_t = f_t.C_{t-1}.i_t \quad (12)$$

C_t = the cell state at time t

- Output Gate

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \quad (13)$$

o_t = the output gate that determines the output of the LSTM at time t

- Hidden State

$$h_t = o_t.\tanh(C_t) \quad (14)$$

h_t = the hidden state or output of the LSTM

IV. Experimental Result and Analysis

During the training of our EDA using both LSTM and GRU models, we began by performing an exploratory analysis on the label groups found in Train.csv, examining their data distributions. The results before using EDA are presented in Fig. 2. While the results after using EDA are presented in Fig. 3.

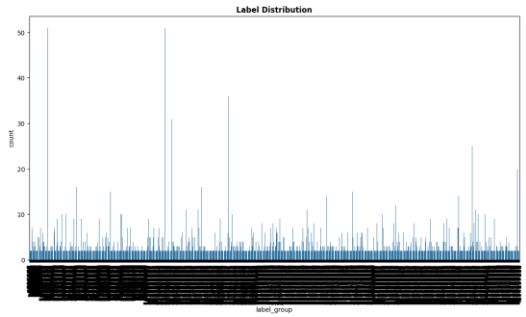


Fig. 2. The distribution of label_group in Train.csv without performing EDA.

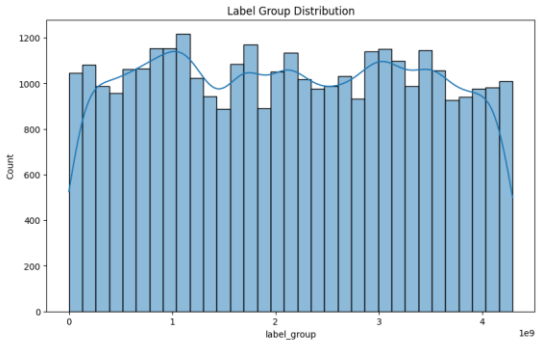


Fig. 3. The distribution of label_group in Train.csv with performing EDA.

From Fig. 2, the figure presents a bar chart. The horizontal axis (x-axis) denotes the label_groups, while the vertical axis (y-axis) displays the count or frequency of occurrences for each label_group. The distribution of label_groups is notably skewed and unbalanced. Some label groups exhibit very high frequencies, surpassing 50, whereas the majority have significantly lower frequencies, frequently below 10, with many falling under 5. The presence of several tall bars indicates that certain label groups are particularly dominant within the dataset. This suggests a bias in the data towards specific labels. Such an imbalanced distribution of labels has important implications for machine learning modeling. If a model were trained directly on this data, it would likely favor the dominant label groups and perform poorly on those that occur less frequently.

From Fig. 3, this figure illustrates the distribution of label_groups in the Train.csv dataset after conducting Exploratory Data Analysis (EDA). Unlike Fig. 2, which shows a significant imbalance, this figure presents a much more balanced distribution. The histogram displays the frequency of each label_group value range as bars, while the KDE (blue line) provides a smoother estimate of the variable's probability distribution. Compared to Fig. 2, the distribution of label_groups here is notably more uniform, with no excessively tall bars. This indicates that the occurrence frequency of each label_group value range is relatively equal. Such a balanced distribution has positive implications for machine learning modeling. The main difference between Fig. 2 and Fig. 3 is the degree of imbalance: Fig. 2 reveals extreme disparity, where some label_groups are highly prevalent while most others are rarely observed. In contrast, Fig. 3 illustrates a more equitable distribution following EDA. This suggests that EDA has resulted in a significant transformation of the data, possibly involving clustering or other modifications to the label_groups.

The test results to explore the distribution of text length in title_image in text.csv can be shown in Fig. 4.

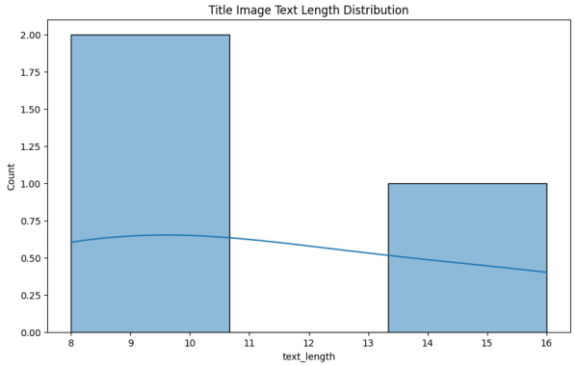


Fig. 4. Title image text length distribution.

Fig. 4 depicts the distribution of lengths for image caption texts. The X-axis represents text lengths, which range from about 8 to 16 characters or words. This shows that the lengths of the image captions fall within this range. The Y-axis indicates the frequency of occurrences for each range of text lengths.

	label_group
count	3.425000e+04
mean	2.128611e+09
std	1.234630e+09
min	2.580470e+05
25%	1.050720e+09
50%	2.120410e+09
75%	3.187910e+09
max	4.294197e+09

Fig. 5. The results of the EDA calculation.

Fig. 5 shows the EDA calculation results. Fig. 6 displays the results of the Exploratory Data Analysis (EDA) for the label_group variable. Below is a detailed explanation of the statistics:

- Count: The value 3.425000e+04 (34,250) indicates that there are 34,250 data points or observations in the label_group variable, representing the total number of entries analyzed.
- Mean: The value 2.128611e+09 (2,128,611,000) represents the average of all label_group values, providing insight into the central tendency of the data.
- Standard Deviation (std): The value of 1.234630e+09 (1,234,630,000) indicates the extent to which the data varies from the mean. A high standard deviation implies that the data points are significantly dispersed, whereas a

low standard deviation suggests that they are closely grouped around the mean. In this instance, the large standard deviation in relation to the mean indicates substantial variability within the data.

- Minimum (min): The value 2.580470e+05 (258,047) denotes the smallest value in the label_group variable.
- First Quartile (25%): The value 1.050720e+09 (1,050,720,000) represents the first quartile. This indicates that 25% of the data falls below or is equal to this value.
- Median (50% / Second Quartile Q2): The value 2.120410e+09 (2,120,410,000) is the median or second quartile. This is the midpoint of the dataset; half of the values are below this point and half are above it. Notably, the median (2.120410e+09) is very close to the mean (2.128611e+09), suggesting a relatively symmetric distribution despite a large standard deviation.
- Third Quartile (75%): The value 3.187910e+09 (3,187,910,000) indicates that 75% of the data has values less than or equal to this figure.
- Maximum (max): The value 4.294197e+09 (4,294,197,000) represents the largest value in the label_group variable.

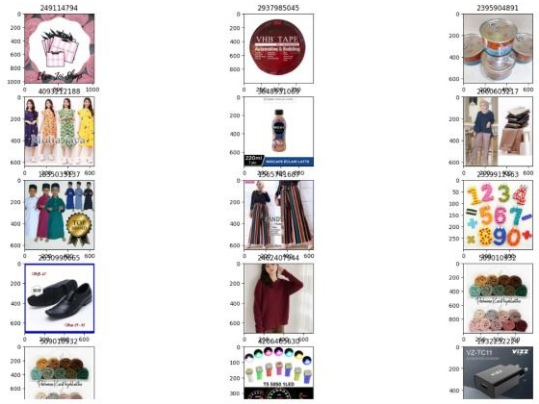


Fig. 6. Example image from label_group.

After the EDA process results, continued with the RNN process which is continued by carrying out the Split Data process for Training and Validation.

TABLE I. RESULTS OF THE SPLIT DATA PROCESS FOR TRAINING AND VALIDATION

Description	Result
Training data shape	(27400, 100) (27400,)
Validation data shape	(6850, 100) (6850,)

Table I shows the results of dividing the training data into Training Data and Validation Data. The results of the Training Data (X_train, y_train) are shown in Table I: (27400, 100) (27400,) (27400, 100). This signifies that the training dataset consists of 27,400 samples or observations, with each sample containing 100 features or time steps (27400,). This represents the shape of the labels or target variables for the training data, indicating that there are 27,400 labels, corresponding to one label for each sample in the training set.

The results of the Validation Data (X_val, y_val) are shown in Table I: TABLE I. (6850, 100) (6850,). (6850, 100). This indicates that the validation dataset consists of 6,850 samples, with each sample containing 100 features or time steps, similar to the training data. The validation data is utilized to assess the model's performance during training and to mitigate the risk of overfitting. (6850,): This represents the shape of the labels or target variables for the validation data, indicating there are 6,850 labels—one corresponding to each validation sample.

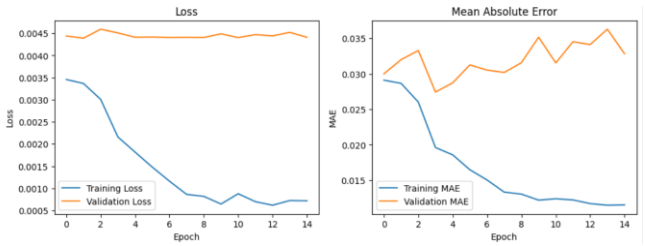


Fig. 7. Training Loss and MAE results with LSTM model without EDA.

In Fig. 7, the document exhibits two graphical representations that monitor the efficacy of the model throughout the training process: the Loss graph and the Mean Absolute Error (MAE) graph, both of which are charted in relation to the epochs.

1) Loss graph: X-axis (Epoch): Represents the number of training epochs, ranging from 0 to 14. Each epoch corresponds to a complete iteration in which the model processes the entire training dataset. Y-axis (Loss): Displays the loss value, which measures the discrepancy between the model's predictions and the actual values. A lower loss value indicates better model performance. Training Loss (Blue Line): Sharp Decline at the Start: The training loss experiences a rapid decrease from approximately 0.0035 at epoch 0 to below 0.001 around epoch 5. This indicates that the model is quickly learning and enhancing its performance on the training data in the early stages. Relatively Stable After Epoch 5: Following epoch 5, while the training loss continues to decline, the rate of decrease becomes less significant and exhibits slight fluctuations. Validation Loss (Orange Line): Generally Stable: The validation loss remains relatively stable at around 0.0045 throughout the training process. Although there are minor fluctuations, there is no clear downward trend as seen in the training loss. A notable gap exists between training loss and validation loss, with the training loss being significantly lower than the validation loss.

2) Mean absolute error (MAE) graph: X-Axis (Epoch): Similar to the loss graph, this axis represents the number of training epochs. Y-Axis (MAE): This axis indicates the MAE value, which measures the average absolute difference between the model's predictions and the actual values. A lower MAE signifies better model performance. Training MAE (Blue Line): Sharp Decline at the Start: Like the training loss, the training MAE also decreases rapidly during the initial stages of training. Relatively Stable After Epoch 5: After epoch 5, the training

MAE stabilizes and shows only minor fluctuations. Validation MAE (Orange Line): Initial Increase Followed by Fluctuations: The validation MAE experiences a slight increase at the beginning of training and then tends to fluctuate around 0.032 to 0.036, without a consistent downward trend. There is a considerable difference between the training MAE and validation MAE, with the training MAE consistently being lower than the validation MAE.

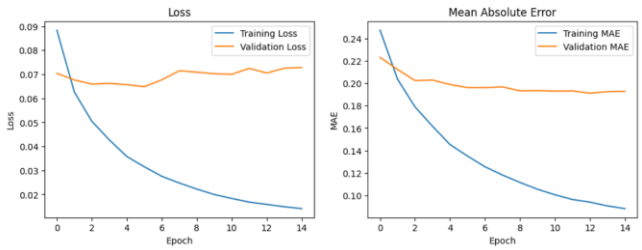


Fig. 8. Training Loss and MAE results with the LSTM model after using EDA.

Fig. 8 presents two graphs that track the model's performance during training, utilizing the Loss and Mean Absolute Error (MAE) metrics plotted against epochs. These metrics are commonly employed to assess regression models. Below are detailed descriptions of each graph.

3) Loss graph: X-axis (Epoch): Displays the number of training epochs, ranging from 0 to 14. Each epoch represents a complete iteration in which the model processes the entire training dataset. Y-axis (Loss): Represents the loss value, which measures the extent to which the model's predictions deviate from the actual values. A lower loss value indicates better model performance. Training Loss (Blue Line): Rapid Decline at the Start: The training loss decreases quickly and consistently from approximately 0.09 at epoch 0 to below 0.02 at epoch 14. This suggests that the model is effectively learning and enhancing its performance on the training data. Consistent Decrease: This steady decline indicates that the model continues to learn throughout the epochs. Validation Loss (Orange Line): Relatively Stable After Initial Decrease: The validation loss shows a slight decrease at the beginning of training, but after a few epochs, it stabilizes and fluctuates slightly around 0.07. There is a notable difference between training loss and validation loss, with training loss consistently lower than validation loss. This difference remains small and stable from the midpoint to the end of the epochs.

4) Mean absolute error (MAE) graph: X-axis (Epoch): Similar to the loss graph, this axis indicates the number of training epochs. Y-axis (MAE): Displays the MAE value, which measures the average absolute difference between the model's predictions and actual values. A lower MAE signifies better model performance. Training MAE (Blue Line): Sharp and Consistent Decline: The training MAE decreases rapidly and steadily from around 0.25 at epoch 0 to below 0.10 at epoch 14. This trend parallels the decrease in training loss, indicating an improvement in model performance on the training data. Validation MAE (Orange Line): Relatively Stable After Initial Decrease: Similar to validation loss, validation MAE experiences a slight initial decrease before remaining relatively stable, fluctuating around 0.20 from mid-epoch to the end. A significant difference exists between training MAE and validation MAE, with training MAE consistently lower. This difference is small and stable from mid-epoch to end.

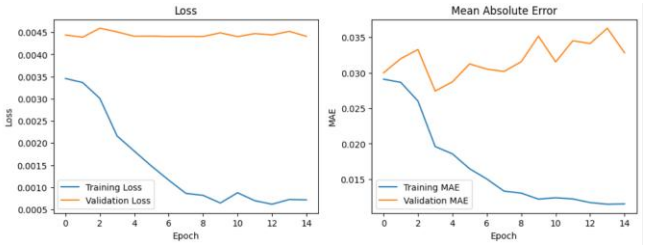


Fig. 9. Training Loss and MAE results with the GRU model after using EDA.

Fig. 9 below presents two graphs that track the model's performance during training, utilizing Loss and Mean Absolute Error (MAE) metrics plotted against epochs. These metrics are commonly employed to assess regression models. Below is a detailed description:

5) Loss graph: X-axis (Epoch): Displays the number of training epochs, ranging from 0 to 14. Each epoch represents a complete iteration during which the model processes the entire training dataset. Y-axis (Loss): Represents the loss value, which quantifies how poorly the model's predictions align with the actual values. A lower loss value indicates better model performance. Training Loss (Blue Line): Sharp Early Decline: The training loss decreases rapidly from approximately 0.0035 at epoch 0 to below 0.001 around epoch 5, indicating that the model is quickly learning and enhancing its performance on the training data in the early stages. Relatively Stable After Epoch 5: After epoch 5, while the training loss continues to decline, the rate of decrease becomes less significant and shows slight fluctuations. Validation Loss (Orange Line): Generally Stable: The validation loss remains relatively stable at around 0.0045 throughout the training process, exhibiting minor fluctuations without a clear downward trend as seen in the training loss. A notable difference exists between training loss and validation loss, with training loss consistently being much lower than validation loss.

6) Mean absolute error (MAE) graph: X-axis (Epoch): Similar to the loss graph, this axis indicates the number of training epochs. Y-axis (MAE): Displays the MAE value, which measures the average absolute difference between the model's predictions and actual values. A lower MAE signifies improved model performance. Training MAE (Blue Line): Sharp Decline at the Start: Like the training loss, the training MAE decreases rapidly during the initial phase of training. Relatively Stable After Epoch 5: Following epoch 5, the training MAE also stabilizes and exhibits only minor fluctuations. Validation MAE (Orange Line): Initial Increase Followed by Fluctuations: The validation MAE shows a slight increase at the beginning of training before fluctuating around

0.042 to 0.046, without a consistent downward trend. There is a significant difference between training MAE and validation MAE, with training MAE consistently lower than validation MAE.

7) The model evaluation results highlight two key metrics: Mean Squared Error (MSE) and Mean Absolute Error (MAE). MSE quantifies the average of the squared differences between the model's predictions and the actual values, with a lower MSE indicating greater accuracy in predicting the target value. On the other hand, MAE assesses the average of the absolute differences between the model's predictions and the actual values, where a lower MAE also signifies improved accuracy in predicting the target value. In this study, three tests were conducted, and the results are presented in Table II.

TABLE II. THE MODEL EVALUATION RESULTS

Epoch Value	Test Results	
	MSE	MAE
LSTM without EDA	2.7837	2.818866014480591
LSTM with EDA	0.07274550944566727	0.1918681412935257
GRU with EDA	0.0053	0.046813491731882095

From Table II, it can be observed that the results for the LSTM model without EDA show a significant disparity between the training metrics (both loss and MAE) and the validation metrics, indicating a strong likelihood of overfitting. The model performs exceptionally well on the training data but struggles with the validation data that it has not encountered before. In contrast, the results for the LSTM model with EDA demonstrate a consistent decline in both training loss and training MAE, suggesting effective learning on the training data. Although there is a difference between the training and validation metrics, this difference is relatively small and stable from the midpoint of the epochs to the end, indicating that overfitting may be minimal or effectively managed. On the other hand, the results for the GRU model with EDA reveal a considerable difference between the training MAE and validation MAE, with training MAE consistently lower than validation MAE. This significant gap between the training metrics (both loss and MAE) and validation metrics strongly suggests overfitting. The model learns very well from the training data but performs poorly on unseen validation data, as evidenced by the continuous decrease in training loss/MAE while validation loss/MAE remains stable or even increases.

V. CONCLUSION

The extensive results of this study indicate that EDA significantly improves the preparation of sequential data for processing by RNNs, allowing essential patterns in the data to be identified and incorporated into the modeling process. By employing EDA techniques specifically for LSTM and GRU models, the risk of overfitting can be reduced. As a result, EDA positively impacts the performance, stability, and interpretability of the RNN model, helping to minimize biases and variances in predictions. The findings—loss: 0.0724; MAE: 0.1913—at epoch 15 show that, in general, higher epoch numbers correspond to lower loss and MAE values. However, it is important to note that if training loss continues to decrease while validation loss begins to rise, this may indicate overfitting. This situation suggests that the model has become too dependent on the specifics of the training data and is failing to develop the robust capabilities needed for accurate predictions on new datasets.

REFERENCES

[1] T. Lips, “Enhanced Sales Management: Using Digital Forecasting,” in Performance Management in Retail and the Consumer Goods Industry, 2019, pp. 247–256.

[2] H. Fischer, S. Seidenstricker, and J. Poeppelbuss, “The triggers and consequences of digital sales: a systematic literature review,” J. Pers. Sell. Sales Manag., vol. 43, no. 1, pp. 5–23, 2023.

[3] E. Widiastuti, M. Nugroho, and A. Halik, “The Effect of Service Quality, Brand Image and Social Media on Choosing Decisions With E-Wom and Attitude as Intervening Variables on Students of Private …,” The Seybold Report. admin369.seyboldreport.org, 2023.

[4] S. M. Amaninder Kaur, “Digital Transformation in Business Era,” Manag. Issues Digit. Transform. Glob. Mod. Corp., 2021.

[5] I. V. Pustokhina and D. A. Pustokhin, “A Comparative Analysis of Traditional Forecasting Methods and Machine Learning Techniques for Sales Prediction in E-commerce,” Am. J. Bus. Oper. Res., vol. 10, no. 2, pp. 39–51, 2023.

[6] C. Zhang, Y. X. Tian, and Z. P. Fan, “Forecasting sales using online review and search engine data: A method based on PCA–DSFOA–BPNN,” Int. J. Forecast., vol. 38, no. 3, pp. 1005–1024, 2022.

[7] K. Saraswathi, N. T. Renukadevi, S. Nandhinidevi, S. Gayathridevi, and P. Naveen, “Sales prediction using machine learning approaches,” AIP Conf. Proc., vol. 2387, no. June, pp. 3619–3625, 2021.

[8] S. Cheriyan, S. Ibrahim, S. Mohanan, and S. Treesa, “Intelligent Sales Prediction Using Machine Learning Techniques,” Proc. - 2018 Int. Conf. Comput. Electron. Commun. Eng. iCCECE 2018, pp. 53–58, 2018.

[9] C. A. Leke and T. Marwala, “Introduction to Deep Learning BT - Deep Learning and Missing Data in Engineering Systems,” C. A. Leke and T. Marwala, Eds. Cham: Springer International Publishing, 2019, pp. 21–40.

[10] R. Hanocka and H. T. D. Liu, “An introduction to deep learning on meshes,” ACM SIGGRAPH 2021 Courses, SIGGRAPH 2021, 2021.

[11] P. M. J. Samonte, E. Britanico, K. E. M. Antonio, J. E. J. Dela Vega, T. J. P. Espejo, and D. C. Samonte, “Applying Deep Learning for the Prediction of Retail Store Sales,” no. 20, pp. 1–11, 2023.

[12] K. Geetha, “Deep Learning and Sentiment Analysis Improve E-commerce Sales Prediction,” 2023 International Conference on Data Science and Network Security, ICDSNS 2023. 2023.

[13] X. Yin, “Prediction of Merchandise Sales on E-Commerce Platforms Based on Data Mining and Deep Learning,” Sci. Program., vol. 2021, 2021.

[14] D. Wang, H. Xiao, Q. Sun, and Y. Chen, “The application of deep learning algorithm in marketing intelligence,” Proc. 2019 IEEE 3rd Adv. Inf. Manag. Commun. Electron. Autom. Control Conf. IMCEC 2019, no. Imcec, pp. 1356–1360, 2019.

[15] M. P. Geetha, “Deep learning architecture towards consumer buying behaviour prediction using multitask learning paradigm,” J. Intell. Fuzzy Syst., vol. 46, no. 1, pp. 1341–1357, 2024.

[16] H. Zhu, “A deep learning based hybrid model for sales prediction of E-commerce with sentiment analysis,” Proceedings - 2021 2nd International Conference on Computing and Data Science, CDS 2021. pp. 493–497, 2021.

[17] S. Wang, “M-GAN-XGBOOST model for sales prediction and precision marketing strategy making of each product in online stores,” Data Technol. Appl., vol. 55, no. 5, pp. 749–770, 2021.

[18] X. Ma, “E-Commerce Review Sentiment Analysis and Purchase Intention Prediction Based on Deep Learning Technology,” J. Organ. End User Comput., vol. 36, no. 1, pp. 1–29, 2024.

[19] D. Jelonek, "Big Data Analytics in the Management of Business," MATEC Web Conf., vol. 125, pp. 1–6, 2017.

[20] S. M. Ansari, "A Review Paper On Big Data Analytics in Cloud," Int. J. Adv. Res. Sci. Commun. Technol., vol. 4, no. 11, pp. 477–483, 2021.

[21] H. Alsghaier, "The Importance of Big Data Analytics in Business: A Case Study," Am. J. Softw. Eng. Appl., vol. 6, no. 4, p. 111, 2017.

[22] A. Ayuningtyas, S. Mokodenseho, A. M. Aziz, D. Nugraheny, and N. D. Retnowati, "Big Data Analysis and Its Utilization for Business Decision-Making," West Sci. Inf. Syst. Technol., vol. 1, no. 01, pp. 10–18, 2023.

[23] K. B. Singh and M. A. Arat, "Deep Learning in the Automotive Industry: Recent Advances and Application Examples," pp. 1–14, 2019.

[24] Q. Xia, P. Jiang, F. Sun, Y. Zhang, X. Wang, and Z. Sui, "Modeling consumer buying decision for recommendation based on multi-task deep learning," Int. Conf. Inf. Knowl. Manag. Proc., pp. 1703–1706, 2018.

[25] Nisha, "Customer Behavior Prediction using Deep Learning Techniques for Online Purchasing," 2023 2nd International Conference for Innovation in Technology, INOCON 2023. 2023.

[26] W. D. Liu, "Application of Computer Vision on E-Commerce Platforms and Its Impact on Sales Forecasting," J. Organ. End User Comput., vol. 36, no. 1, 2024.

[27] K. Zhao and C. Wang, "Sales Forecast in E-commerce using Convolutional Neural Network," Retrieved from http//arxiv.org/abs/1708.07946, 2017.

[28] Y. Qi, C. Li, H. Deng, M. Cai, Y. Qi, and Y. Deng, "A Deep Neural Framework for Sales Forecasting in E-Commerce," pp. 299–308, 2019.

[29] C. Yang and Z. Liu, "Application of Computer Vision In Electronic Commerce," 2021.

[30] M. A. Kassem, "A novel deep learning model for detection of inconsistency in e-commerce websites," Neural Comput. Appl., 2024.

[31] S. Wang, C. Liu, X. Gao, H. Qu, and W. Xu, "Session-Based Fraud Detection in Online E-Commerce Transactions Using Recurrent Neural Networks," Springer Nat. Link, vol. 10536, no. Computer Science, pp. 241–252, 2017.

A Rapid Drift Modeling Method Based on Portable LiDAR Scanner

Zhao Huijun[1], Liu Chao[2], Qi Yunpu[3], Song Zhanglun[4], Xia Xu[5]
Norin Mining Limited, Beijing 100053, China[1, 2, 3, 4]
DIMINE Co., Ltd., Changsha 410083, China[5]

***Abstract*—Traditional measurement methods in underground mining tunnels have faced inefficiencies, limited accuracy, and operational challenges, consuming significant time and labor in complex environments. These limitations severely restrict the efficiency and quality of mine management and engineering design. To enhance the efficiency and accuracy of 3D modeling in underground tunnels, this study combines portable 3D LiDAR scanning technology with simultaneous localization and mapping. This integration enables autonomous positioning and efficient modeling without external positioning signals. The proposed approach effectively acquires high-resolution 3D data in complex environments, ensuring data accuracy and model reliability. High-resolution scanning of multiple critical areas was conducted on-site, with inertial navigation systems correcting the device's pose information. Automated data processing software was used for filtering, denoising, and modeling the collected data, leading to precise 3D tunnel models. Validation results indicate that portable laser scanning technology offers significant advantages in efficiency, accuracy, and safety, meeting the geological surveying and engineering needs of mining operations. The application of portable 3D laser scanning technology demonstrates considerable benefits in the rapid modeling of underground tunnels, providing effective technical support to improve mine management efficiency and safety. It also reveals broad application prospects.**

Keywords*—Underground mining; 3D modeling; portable 3D laser scanning; simultaneous localization and mapping (SLAM); mine surveying; inertial measurement unit (IMU)

I. Introduction

As mining resource development deepens, the scale and complexity of underground tunnel networks increase, raising higher demands for geological measurement and engineering management in mines[1], [2], [3]. High-precision 3D modeling forms the foundation of safe mining production and is crucial for optimizing engineering design and enhancing management efficiency [4], [5]. However, achieving efficient and accurate 3D modeling in the complex and harsh conditions of underground tunnels presents significant challenges.

Traditional surveying methods rely heavily on total stations and distance measuring devices. These methods often require manual point setting and observation, which are cumbersome and time-consuming, especially in poorly lit and dusty underground environments [6], [7]. Such conditions can compromise the efficiency and accuracy of data collection. Additionally, these methods depend significantly on the experience of survey personnel, making them prone to human error and limiting rapid responses to unexpected situations. Although static 3D laser scanning technology has significantly improved measurement accuracy, its application in underground spaces remains constrained by the complexity of equipment setup and positioning, which reduces flexibility and efficiency.

Recent advancements in 3D laser scanning technology, particularly the emergence of portable laser scanning devices, have provided an efficient and flexible solution for 3D modeling in complex environments. These portable devices utilize a non-contact measurement approach, enabling the rapid acquisition of high-resolution point cloud data. When combined with simultaneous localization and mapping (SLAM) technology, these devices achieve autonomous positioning and mapping in GPS-denied environments. Due to these technical advantages, portable laser scanning devices have increasingly been applied in fields such as construction [8], [9], [10], geological exploration [11], [12], [13], and autonomous driving [14], [15], [16]. Despite such advancements, research and application in underground tunnel environments remain limited, necessitating further exploration of portable laser scanning technology in these settings.

In this context, this study proposes a technical solution for rapid modeling of underground mine tunnels by integrating portable 3D laser scanning devices with inertial measurement units (IMU). By combining SLAM algorithms with inertial data in complex underground environments, this approach addresses laser radar positioning error accumulation and generates high-precision 3D models through automated data processing. The objective of this study is to validate the feasibility of portable 3D laser scanning technology in underground tunnel modeling and evaluate its advantages in terms of efficiency, precision, and applicability. Through testing and analysis of actual mining projects, the performance of this technology is assessed, and its potential applications in mining management and engineering design are further explored. This research provides an effective technical pathway for efficient modeling of underground tunnels and contributes to advancing 3D modeling technology in complex environments.

The remainder of this paper is organized as follows: Section II reviews related works and identifies research gaps. Section III details the proposed methodology integrating portable LiDAR, SLAM, and IMU. Section IV presents experimental results from engineering applications. Section V discusses the findings and compares with existing methods. Finally, Section VI concludes the study and outlines future directions.

II. Related Work

In recent years, research on 3D modeling of mining tunnels has gained traction, focusing on the application of 3D laser scanning technology, point cloud data processing and analysis, and innovations in positioning and navigation technologies. Traditional surveying methods for mining tunnels often rely on total stations and distance measuring devices. However, these methods face challenges such as inadequate measurement accuracy, cumbersome operations, and high time costs. In contrast, 3D laser scanning technology enables the rapid acquisition of high-precision spatial information, making it the preferred method for mining tunnel measurement and modeling [17], [18], [19]. Particularly in complex underground environments, laser radar can obtain data non-contact, overcoming the limitations of traditional surveying methods.

The introduction of SLAM technology has significantly enhanced data collection accuracy and efficiency in mining tunnel 3D modeling [20], [21], [22]. By continuously updating the device's pose, SLAM avoids the reliance on external positioning signals, which are often unavailable in GPS-denied underground environments.

IMU have also played a crucial role in mining 3D modeling by providing real-time motion state information [23], [24]. This data is vital for improving the stability and accuracy of SLAM systems in dynamic environments.

Regarding point cloud data processing, extracting useful information from large-scale point cloud datasets and removing noise remains a significant challenge [25], [26], [27], [28]. Existing studies have addressed this through techniques such as point cloud filtering, registration, and down-sampling. Point cloud filtering eliminates noise points generated during scanning, thereby improving data quality. Point cloud registration aligns data collected from different perspectives or times to construct continuous and consistent 3D models.

Despite substantial progress in the field of mining tunnel 3D modeling, the integration of portable laser scanning technology with SLAM and IMU still faces challenges. It is crucial to further minimize positioning errors in confined and complex spaces and ensure system stability in harsh environments. The maturity of the technology and its widespread application in the field require further validation and optimization. In Table I, the limitations of existing 3D modeling methods are summarized.

TABLE I. Limitations of Existing 3D Modeling Methods

Method	*Accuracy*	*Efficiency*	*Positioning Dependency*	*Environmental Adaptability*
Total Station	Medium	Low	High	Poor (dusty/low-light)
Static LiDAR	High	Medium	Medium	Moderate
SLAM-only Systems	Medium	High	None	Good (limited in dynamics)

While prior studies have advanced SLAM and IMU integration, three key limitations persist: (1) Cumulative errors in SLAM-based systems under prolonged operation, (2) Inadequate sensor fusion strategies for dynamic underground conditions, and (3) Limited validation in large-scale mining networks. Our approach addresses these through tight LiDAR-IMU coupling with error compensation (Section III.D) and field validation in 2.3km tunnel networks (Section IV).

III. Rapid Tunnel Modeling

A. Portable 3D LiDAR Scanning

Portable 3D laser scanning devices are developed based on SLAM technology, as shown in Fig. 1. Initially proposed by Smith and Cheeseman in 1986 to address spatial uncertainty estimation, SLAM primarily solves navigation and localization issues in unknown environments. It determines the device's position and orientation by observing features such as corners and columns, incrementally constructing an environmental map based on positional changes. SLAM enables concurrent localization and mapping, making it particularly suitable for complex underground mining environments where GPS signals are unavailable.

Fig. 1. Portable 3D laser scanning device.

Portable 3D scanning devices perform rapid and continuous scanning of complex environments while in motion, automatically capturing accurate three-dimensional spatial information. Compared to traditional fixed surveying techniques, portable devices address frequent setup issues caused by laser line-of-sight limitations, significantly enhancing fieldwork efficiency. These devices reduce operator labor intensity and safety risks while enabling extensive continuous 3D scanning tasks in underground mines.

In underground mining applications, portable 3D laser scanning devices integrate SLAM technology for autonomous localization and mapping in environments without external positioning signals, ensuring high-precision surveying capabilities. Their compact, lightweight, and ergonomic designs make them easy to transport and operate in confined or complex environments. Additionally, the integration of complementary sensors such as gyroscopes, accelerometers, and GPS enhances the collection of positioning and orientation data, substantially improving scanning accuracy and data reliability. These devices are capable of scanning various underground features, including tunnels, voids, and chutes, supporting tasks such as dimension measurement, volume calculations, morphology analysis, over-excavation and under-excavation assessments, engineering

quantity estimations, surface area measurements, and extraction of 3D contours and tunnel profiles.

B. Spatial Calibration Methods

The working principle of a 3D laser scanning system relies on laser ranging technology, which measures the time difference between laser emission and reflection to calculate the distance to an object's surface. By emitting a laser beam toward the object and receiving its reflected signal, the scanner determines the distance d based on the time-of-flight formula, as shown in Eq. (1).

$$d = \frac{c \cdot \Delta t}{2} \tag{1}$$

Where, c is the speed of light, and Δt is the time difference between the emission and reception of the laser.

To determine the three-dimensional coordinates of points on the object's surface, it is essential to measure not only the distance but also the emission angles of the laser beam. The laser scanner is typically mounted on a rotating device, allowing the laser beam to cover the entire scanning area through rotation. The three-dimensional coordinates (x, y, z) of each laser point can be calculated using the measured pitch angle θ, yaw angle ϕ, and the measured distance d, as shown in Eq. (2).

$$\begin{cases} x = d \cdot \cos(\theta) \cdot \cos(\phi) \\ y = d \cdot \cos(\theta) \cdot \sin(\phi) \\ z = d \cdot \sin(\theta) \end{cases} \tag{2}$$

Where, θ is the pitch angle of the laser beam, and ϕ is the yaw angle.

In practical applications, the laser scanner may move or rotate, necessitating consideration of its spatial position and orientation. Assuming the position of the scanner in space is PP and its rotation matrix is R, the three-dimensional coordinates (X, Y, Z) of the object's surface can be corrected using Eq. (3).

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = R \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix} \tag{3}$$

Where, R is the rotation matrix.

By repeatedly performing this process, the scanner can calculate the three-dimensional coordinates of each laser point, thereby constructing the complete 3D point cloud data of the object or scene.

C. Mobile Scanning Method

The mobile scanning method uses highly integrated portable 3D laser scanning equipment to collect environmental data and generate 3D point cloud models during movement. The structure of the mobile scanning device, shown in Fig. 2, comprises a probe and a backpack. These components are interconnected via various communication protocols, including Ethernet, RS-485, RS-232, and GPIO control.

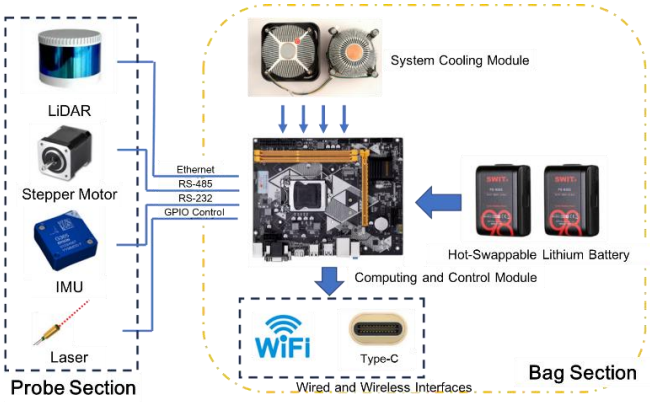


Fig. 2. Mobile scanning device structure.

The probe includes a LiDAR, stepper motor, inertial navigation module, and laser rangefinder, responsible for generating high-precision 3D point cloud data. The LiDAR measures the distance and shape of target objects, the stepper motor controls the scanning angle and rotation, the inertial navigation module provides device orientation information, and the laser rangefinder is used for accurate distance measurement.

The backpack focuses on performance and portability, containing a system cooling module, hot-swappable lithium battery, high-performance computing and control module, as well as wireless and wired interfaces. The cooling module ensures the stability of long-term device operation, the lithium battery provides continuous power, and the high-performance computing and control module processes data in real-time to generate precise point cloud models. The wireless and wired interfaces facilitate data transmission and device communication, supporting remote monitoring and data synchronization.

The mobile scanning process is illustrated in Fig. 3 and consists of four main steps:

1) Initial scanning and point cloud acquisition: Activate the LiDAR and scan the surrounding 3D space with the probe to capture the initial point cloud dataset, denoted as P_0.

2) Mobile remeasurement and data update: After the probe moves, use the LiDAR to rescan the previous area, obtaining the updated point cloud data denoted as P_1. This process continues to acquire new datasets P_2, P_3, and so on.

3) Trajectory estimation and spatial reconstruction: By comparing the two consecutive point cloud datasets P_t and P_{t+1}, estimate the probe's displacement and rotation by minimizing $T_{t \to t+1} = \arg\min_T \sum \left\| P_i^{t+1} - T \cdot P_i^t \right\|^2$, where T is the transformation matrix. The SLAM pipeline utilizes a feature-based Iterative Closest Point (ICP) algorithm that starts with feature extraction, where edge features (tunnel ribs) and planar features (tunnel walls) are extracted using curvature thresholds (edge: curvature > 0.1; planar: curvature < 0.05). Following this, dynamic object removal is performed through statistical outlier removal (50-neighbor points, 1.5 σ threshold) and DBSCAN clustering (ε=0.3m, minPts=15).

4) Point cloud registration and data fusion: Utilize the ICP algorithm to accurately align and fuse datasets $P_0, P_1, \ldots, P_t$, resulting in the creation of a continuous point cloud model M. Fine registration minimizes point-to-plane residuals using Huber loss with a threshold of δ=0.2m, expressed as $E = \sum \left| n_j^T (Rp_i + t - q_j) \right|^2$, where n_j represents the normal vector of the target plane j.

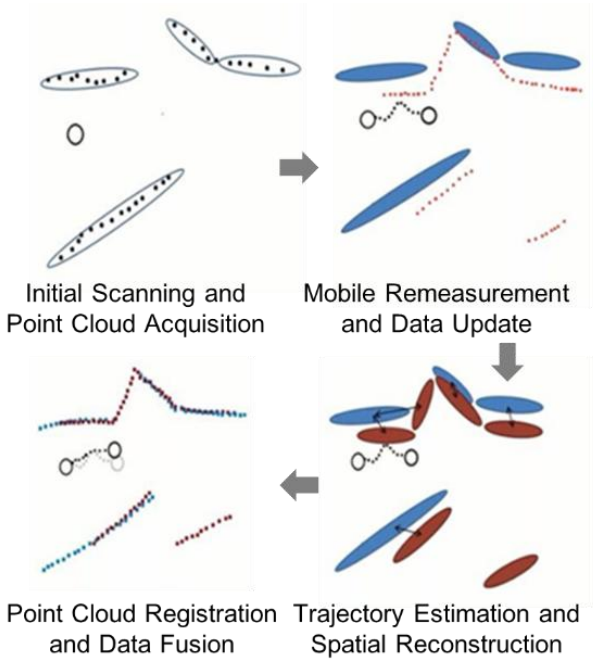


Fig. 3. Mobile scanning process.

D. Inertial State Integration

To mitigate positioning errors and inaccuracies caused by dynamic motion, an IMU is integrated into the system for spatial calibration. Traditional LiDAR systems often require stable platforms, but portable devices operating in dynamic environments can encounter vibrations and irregular movements that impact data accuracy. The IMU enhances measurement robustness by providing supplementary pose data.

The LiDAR is mounted on a linkage mechanism with dual rotational axes, as shown in Fig. 4. This configuration enables continuous acquisition of both laser ranging and inertial data, providing real-time six degrees of freedom (6DOF) trajectory estimation.

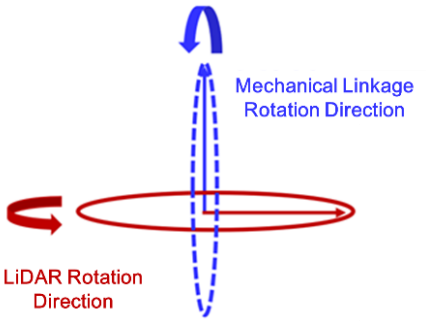


Fig. 4. Device operation rotation diagram.

By utilizing encoder information, the specific rotation angle of the LiDAR probe relative to the stepper motor at each moment can be obtained. Let the rotation angle of the LiDAR with respect to the stepper motor be denoted as $\theta(t)$. The initial point cloud data P is then corrected to align its coordinate system with that of the stepper motor. The corrected point cloud data is represented as P', with the transformation given by $P' = R(\theta(t)) \cdot P$, where $R(\theta(t))$ is the rotation matrix based on the angle $\theta(t)$.

The corrected point cloud data is tightly integrated with inertial information. Pre-integration of the inertial data is performed to accumulate the sensor's motion trajectory. This pre-integration process primarily relies on the acceleration $a(t)$ and angular velocity $\omega(t)$ measured by the IMU to compute increments in position and orientation, as shown in Eq. (4).

$$\begin{cases} \Delta p = \int_{t_0}^{t_1} v(t) \cdot dt \\ \Delta v = \int_{t_0}^{t_1} a(t) \cdot dt \\ \Delta R = \int_{t_0}^{t_1} \omega(t) \cdot dt \end{cases} \tag{4}$$

The IMU-LiDAR fusion utilizes a Kalman filter, which includes a state vector representing position, velocity, quaternion, and the biases from the accelerometer and gyroscope. During the update process, the IMU performs predictions at a high frequency, while the LiDAR provides corrections at a lower frequency using feature residuals. Furthermore, bias compensation is implemented through online calibration via sliding window optimization.

Combining the point cloud constructed through the mobile scanning method with the corrected point cloud data and inertial data allows for real-time estimation of the sensor's 6DOF trajectory.

E. 3D Model Construction

The method for constructing a 3D model using LiDAR and IMU primarily consists of tightly coupled joint optimization of the LiDAR inertial odometry and point cloud mapping, as illustrated in Fig. 5.

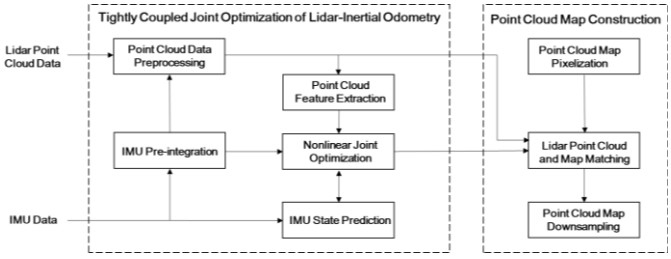


Fig. 5. System working principle.

In the tightly coupled joint optimization of the LiDAR inertial odometry, the IMU data is pre-integrated, and both intra-frame and inter-frame motions of the LiDAR are computed. The

calculated intra-frame motion is then used to preprocess the point cloud data, mitigating the effects of motion distortion through motion compensation using the IMU-derived trajectory. Additionally, ground segmentation is performed using RANSAC plane fitting with a defined inlier threshold, enabling the separation of ground points from other data. After this preprocessing, voxel downsampling is applied with differing resolutions for near-field and far-field regions to balance computational efficiency and detail retention. Feature extraction is then performed on the processed point cloud data, alongside IMU state prediction. Finally, based on the established IMU constraints and LiDAR constraints, a nonlinear optimization model is constructed to obtain the state variables required for the mapping process. The optimization process is supported by global techniques facilitated by the Ceres solver using the Levenberg-Marquardt algorithm, ensuring robust and accurate results.

In the point cloud mapping phase, the point cloud map is first voxelized. Subsequently, the optimized results from the odometry phase are used to match the LiDAR point cloud with the point cloud map. Lastly, the point cloud map undergoes downsampling to reduce its size.

IV. Engineering Applications

A. Project Background

The CMC project in the Democratic Republic of the Congo is a complex underground mining development system engineering project. The underground tunnel network includes multiple key areas and structures, with the areas to be scanned depicted in Fig. 6. These areas encompass light vehicle ramps, heavy vehicle ramps, and various sections such as the 1110, 940, and 990 levels, as well as the measures ramp. Additionally, the project involves significant underground facilities including ventilation shafts, auxiliary shafts, and water chambers.

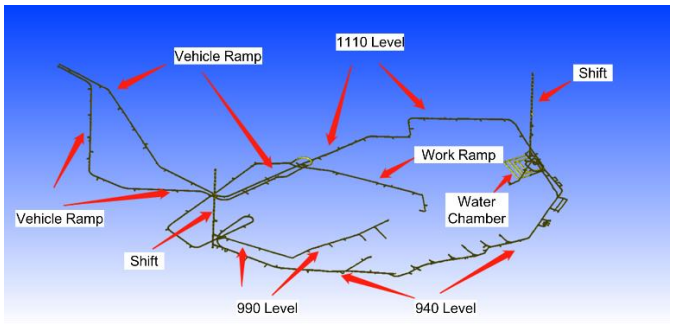


Fig. 6. Overall view of the underground mine survey area.

The geographical distribution and functional positioning of these tunnels and facilities result in a considerable project length, traversing various geological and structural layers. In such a complex underground environment, traditional measurement and modeling methods may not meet the required accuracy and efficiency standards. Therefore, a rapid tunnel modeling approach based on portable LiDAR technology is employed to achieve high-precision 3D positioning of the underground space.

B. Field Data Collection

Due to the complexity and long distances of the underground tunnel system, multiple data collection methods were employed to adapt to different environmental conditions and scenarios, as illustrated in Fig. 7. In accessible areas, operators conducted scanning with handheld devices. In the light and heavy vehicle ramp areas, scanning was performed from vehicles to enhance collection efficiency. In the 940 and 1110 level, where rainfall was prevalent, handheld scanning was utilized, with operators using umbrellas to shield against water spray, particularly in the 940 level. The device sensor continuously rotated at a steady speed to perform panoramic scanning, covering a distance of 100 meters, allowing for complete scanning with a single pass through the scene. For hazardous areas inaccessible to personnel, such as chutes and mined-out zones, drones or extension poles equipped with scanning devices were employed. These methods ensured high-quality data acquisition even in complex and dangerous environments.

Fig. 7. Handheld scanning and vehicle-mounted handheld scanning.

Data collection was conducted in segments due to the overall length of the underground tunnels. Each segment's data was distinguished by different colors, facilitating subsequent stitching and adjustments. To align the data with the geodetic coordinate system, the equipment needed to be calibrated to control points within the tunnels. In accessible areas, such as the tunnel floor, alignment was achieved by overlapping the crosshair on the device's base with the control point. In areas inaccessible to personnel, such as the tunnel ceiling, a visible laser on the device was used to aim at the control point. The positions of the control points were recorded on a mobile control device. This laser anchoring method did not require additional personnel, resulting in higher accuracy and allowing for single-person operation.

C. Indoor Data Processing

After data collection, the coordinates of the control points needed to be entered on the mobile control device. In the mobile control device's menu, the system for analysis was selected, and by clicking “Retrieve Data,” all collected data could be viewed. These data packets were named according to the collection time, with corresponding control point information displayed on the right side of each packet. By clicking on the control point information for a specific data packet, a window for inputting the control point coordinates would appear, where the geodetic coordinates (in meters) for the corresponding control point could be entered. After inputting the control point coordinates, the

relevant collected data was selected on the left side, followed by clicking the "Analyze" button to initiate data processing. The control point input interface on the mobile device is shown in Fig. 8.

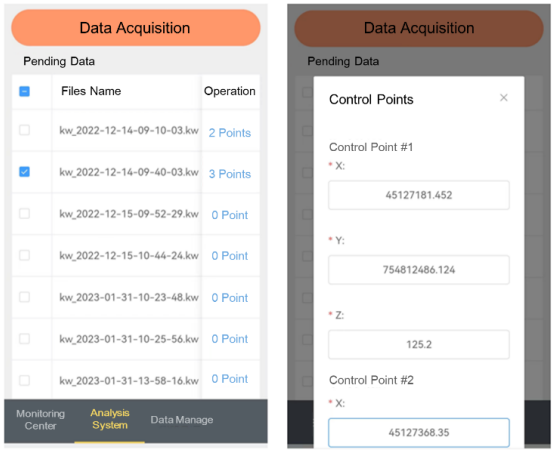


Fig. 8. Mobile interface for control point input.

After the data processing was completed, the data is transmitted to the PC via a network. The data folder contains several files, including filtermap.dxf, ControlPoint.txt, map.las, filter_map.las, map.pcd, filtermap.pcd, filtermap.ply, and Path.txt. Among these, filtermap.dxf contains the 3D model data generated automatically by the device, ControlPoint.txt records the geodetic and control point coordinates, while map.las and filter_map.las represent the original and decimated point cloud data of the map, respectively. Additionally, map.pcd and filtermap.pcd contain the original and decimated point cloud data in PCD format, filtermap.ply is the 3D object format point cloud data, and Path.txt contains the trajectory data of the device.

The LAS point cloud data undergoes filtering and denoising processes to eliminate noise generated by moving objects within the tunnel. This process retains valid points while reducing the overall data size to approximately one-tenth of the original. After filtering and denoising, a 3D model of the tunnel is created, generating a closed triangular mesh model that preserves the original morphology of the tunnel, as shown in Fig. 9.

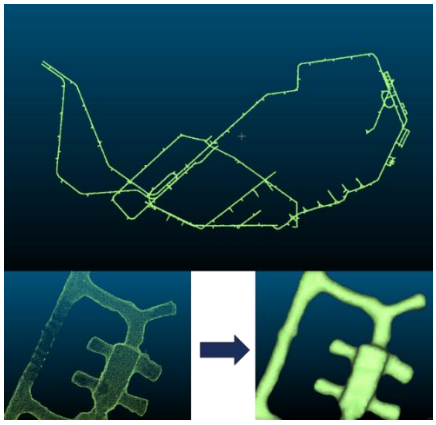


Fig. 9. Drift 3D model.

V. Results

A. Accuracy Validation

Once the tunnel's 3D model is established in point cloud processing software, it can be saved in DXF format and imported into 3D mining software such as Dimine for further processing. This facilitates a series of tasks related to geological surveying and mining operations, including ore body modeling, tunnel design, resource estimation, and extraction planning. By comparing the 3D model with the CAD base map in detail, the accuracy and completeness of the model can be analyzed, as shown in Fig. 10.

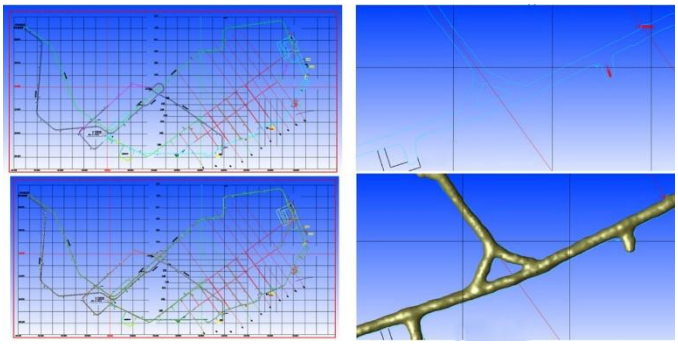


Fig. 10. Comparison of CAD drawing and model.

First, the CAD base map and the DXF format 3D model data are imported separately into the software. The comparison tool is then used to overlay the 3D model with the CAD base map to check their alignment. If discrepancies are identified between the 3D model and the CAD base map, adjustments and corrections can be made based on the comparison results to ensure high precision of the model.

The accuracy assessment involves comparing the 3D model with the actual tunnel, measuring deviations in both longitudinal and lateral positions. Several representative measurement points were selected at different locations within the tunnel, and actual measurements were taken. These data were then compared with the corresponding data in the 3D model, as shown in Table II.

TABLE II. Comparison between Measured Model and Actual Measurements

Measurement Point	*Measurement Content*	*Actual Value (m)*	*Measured Value (m)*	*Absolute Error (m)*	*Relative Error (%)*
Vehicle Ramp	Longitudinal Deviation	269.87	270.12	-0.25	0.09
	Lateral Deviation	4.72	4.74	-0.02	0.42
Water Chamber	Longitudinal Deviation	15.18	15.14	0.04	0.26
	Lateral Deviation	4.21	4.22	-0.01	0.24
Stope Substation	Longitudinal Deviation	5.37	5.36	0.01	0.19
	Lateral Deviation	4.92	4.93	-0.01	0.20

The results demonstrate that the absolute errors across all test points were minimal, with relative errors ranging from 0.09% to 0.42%, ensuring that the generated 3D models meet precision requirements for mining applications. These findings indicate the reliability of the presented method in achieving high-accuracy 3D modeling in complex underground environments.

Beyond validation through direct measurements, 3D models were also compared against original design CAD drawings within mining software (e.g., DIMINE). The overlay alignment evaluation revealed minimal discrepancies, confirming the completeness and consistency of the 3D models with the engineering designs (Fig. 10). This validation process underscores the potential of the proposed method for enhancing tunnel design implementation and quality control.

B. Efficiency Analysis

The efficiency of the proposed method was assessed by examining the time required for data collection and processing compared to traditional methods such as total station measurements and static LiDAR scanning. The portable LiDAR system significantly reduced fieldwork times due to its ability to capture spatial data continuously and autonomously while in motion. Specifically, the system completed a 2.3 km tunnel network survey in approximately 6 hours, including setup, scanning, and preliminary control point calibration. Traditional methods would typically require at least 2–3 days for similar coverage, highlighting the advantages of the portable LiDAR solution in terms of operational efficiency.

Furthermore, the automated data processing workflow reduced the time needed for point cloud filtering, denoising, and stitching by at least 30% compared to conventional software pipelines. The integration of SLAM and IMU technologies eliminated the dependency on fixed reference points, further streamlining data collection in environments with limited access or visibility.

C. Safety and Adaptability

The adaptability and safety of the proposed method were demonstrated through its successful deployment in hazardous and challenging underground environments. The compact and portable design of the scanning device facilitated access to narrow and confined spaces, eliminating the need for extensive manual setups. Additionally, automated scanning from mobile platforms (e.g., vehicles) or remote systems (e.g., drones) allowed for data collection in areas that were otherwise inaccessible or unsafe for personnel, such as waterlogged tunnels and mined-out zones.

The flexibility of the system was further underscored by its performance in varying environmental conditions. For instance, handheld scanning with umbrellas in the 940 level ensured high-quality data acquisition despite water spray and poor lighting. These capabilities emphasize the robustness of the system in adapting to diverse underground scenarios.

D. Discussion

The proposed portable LiDAR-based rapid modeling method demonstrates significant advancements in accuracy, efficiency, and adaptability for underground mining applications. The integration of SLAM and IMU technologies effectively addresses common challenges in GPS-denied environments, ensuring reliable data collection and high-resolution 3D modeling. Compared to traditional surveying methods, the proposed approach achieves superior time efficiency and operational flexibility while maintaining precision, as evidenced by relative errors below 0.5% and successful deployment in complex tunnel networks. Furthermore, the system's portability and ability to operate under adverse conditions, such as low light, dust, and water spray, highlight its robustness and practical value for diverse underground scenarios.

Despite these advantages, residual challenges remain. Pose error accumulation during extended operations and computational demands for large-scale point cloud processing represent areas for improvement. Environmental factors, such as extreme humidity or dust, may also introduce noise, impacting data quality. Future research should focus on optimizing SLAM algorithms, developing real-time processing capabilities, and integrating complementary sensors to enhance system performance. Additionally, further miniaturization and automation could expand its applications across other industries.

Overall, this study provides a comprehensive solution for efficient and accurate tunnel modeling, addressing critical limitations of traditional methods and laying a foundation for broader adoption of portable LiDAR technologies in mining and beyond.

VI. Conclusion

This study addresses the challenge of rapid 3D modeling of underground mining tunnels by proposing an innovative solution based on portable 3D laser scanning technology and SLAM techniques. Experimental validation demonstrates the feasibility and effectiveness of this approach in complex environments. By integrating portable 3D laser scanning technology with SLAM algorithms and IMU, this research successfully achieves efficient 3D modeling of underground tunnels without relying on external positioning signals. This method effectively overcomes the limitations of traditional measurement techniques, such as low efficiency and restricted accuracy in underground settings, significantly enhancing data collection speed and quality while providing technical support for high-resolution modeling of complex tunnel networks. The findings indicate that portable devices can maintain measurement stability and precision even in adverse underground conditions characterized by poor lighting and air quality, offering new insights for 3D modeling in other complex subterranean environments. The generated 3D models meet practical engineering requirements in terms of accuracy, completeness, and visualization, making them applicable in various domains, including geological surveying, engineering design, construction acceptance, and safety management. Moreover, this study demonstrates the flexibility and practicality of the technology in complex scenarios, highlighting its ease of operation and high degree of automation in data processing, thus providing crucial technical support for digital mining and intelligent mine management.

However, this research has certain limitations, and future studies will focus on optimizing SLAM algorithms to reduce pose error accumulation, as well as integrating multi-sensor

fusion technologies (e.g., collaborative use of vision and LiDAR) to enhance data collection accuracy. Building on this foundation, the development of more compact and efficient portable devices to meet the surveying needs of various complex environments will be pursued.

ACKNOWLEDGMENT

This research was supported by the Key Research and Development Program of Hunan Province (Grant No: 2022GK2061).

REFERENCES

[1] S. Tavani et al., "Smartphone assisted fieldwork: Towards the digital transition of geoscience fieldwork using LiDAR-equipped iPhones," Earth-Science Reviews, vol. 227, p. 103969, Apr. 2022, doi: 10.1016/j.earscirev.2022.103969.

[2] T. G. Garrison et al., "Assessing the lidar revolution in the Maya lowlands: A geographic approach to understanding feature classification accuracy," Progress in Physical Geography: Earth and Environment, vol. 47, no. 2, pp. 270–292, Apr. 2023, doi: 10.1177/03091333221138050.

[3] P. Padmanabhan, C. Zhang, and E. Charbon, "Modeling and Analysis of a Direct Time-of-Flight Sensor Architecture for LiDAR Applications," Sensors, vol. 19, no. 24, Art. no. 24, Jan. 2019, doi: 10.3390/s19245464.

[4] L. Zhao, M. Zhang, and X. Jin, "Construction and application of a high precision 3D simulation model for geomechanics of the complex coal seam," Sci Rep, vol. 11, no. 1, p. 21374, Nov. 2021, doi: 10.1038/s41598-021-00709-5.

[5] L. Chen et al., "High-Precision Positioning, Perception and Safe Navigation for Automated Heavy-Duty Mining Trucks," IEEE Transactions on Intelligent Vehicles, vol. 9, no. 4, pp. 4644–4656, Apr. 2024, doi: 10.1109/TIV.2024.3375273.

[6] M. Laguillo, P. Segarra, J. A. Sanchidrián, and F. Beitia, "A novel borehole surveying system for underground mining: Design and performance assessment," Measurement, vol. 194, p. 111021, May 2022, doi: 10.1016/j.measurement.2022.111021.

[7] S. Kahraman, J. Rostami, and A. Naeimipour, "Review of Ground Characterization by Using Instrumented Drills for Underground Mining and Construction," Rock Mech Rock Eng, vol. 49, no. 2, pp. 585–602, Feb. 2016, doi: 10.1007/s00603-015-0756-4.

[8] A. Aryan, F. Bosché, and P. Tang, "Planning for terrestrial laser scanning in construction: A review," Automation in Construction, vol. 125, p. 103551, May 2021, doi: 10.1016/j.autcon.2021.103551.

[9] C. Wu, Y. Yuan, Y. Tang, and B. Tian, "Application of Terrestrial Laser Scanning (TLS) in the Architecture, Engineering and Construction (AEC) Industry," Sensors, vol. 22, no. 1, Art. no. 1, Jan. 2022, doi: 10.3390/s22010265.

[10] A. Piekarczuk, A. Mazurek, J. Szer, and I. Szer, "A Case Study of 3D Scanning Techniques in Civil Engineering Using the Terrestrial Laser Scanning Technique," Buildings, vol. 14, no. 12, Art. no. 12, Dec. 2024, doi: 10.3390/buildings14123703.

[11] S. Kumar Singh, B. Pratap Banerjee, and S. Raval, "A review of laser scanning for geological and geotechnical applications in underground mining," International Journal of Mining Science and Technology, vol. 33, no. 2, pp. 133–154, Feb. 2023, doi: 10.1016/j.ijmst.2022.09.022.

[12] R. Hudson, F. Faraj, and G. Fotopoulos, "Review of close-range three-dimensional laser scanning of geological hand samples," Earth-Science Reviews, vol. 210, p. 103321, Nov. 2020, doi: 10.1016/j.earscirev.2020.103321.

[13] J. Telling, A. Lyda, P. Hartzell, and C. Glennie, "Review of Earth science research using terrestrial laser scanning," Earth-Science Reviews, vol. 169, pp. 35–68, Jun. 2017, doi: 10.1016/j.earscirev.2017.04.007.

[14] J. Li et al., "Real-time self-driving car navigation and obstacle avoidance using mobile 3D laser scanner and GNSS," Multimed Tools Appl, vol. 76, no. 21, pp. 23017–23039, Nov. 2017, doi: 10.1007/s11042-016-4211-7.

[15] X. Lian et al., "Biomass Calculations of Individual Trees Based on Unmanned Aerial Vehicle Multispectral Imagery and Laser Scanning Combined with Terrestrial Laser Scanning in Complex Stands," Remote Sensing, vol. 14, no. 19, Art. no. 19, Jan. 2022, doi: 10.3390/rs14194715.

[16] M. J. Sumnall et al., "Effect of varied unmanned aerial vehicle laser scanning pulse density on accurately quantifying forest structure," International Journal of Remote Sensing, vol. 43, no. 2, pp. 721–750, Jan. 2022, doi: 10.1080/01431161.2021.2023229.

[17] C. Zhang, "Mine laneway 3D reconstruction based on photogrammetry," Transactions of Nonferrous Metals Society of China, vol. 21, pp. s686–s691, Dec. 2011, doi: 10.1016/S1003-6326(12)61663-X.

[18] A. Adamek, J. Będkowski, P. Kamiński, R. Pasek, M. Pełka, and J. Zawiślak, "Method for Underground Mining Shaft Sensor Data Collection," Sensors, vol. 24, no. 13, Art. no. 13, Jan. 2024, doi: 10.3390/s24134119.

[19] X. Lian and H. Hu, "Terrestrial laser scanning monitoring and spatial analysis of ground disaster in Gaoyang coal mine in Shanxi, China: a technical note," Environ Earth Sci, vol. 76, no. 7, p. 287, Apr. 2017, doi: 10.1007/s12665-017-6609-6.

[20] L. Fahle, E. A. Holley, G. Walton, A. J. Petruska, and J. F. Brune, "Analysis of SLAM-Based Lidar Data Quality Metrics for Geotechnical Underground Monitoring," Mining, Metallurgy & Exploration, vol. 39, no. 5, pp. 1939–1960, Oct. 2022, doi: 10.1007/s42461-022-00664-3.

[21] Z. Ren, L. Wang, and L. Bi, "Robust GICP-Based 3D LiDAR SLAM for Underground Mining Environment," Sensors, vol. 19, no. 13, Art. no. 13, Jan. 2019, doi: 10.3390/s19132915.

[22] A. Ellmann, K. Kütimets, S. Varbla, E. Väli, and S. Kanter, "Advancements in underground mine surveys by using SLAM-enabled handheld laser scanners," Survey Review, vol. 54, no. 385, pp. 363–374, Jul. 2022, doi: 10.1080/00396265.2021.1944545.

[23] L. Wang, S. Zhang, J. Qi, H. Chen, and R. Yuan, "Research on IMU-Assisted UWB-Based Positioning Algorithm in Underground Coal Mines," Micromachines, vol. 14, no. 7, Art. no. 7, Jul. 2023, doi: 10.3390/mi14071481.

[24] M.-G. Li, H. Zhu, S.-Z. You, and C.-Q. Tang, "UWB-Based Localization System Aided With Inertial Sensor for Underground Coal Mine Applications," IEEE Sensors Journal, vol. 20, no. 12, pp. 6652–6669, Jun. 2020, doi: 10.1109/JSEN.2020.2976097.

[25] S. Tong, Q. Jia, N. Song, W. Zhou, T. Duan, and C. Bao, "Determination of gold(III) and palladium(II) in mine samples by cloud point extraction preconcentration coupled with flame atomic absorption spectrometry," Microchim Acta, vol. 172, no. 1, pp. 95–102, Feb. 2011, doi: 10.1007/s00604-010-0466-2.

[26] Y. Wang, W. Tu, and H. Li, "Fragmentation calculation method for blast muck piles in open-pit copper mines based on three-dimensional laser point cloud data," International Journal of Applied Earth Observation and Geoinformation, vol. 100, p. 102338, Aug. 2021, doi: 10.1016/j.jag.2021.102338.

[27] S.-J. Lee and S.-O. Choi, "Analyzing the Stability of Underground Mines Using 3D Point Cloud Data and Discontinuum Numerical Analysis," Sustainability, vol. 11, no. 4, Art. no. 4, Jan. 2019, doi: 10.3390/su11040945.

[28] M. Chen, Y. Feng, S. Wang, and Q. Liang, "A Mine Intersection Recognition Method Based on Geometric Invariant Point Detection Using 3D Point Cloud," IEEE Robotics and Automation Letters, vol. 7, no. 4, pp. 11934–11941, Oct. 2022, doi: 10.1109/LRA.2022.3208366.

Dialogue-Based Disease Diagnosis Using Hierarchical Reinforcement Learning with Multi-Expert Feedback

Shi Li, Xueyao Sun
College of Computer and Control Engineering, Northeast Forestry University, Harbin, China

***Abstract*—In order to minimize the stochasticity of agents used in disease diagnosis within the dialogue system, and to enable them to interact with users based on the inherent connections between symptoms and diseases, while simultaneously addressing the issue of limited medical data, we propose the Hierarchical Reinforcement Learning with Multi-expert Feedback framework. The framework constructs a reward model in the lower-level networks of the hierarchical structure. Here, the discriminator leveraging the concept of adversarial networks generates rewards by evaluating the authenticity of symptom query sequences generated by the agent, and the large language model of human experts synthesizes various factors to assess the reasonableness of the agent's current symptom queries, thereby guiding the learning of the policy network. The algorithm addresses the deficiencies in data characteristics and improves the policy's capability to leverage feature information, thus making the process of disease diagnosis more aligned with clinical practice. Experimental results demonstrate that the proposed framework achieves diagnostic success rates of 61.5% on synthetic datasets and 84.4% on real-world datasets, while requiring fewer dialogue turns on average. Both metrics surpass those of conventional approaches, further indicating the framework's strong generalization ability.**

***Keywords*—Disease diagnosis; dialogue system; large language model; reinforcement learning; reward model; adversarial network; dialogue agent**

I. Introduction

Due to the broad application prospects and significant commercial benefits of task-oriented dialogue systems [1], researchers have introduced them into the medical field, forming research focused on medical dialogue systems to alleviate the problem of strained medical resources [2].The rapid development of deep learning has propelled advancements in the field of disease diagnosis, and the application of deep reinforcement learning in conversational disease diagnosis research is gradually emerging [3]. However, deep reinforcement learning still faces two major challenges in this field: first, it relies too heavily on random sampling [4] and lacks a universal reward and punishment mechanism adaptable to different application scenarios. For example, Kao et al. [5]use +1 and 0 as reward values, whereas Zhong et al. [6] set substantial rewards and penalties of +20 and -100. Second, medical data resources are scarce and valuable, and the available feature information is also very limited, which greatly restricts research in disease diagnosis.

In response to the aforementioned challenges, this study introduces a Multi-Expert Feedback-based Hierarchical Reinforcement Learning framework (ME-RL), utilizing a hierarchical structure to manage complex sequential decision-making issues, and leveraging the principles of Generative Adversarial Networks, it establishes an adversarial workflow comprising a generator and a discriminator. Utilizing the evaluations from the discriminator and the feedback from the large language model of human experts (AIExpert) as low-level rewards, the framework optimizes the policy and improves the agent's learning efficiency and performance within dynamic environments. In addition, a novel joint evaluation metric is introduced, which simultaneously takes into account the diagnostic success rate and the average number of dialogue turns in disease diagnosis tasks, thereby offering a more comprehensive assessment of the model's effectiveness.

The remainder of this paper is organized as follows: Section II introduces the related work. Section III provides an overview of the methods. Section IV introduces the experimental design. Section V presents the experimental results and related analysis. Section VI summarizes the paper and proposes future research directions.

II. Related Work

Task-oriented dialogue systems have received widespread attention in recent years, encompassing areas such as ticket booking services and e-commerce [7], [8]. In the context of strained medical resources, researchers have begun exploring how to integrate dialogue systems into the smart healthcare field to improve doctors' time utilization and provide patients with preliminary diagnostic references [9]. Through continuous development, conversational disease diagnosis systems have demonstrated significant potential in simplifying diagnostic processes, reducing costs, and efficiently collecting patient medical history information [10].

Due to the fact that conversational disease diagnosis involves multiple consecutive time stages and interactive steps, its decision-making characteristics are highly compatible with the applicability of reinforcement learning (RL) methods. Researchers model the disease diagnosis task as a Markov Decision Process and utilize reinforcement learning methods to optimize policy learning [6]. By employing policy gradient algorithms, the agent can make decisions based on the specific symptom information provided by patients.

With the continuous advancement of deep learning, many researchers have integrated deep learning models [11], [12] for disease diagnosis prediction. Deep reinforcement learning, which combines deep learning and reinforcement learning, has also been rapidly applied to research on conversational disease diagnosis [13]. In the context of symptom-oriented disease diagnosis dialogue systems, researchers such as Wei et al. [14] have designed an automatic diagnostic system based on human-computer dialogue. This system models the symptom inquiry and disease diagnosis processes as a Markov Decision Process and employs a Deep Q-Network (DQN) policy learning algorithm. In each interaction cycle, the intelligent agent can choose to inquire about specific symptoms or make diagnostic judgments based on the situation. The agent learns the optimal strategy by continuously optimizing to maximize the expected simulated rewards. The dialogue system constructed in this manner is relatively comprehensive, marking a preliminary breakthrough in addressing automatic diagnostic medical dialogue problems. Since then, reinforcement learning has gradually become the mainstream choice among researchers in this field [15].

To further enhance the performance of automatic diagnostic models, Chen et al. [16] proposed a multi-action policy representation method, enabling agents to timely recommend medical examination items to assist in diagnosis. Xu et al. [2] further introduced the KR-DS system, which integrates Knowledge Path-based Deep Q-Networks (DQN). This system incorporates knowledge branches and knowledge routing branches that capture the associations between diseases and symptoms within the deep reinforcement learning framework, thereby considering external probabilistic symptom information related to the reinforcement learning framework. This integration enhances the rationality and accuracy of medical dialogue decision-making. Tiwari et al. [17] designed a hierarchical reinforcement learning framework that integrates a knowledge-driven mechanism. By embedding a Potential Candidate Module (PCM), the framework enhances the agent's efficiency in organizing and probing symptom information. Yan et al. [18] addressed the challenges of insufficient diagnostic evidence and irrelevant symptom inquiries by constructing a more comprehensive medical dialogue dataset and combining experiential diagnostic knowledge with a medical knowledge graph. Subsequently, Yan et al. [19] proposed the EIRAD dialogue system based on a medical knowledge graph, which enhances the interpretability and accuracy of the diagnostic process by utilizing the topological structure of the knowledge graph.

Existing methods have laid the foundation for the optimization and development of dialogue-based disease diagnosis systems. However, they still largely rely on random sampling, lack a universal reward-punishment mechanism adaptable to different application scenarios, and are overly dependent on scarce medical data resources. To address these limitations, the framework proposed in this paper integrates hierarchical reinforcement learning, adversarial networks, and large language models, constructing a general and effective reward model. Additionally, it leverages the rich database of large language models to mitigate the issue of sparse data features, enabling the dialogue agent to generate reasonable disease query sequences, thereby enhancing the performance of the disease diagnosis system. Experimental results demonstrate that, compared to traditional learning methods, the proposed approach significantly improves diagnostic accuracy, reduces dialogue turns, and enhances the generalization capability of the model. Table I provides a summary and comparison of existing research, highlighting their methodologies, strengths and weaknesses.

TABLE I. RELATED WORK COMPREHENSIVE

Methodology	Strengths	Weaknesses	Reference
RL, Knowledge graph	Leveraging the relationships between symptoms and diseases	Lack of a universal reward model	[2]
RL, Prioritization of severe pathologies	Exploration-confirmation framework, Severe pathology prioritization	Lack of a universal reward model, Dependency on dataset quality	[3]
RL, Feature rebuilding	Sparse feature exploration, Faster diagnosis	Lack of a universal reward model	[4]
RL, Context-aware symptom checking	Context-aware diagnosis, Generalization across tasks	Lack of a universal reward model	[5]
RL	Novel dataset, Task-oriented framework	Dependency on dataset quality	[14]
RL, Label-guided exploration	Efficient exploration, Multiple test suggestions	Lack of real-world dataset validation	[16]
RL, Potential candidate module	Context-aware and knowledge-guided investigation	Lack of real-world dataset validation	[17]
RL, Medical knowledge graph	Interpretability, Integration of medical knowledge	Dependency on data quality	[19]

III. METHOD DESIGN

A. Overall Framework

The structure of the disease diagnosis model based on Multi-Expert Feedback Hierarchical Reinforcement Learning is illustrated in Fig. 1. Its core component is the diagnostic agent, which is divided into a two-tier structure comprising high-level and low-level policies, and consists of four key modules: controller, generator, evaluator, and classifier. Furthermore, it is equipped with a user simulator. Among them, the controller is responsible for scheduling the generator or classifier to operate. The generator is tasked with inquiring about potential symptoms from the patient, the evaluator assesses the rationality of the actions chosen by the generator and provides rewards, while the classifier is used to inform the user simulator of possible disease types based on the currently collected information. The user simulator is designed to mimic patient behavior, interact with the agent, and return intrinsic rewards for the disease classifier. The sum of the rewards obtained by the generator and the disease classifier constitutes the external rewards received by the controller.

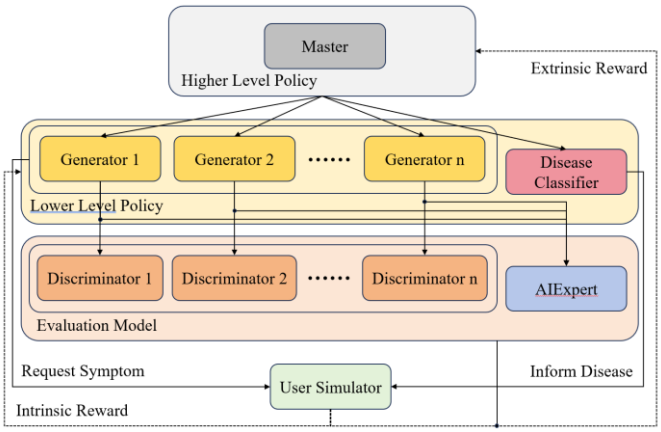


Fig. 1. Overall framework of ME-RL.

B. Dialogue Policy

1) Agent policy: The automatic diagnosis tasks can be formulated as a Markov Decision Process and utilize reinforcement learning to optimize dialogue policies. In this framework, the policy includes the state space S, action space A, state transition probabilities TP, and reward model R. Each state (s, $s \in S$) in the state space is composed of several elements, including the current state of the dialogue agent, the symptoms pending inquiry, user symptom feedback, a list of all confirmed symptoms, the number of dialogue rounds, and reward signals. The action space includes all possible disease diagnosis options and symptom inquiry actions, with each action corresponding to the examination of a particular symptom or the judgment of a disease. State transition probabilities dictate how the next state s_{t+1} is formed given the current state s_t and the action a taken, represented as s_{t+1} = TP(s_t, a). The reward model offers the probability of receiving an immediate reward r_{t+1} after taking action a_t at time step t, denoted as r_{t+1} = R(s_t, a_t).

The agent's goal is to seek an optimal dialogue strategy that maximizes the total accumulated reward R during one round of interaction [20]. The formula for the total reward R is presented below:

$$R = \sum_{i=1}^{N} \sum_{t=0}^{T} \gamma^t \cdot r_t \tag{1}$$

where N represents the total number of dialogues in a round, T represents the total number of dialogue turns in the current conversation, γ^t represents the discount factor, and r_t represents the immediate reward received by the agent at the t-th time step. $\sum_{t=0}^{T} \gamma^t \cdot r_t$ can be approximated as [21]:

$$Q^*(s, a) = E_{s'}[r + \gamma \cdot \max_{a' \in A} Q^*(s', a')] \tag{2}$$

where Q(s, a) denotes the Q-function representing the state-action value, which is used to estimate the value acquired after performing action a in the current state s, leading to a new state s'. During the training process, the estimation of Q(s, a) is progressively updated based on experience and learning, until it converges and stabilizes.

2) User simulation: The user simulator is designed to emulate the interaction process between real patients and the automated diagnostic system. At the beginning of each dialogue session, the simulator randomly selects a user objective from the experimental dataset as the basis for the simulation. Each user objective comprises two types of symptom information: one is the explicit symptoms clearly expressed by the user, and the other is the implicit symptoms that can only be uncovered through dialogue interaction. Additionally, the user objective includes the actual disease labels. At the initiation of the dialogue, the diagnostic agent directly receives the explicit symptoms as the initial information. Subsequently, the user simulator engages in further dialogue interactions with the agent based on the implicit symptoms. During the dialogue process, the simulated user will respond based on whether the symptom currently inquired by the agent exists in their objective: correctly, incorrectly, or with uncertainty. When the agent successfully diagnoses the correct disease within the maximum number of rounds, the dialogue is considered a successful interaction; otherwise, it is deemed a failure.

C. Hierarchical Structure

Considering that the agent needs to handle a vast action selection space in complex environments, a hierarchical structure with two levels of policies has been constructed, as shown in Fig. 2.

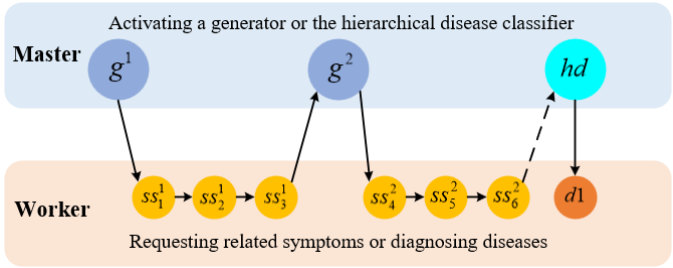


Fig. 2. Interaction of two layer strategies.

3) Higher policy: The controller manages the high-level policies, with its core function being to schedule and activate lower-level policies, and the action space is denoted as $A_m = \{g_i \mid i = 1,2,...,n\} \cup \{d\}$. Here, g_i represents the activation of generator G_i, and action d represents the activation of the disease classifier. At time step t, the controller takes the current dialogue state s_t as input and, based on the policy function π_m, decides which action a_m ($a_m \in A_m$) to take, thereby leading to a new state s_{t+1}.Once the controller activates a lower-level agent, the agent will engage in multiple rounds of interaction with the user until its subtask is completed. The learning problem of the controller can be viewed as a Semi-Markov Decision Process. At the t-th time step, after the controller executes action a_m, the immediate reward r_m it receives can be expressed as the discounted sum of all external rewards accumulated by the

lower-level policies it has triggered while performing a series of actions a_i. Therefore, the external reward for the controller is calculated as follows:

$$r_m = \begin{cases} \sum_{t_1=1}^{T_1} \gamma_m r_{t_1}^{g}, & if\ a_m = g_i \\ r_d, & if\ a_m = d \end{cases} \quad (3)$$

where $i = 1,\ldots,n$, r_m represents the external rewards obtained by the high-level policy at the t-th round, γ_m is the discount factor of the controller, T_1 is the number of original actions of the generator, and g_i and d are the actions to activate generator G_i and classifier D, respectively. When the generator is activated by the controller's action a_m, it will select appropriate actions a_i based on the current state to query symptoms from the patient based on the current disease group, thereby activating the evaluator to provide corresponding intrinsic rewards $r_{t_1}^{g}$. If the disease classifier is triggered, it will perform disease identification and return rewards r_d based on the diagnostic results. The objective of the high-level agent is to maximize external rewards within a round through a Semi-Markov Decision Process, thus allowing us to derive the controller's loss function [22]:

$$L(\theta_m) = [(r_m + \gamma_m \max_{a'} Q_m(s', a', \theta_m^-)) - Q_m(s, a, \theta_m)]^2 \quad (4)$$

where $L(\theta_m)$ represents the loss of the controller at time step t, while θ_m and θ_m^- are the frozen parameters during the current and past training iterations, respectively.

4) Lower policy: The low-level policy consists of a generator and a disease classifier, with an action space comprising symptom inquiries and disease classification actions. The low-level policy is centrally scheduled by the high-level policy, which provides the current environmental state (s_t) as input to guide the low-level policy in selecting the most appropriate basic action a_i. Once action a_i is executed, the dialogue environment state updates to s_{t+1}, and the low-level agent immediately receives rewards or penalties (r_t^{g}) based on the suitability of the current action a_i for the current state s_t. Because the rewards obtained by the low-level policy are direct results of the actions currently being executed, these rewards are termed intrinsic rewards. The calculation of intrinsic rewards is as follows:

$$r_t^{g} = w_{dis} r_t^{dis} + w_{ai} r_t^{ai} \quad (5)$$

where r_t^{dis} represents the reward provided by the discriminator, r_t^{ai} represents the feedback evaluation from AIExpert, w_{dis} and w_{ai} denote their respective weights. Since the SD dataset does not include real patient-doctor symptom inquiry sequences, w_{dis} is set to 0.4 and w_{ai} is set to 0.6.When using the MZ4 dataset, which includes real patient-doctor symptom inquiry sequences, w_{dis} and w_{ai} are both set to 0.4.

The generator is described in detail in subsequent chapters. The classifier adopts a hierarchical classifier structure, consisting of group classifiers and specific disease classifiers. When the disease classifier is activated by the controller, the currently confirmed symptoms are input into the disease classifier. The group classifier then designates the appropriate specific disease classifier, which conducts a detailed diagnosis based on the current disease group. Each specific disease classifier for a particular disease group is constructed as a three-layer neural network structure, including a hidden layer, which is used to accurately identify and diagnose each specific disease within the selected disease group, and to feedback the disease diagnosis result with the highest probability to the user.

D. Evaluation Model Based on Multi-Expert Feedback

The evaluation model based on multi-expert feedback consists of multiple discriminators corresponding to the generator and one AIExpert. The adversarial structure-based evaluation process is illustrated in Fig. 3. Both the generator and the discriminator have been pre-trained. The generator employs the maximum likelihood estimation method, predicting the next possible symptom based on the current state during training, ultimately generating a symptom inquiry sequence (fake symptom sequence). Additionally, a dedicated data repository is established to store both real and fake symptom sequences for pre-training the discriminator. The real symptom sequences in the database are directly sampled from user objectives in the dataset. An example of AIExpert model interaction is shown in Fig. 4, where AIExpert uses interaction history as the environmental information for large language model (LLM) scoring. Firstly, the RL model generates the next symptom to inquire, which, along with the interaction history, is processed into a prompt. Subsequently, the LLM uses this prompt to generate a score for the queried symptom, and finally, the score is returned to the RL model as a reward.

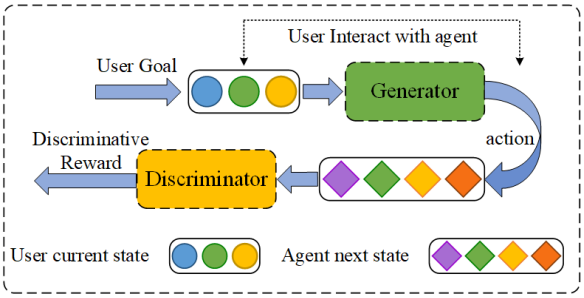


Fig. 3. Adversarial structures in low-level policy.

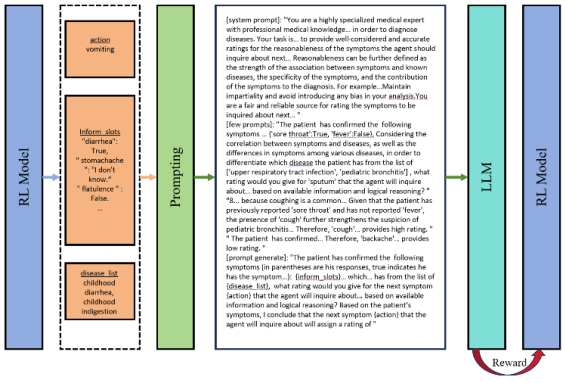


Fig. 4. AIExpert interaction example.

1) Generator: In the context of medical disease diagnosis, the core task of the sequential query generator is to dynamically generate a series of follow-up questions targeting the patient's initial chief complaint. This process aims to simulate a real medical consultation workflow, guiding patients to elaborate in detail on their current medical history and the specific manifestations and variations of each related symptom. Specifically, it involves intelligently generating the next most reasonable and diagnostically beneficial question based on the initial information provided by the patient, resulting in patient responses of confirmation, denial, or uncertainty.

The generator is a low-level strategy within the agent. Once the high-level strategy selects symptom inquiries, generator G_i is activated and interacts with the patient to collect information on specific symptom groups. The action workspace of generator G_i is: $A_{G_i} = \{y \mid y \in Y_i\}$. Here, y represents the requested symptom, and Y_i represents the set of specific disease symptoms corresponding to this generator. At the t-th time step, generator G_i receives the current state s_t as input from the high-level strategy controller and generates the action y_t to inquire about the next symptom ($y_t \in A_{Gi}$). Subsequently, the dialogue state is updated, and generator G_i immediately receives an intrinsic reward. The task of the sequential generator is to construct a coherent and seemingly realistic situational dialogue sequence starting from the user's initial symptom description. Its objective is to "fool" the discriminator, making the discriminator believe that the generated series of questions and answers belong to actual patient-doctor interactions. That is, to generate a symptom inquiry sequence from the initial state in order to maximize its expected return [23]:

$$J(\theta) = \mathrm{E}_{a \sim G_\theta(a|s)}[\mathrm{Q_D}(s,a) \mid \theta] \tag{6}$$

where a represents the inquiry action taken by generator G in the current state s. The rewards are derived from the evaluation model. $Q_D(s, a)$ is the state-action value function, which approximates the value obtained when taking action a in the current state s.

2) Discriminator: Each generator is equipped with its own discriminator, which is pre-trained to assess whether the sequence of queries is genuine. The objective is to maximize the probability of correctly classifying real data and minimize the probability of misclassifying generated data. The output of the discriminator serves as a reward for the generator, encouraging the generator to produce symptom inquiry sequences that are indistinguishable from real symptom sequences. The discriminator is trained using a Deep Neural Network and outputs a single scalar value that represents the probability of a symptom sequence originating from real data rather than from the generator.

For interactive dialogue systems, this paper not only focuses on the quality of the final generated complete symptom sequences but also emphasizes real-time feedback during the diagnostic process. A discriminator model has been trained to evaluate and assign rewards to both fully and partially observed symptom sequences, enabling it to provide immediate rewards for partial symptom information that users gradually disclose during the dialogue process. At the t-th time step, given the current dialogue state s, the generator produces the next inquiry action a. The concatenation of the generated action a and the input s is fed into the discriminator, resulting in a reward from the discriminator [24]:

$$r_t^{dis} = \mathrm{D}(s = Y_{1:t-1}, a = y_t \mid \mathrm{G}(a \mid s)) \tag{7}$$

where $Y_{1:t-1}$ represents the symptom sequence contained in the current state s. y_t represents the next symptom inquiry generated by the generator based on s.

3) AIExpert: The actions generated by the generator in each round receive expert evaluation feedback from AIExpert. AIExpert is a module designed to simulate human experts (doctors) by integrating large language model. AIExpert supplements medical data and knowledge related to disease diagnosis through a vast knowledge base and utilizes the capabilities of LLM to simulate the behavior of experts assisting in diagnosis, thereby achieving a dual purpose.

Each interaction between the RL model and AIExpert generates a new interaction record, which is appended to the interaction history. As time steps progress, the interaction history grows, potentially exceeding the current LLM's context length [25]. To address this issue, before each interaction, history retrieval is performed to retrieve the latest 10 interaction records related to the current disease list from the user's interaction history (including symptoms and related scores) as the RL model's observation. Then, the RL model asks the user about a symptom as its action, and the action and observation are subsequently used to generate the prompt.

Each dialogue generates a prompt containing the detailed information required by the LLM, including the symptom inquiry history corresponding to the current generator, the symptom to be queried, and the current disease list. Additionally, the prompt includes system prompts and example prompts. The system prompt informs the LLM of its positioning as a medical expert and provides guidance on the aspects to consider when generating feedback scores (in addition to the applicability of symptom sequences for known symptoms, it also considers other factors, such as whether the current symptom can help distinguish between diseases in the current disease list). Example prompts serve as a reference for subsequent scoring by the LLM. The purpose of constructing the prompt is to enable the LLM to accurately provide a score for the current queried symptom. The output generated by the LLM based on the prompt is processed and returned to the RL model as a reward from AIExpert feedback.

IV. Experimental Design

A. Dataset Description

This study evaluates all methods on both synthetic and real-world datasets. The description of the synthetic dataset (SD) is shown in Table II. It is a publicly available dataset synthesized from medical library knowledge, with the nine elements being G1, G4, G5, G6, G7, G12, G13, G14, and G19. Each user goal includes one explicit symptom and multiple implicit

symptoms. The description of the real-world dataset MZ-4 [14] is shown in Table III. It is a publicly available dataset constructed from data collected in real medical scenarios. Symptoms extracted from the user's self-report are defined as explicit symptoms, while symptoms discovered through dialogue are defined as implicit symptoms.

TABLE II. OVERVIEW OF SYMCAT-SD-90 DATASET

Entries	Value
# of user goal	30000
# of disease	90
# of groups	9
# of symptoms	266
Avg. # of implicit symptoms	2.6

TABLE III. OVERVIEW OF MZ-4 DATASET

Entries	Value
# of user goal	1733
# of disease	4
# of groups	2
# of symptoms	230
Avg. # of implicit symptoms	5.46

B. Evaluation Metrics

This paper proposes a new combined evaluation metric that assesses the overall performance of the disease automatic diagnosis system in real-world scenarios from two perspectives: accuracy and average dialogue rounds. This combined metric is named ST, and its definition is as follows:

$$\mathrm{ST} = \mathrm{W}_{SR} \cdot SR \times 100 - \mathrm{W}_{Turn} \cdot Turn \tag{8}$$

where *SR* represents the success rate, *Turn* denotes the average number of dialogue rounds, and W_{SR} and W_{Turn} represent their respective weight coefficients. The success rate is a positive evaluation metric for the disease automatic diagnosis system, while fewer dialogue rounds are better. Therefore, in the experiment, W_{SR} is set to 1 and W_{Turn} to 0.5.

This paper uses five metrics to evaluate the effectiveness of the model: the success rate of disease diagnosis (SR), the average number of dialogue rounds (Turn), the ratio of symptoms correctly identified by the agent out of those requested (AMR), the ratio of the number of hidden symptoms identified by the agent to the total number of hidden symptoms in the user's target (SIR), and the ST metric proposed in this paper. The calculation methods for SR, AMR, and SIR are as follows:

$$\mathrm{SR} = \frac{\sum_{i=1}^{EL} DS_i}{EL} \tag{9}$$

$$\mathrm{AMR} = \frac{\sum_{i=1}^{i=EL}\sum_{j=1}^{j=t} \mathrm{m}_i / r_i}{EL} \times 100, \mathrm{m}_i \tag{10}$$

$$\mathrm{SIR} = \frac{\sum_{i=1}^{i=EL} m_i / \mathrm{f}_i}{EL} \times 100 \tag{11}$$

where *EL* represents the total number of dialogues in the simulated dialogue set. *DS* indicates whether the dialogue was successfully completed. *t* represents the total number of dialogue rounds in the *i*-th conversation, while r_i denotes the total number of symptoms asked by the dialogue agent in the *i*-th conversation. m_i represents the cumulative number of symptoms that the agent asked which actually match the patient's condition. f_i indicates the total number of hidden symptoms that the patient actually possesses in the *i*-th dialogue.

C. Baselines

To demonstrate the effectiveness of the proposed method, several benchmark methods were chosen for comparison:

SVM-ex: A model built using Support Vector Machine (SVM) [6], where the input is the one-time encoding of the patient's explicit symptoms, and the output is the disease label. SVM-ex&im, on the other hand, simultaneously considers both the patient's explicit and implicit symptoms, and its performance can be considered as the ideal upper limit achievable by a reinforcement learning-based dialogue agent.

HRL: A hierarchical reinforcement learning framework that employs a two-layer policy structure and a disease classifier [14].

KI-CD: A knowledge-driven hierarchical reinforcement learning framework [17] that introduces a potential candidate module as knowledge assistance and constructs a multi-level disease classifier. The KI-CD_PCM model includes only the potential candidate module without the multi-level disease classifier.

EIRAD: An automatic diagnostic evidence-based dialogue system with interpretable reasoning paths [19], based on the Medical Knowledge Graph, which explicitly utilizes the topological structure of the Medical Knowledge Graph to capture key symptoms of suspected diseases.

D. Experimental Setup

In this paper, 80% of the data in the dataset is used for training the agent, while 20% is used as the test set. Both the master controller and all generators employed a three-layer deep Q-learning network (DQN) with 512 nodes in the hidden layers. All discriminators used a three-layer neural network model, with the input layer dimension matching the number of corresponding symptoms, and the output layer consisting of a single node. The LLM in AIExpert used Llama-2-7B-Chat-GPTQ, a model that has been quantized using GPTQ to allow it to run within a 24GB memory limit without significantly affecting performance. Regarding parameter settings, the master controller and generators shared a discount factor of 0.95 and a learning rate of 0.0005. The learning rate for the discriminators was set higher at 0.01 to allow faster weight adjustment. The learning rate for the disease classifier was also set to 0.0005. During the entire training cycle for the master controller, after every 10 training steps, all generators,

discriminators, and disease classifiers would undergo a training update. Additionally, each cycle consisted of 100 dialogue sessions, and all neural networks were trained using the Adam optimizer.

V. Result and Analysis

A. Performance on SD Dataset

Five metrics were used to evaluate the model's performance on the SD dataset, with the results presented in Table IV. ME-RL_Adv refers to the model that only includes the discriminator, without AIExpert. Compared to the baseline methods, ME-RL showed significant advantages across all metrics. Except for the number of dialogue rounds, the SR, AMR, SIR, and ST increased by 8.08%, 24.72%, 6.15%, and 15.05%, respectively, compared to KI-CD. Among these, the improvement in AMR was particularly notable, indicating a significant enhancement in the agent's ability to identify the symptoms the user is suffering from, i.e. the model can generate more reasonable symptom query sequences. Furthermore, both the ME-RL_Adv and ME-RL_AI models performed worse than ME-RL, highlighting that the multi-expert feedback model, which uses both the discriminator and AIExpert, outperforms the single-feedback evaluation model.

TABLE IV. Performance on SD Dataset with DQN

Model	SR	Turn	AMR(%)	SIR(%)	ST
SVM-ex	0.321	-	-	-	-
HRL	0.504	12.95	10.49	29.56	43.93
KI-CD	0.569	16.11	09.95	48.33	48.85
ME-RL_Adv	0.594	15.32	10.66	38.74	51.74
ME-RL_AI	0.562	13.55	10.02	40.24	49.425
ME-RL	**0.615**	**10.61**	**12.41**	**51.30**	**56.20**
SVM-ex&im	0.732	-	-	-	-

TABLE V. Performance on SD Dataset with Dueling DQN

Model	SR	Turn	AMR(%)	SIR(%)	ST
HRL	0.478	**8.57**	**13.80**	29.24	43.52
KI-CD	0.523	14.71	11.02	40.03	46.79
ME-RL	**0.570**	12..86	10.85	**46.63**	**50.57**

TABLE VI. Performance on SD Dataset with DDQN

Model	SR	Turn	AMR(%)	SIR(%)	ST
HRL	0.426	**7.02**	13.52	22.26	39.09
KI-CD	0.507	10.60	13.70	**39.04**	45.40
ME-RL	**0.544**	7.63	**14.85**	38.91	**50.59**

Experiments were also conducted using Double Deep Q-Network (DDQN) and Dueling Deep Q-Network (Dueling DQN) algorithms to replace the DQN algorithm in the model. The experimental results are reported in Tables V and VI. Compared to other models using these two algorithms, the ME-RL model achieved higher accuracy and overall performance superior to all baseline models.

From the above experiments, it can be observed that the ME-RL model using DQN and DDQN algorithms has a higher average number of dialogue turns compared to the HRL model. This is considered reasonable, as more dialogues between the agent and the patient allow for the collection of additional information about the patient's latent symptoms, ensuring a more thorough investigation and accurate diagnosis. Meanwhile, the diagnostic success rate is the most crucial factor for any automated disease diagnosis system, serving as a prerequisite for the system's usability. Additionally, the ME-RL model consistently achieves the highest overall evaluation metric, further proving its effectiveness.

B. Performance on MZ-4 Dataset

Due to the limited number of disease types in the dataset, a single-layer disease classifier is used for disease diagnosis. The proposed model outperforms all baseline models across three algorithms (DQN, Double DQN, and Dueling DQN), achieving an accuracy rate exceeding 0.8, with an average number of dialogue turns remaining below 5. This indicates that the model is capable of reaching correct diagnostic results in most cases with fewer dialogue turns, demonstrating its outstanding performance. The experimental results are shown in Table VII.

TABLE VII. Performance on MZ-4 Dataset

Model	SR	Turn	ST
EIRAD with DQN	0.770	15.10	69.45
KI-CD_PCM with DQN	0.808	6.16	77.72
ME-RL with DQN	0.844	4.32	82.24
EIRAD with DDQN	0.732	12.23	67.09
KI-CD_PCM with DDQN	0.778	6.37	74.62
ME-RL with DDQN	0.813	4.21	79.20
EIRAD with Dueling DQN	0.720	10.86	66.57
KI-CD_PCM with Dueling DQN	0.757	4.90	73.25
ME-RL with Dueling DQN	0.801	3.02	78.59

C. Error Analysis

To analyze the agent's misdiagnosis cases, all instances where the agent made incorrect diagnoses on the SD dataset were collected, and a confusion matrix for the disease classifier was constructed. Fig. 5 shows the prediction accuracy for nine different disease groups. Upon examining the matrix, it is observed that the color intensity on the diagonal blocks is darker, indicating that the system tends to misclassify diseases into the same broad category. A partial reason for this phenomenon is that some diseases share many similar common symptoms, making it difficult for the classifier to distinguish them. However, it also demonstrates that even if the agent does not precisely predict the exact disease type, it can still correctly distinguish the general category of the disease to some extent, reflecting its diagnostic value.

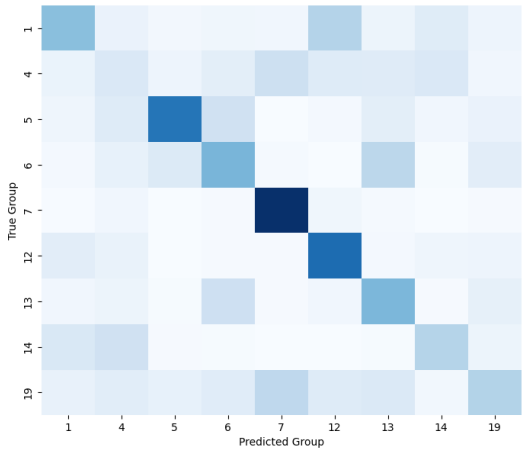


Fig. 5. Confusion matrix between different disease groups.

D. Performance of Different Generators

An analysis of the performance of different generators using the SD dataset is shown in Table VIII. It is observed that the best-performing generators, Generator 1 and Generator 8, correspond to disease groups with the fewest average number of latent symptoms, which could be one of the factors influencing the generator's performance. The average disease classification accuracy corresponding to these generators is higher than all baseline models, indicating that ME-RL facilitates the generation of reasonable symptom query sequences, thereby improving the accuracy of the disease classifier.

TABLE VIII. PERFORMANCE OF DIFFERENT GENERATORS

Model	Accuracy (%)
HRL_Avg	50.3
KI-CD_Avg	75.6
Generator 1(G1)	77.0
Generator 2(G4)	91.6
Generator 3(G5)	69.7
Generator 4(G6)	79.4
Generator 5(G7)	65.3
Generator 6(G12)	67.2
Generator 7(G13)	79.5
Generator 8(G14)	83.5
Generator 9(G19)	78.0

E. Ablation Study

Fig. 6 shows the accuracy curves of different dialogue agents based on ME-RL on the SD dataset, and Table IX displays their performance on the test set. The agent used in the ME-RL model is named the Unique-dis Agent, which indicates that each generator in the low-level agents is assigned a unique discriminator. The dialogue agent with only the evaluation model set for the master controller is named the Eva-Master Agent. The agent with a unified discriminator for all generators is named the Unified-dis Agent. All agents use the same AIExpert in their evaluation models. The analysis reveals that the Adv-Master Agent performs poorly due to the lack of reasonable distribution of low-level agent sequence references, while the Unified-dis Agent and Unique-dis Agent perform significantly better, indicating the validity of setting an evaluation model for low-level agents. The Unique-dis Agent significantly outperforms the Unified-dis Agent, which may be due to the large action space of the generators. Overall, it is reasonable to enhance the model's adversarial capability by assigning a corresponding discriminator to each generator.

TABLE IX. PERFORMANCE ON MZ-4 DATASET

Model	SR	Turn	AMR(%)	SIR(%)	ST
Eva-Master Agent	0.385	**7.12**	02.77	19.6	33.94
Unified-dis Agent	0.501	13.04	09.73	40.25	43.58
Unique-dis Agent	**0.615**	10.61	**12.41**	**51.30**	**56.20**

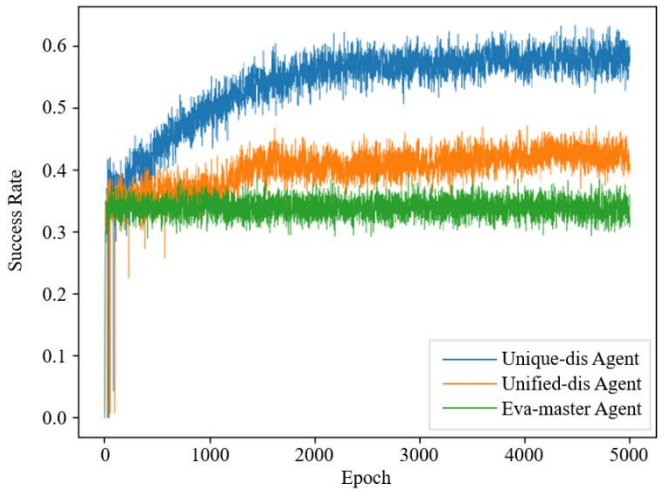


Fig. 6. Policy learning curves.

VI. CONCLUSION

This paper proposes a multi-expert feedback-based hierarchical reinforcement learning framework and successfully applies it to the field of automated disease diagnosis. The hierarchical structure allows the model to make decisions at different levels of abstraction, enhancing the model's effectiveness and modularity. Discriminators are introduced to the generators in the lower-level networks to evaluate symptom query sequences and generate rewards, guiding the agent to generate more targeted inquiry symptoms. The introduction of the AIExpert feedback model enriches the feature information of medical knowledge and data, and the powerful understanding and analytical capabilities of the model provide evaluations simulating human experts, improving the generator's ability to query symptoms. The multi-expert feedback model adjusts weights under different conditions, compensating for the limitations of a single expert and enhancing the model's robustness and generalization capability. Experiments show that, compared with baseline technologies, although the number of dialogue turns has not been reduced, the success rate, AMR, SIR, and the newly proposed comprehensive evaluation metric on the SD dataset have improved by 8.08%, 13.97%, 4.28%, and 5.22%, respectively. Future work will further explore strategies to reduce dialogue

turns in order to improve the model's performance. In future research, efforts will be made to integrate medical knowledge graphs to strengthen the associations between symptoms and diseases, thereby further enhancing the decision-making process and ensuring consistency with established medical knowledge.

Although the results are promising, the proposed framework still faces certain noteworthy limitations. Its effectiveness partially depends on high-quality and diverse medical data, if data are limited in quantity or unevenly distributed, diagnostic accuracy for less common or underserved patient populations may be compromised. Furthermore, the combination of adversarial training and large language models demands substantial computational resources, which may pose feasibility issues in resource-constrained environments. To address these challenges, future research should prioritize enriching and balancing medical datasets, which may involve collaborating with healthcare institutions to collect more diverse patient data or employing data augmentation techniques to better capture rare diseases. Meanwhile, adopting strategies such as model compression, knowledge distillation, and hardware acceleration (e.g. TPUs) can help mitigate high computational demands. Through these improvements, the proposed framework is expected to further enhance adaptability and reliability, thereby exerting a wider clinical impact across a variety of healthcare settings.

REFERENCES

[1] J. Sun, J. Kou, W. Shi, and W. Hou, “A multi-agent collaborative algorithm for task-oriented dialogue systems,” International Journal of Machine Learning and Cybernetics, pp. 1–14, 2024.

[2] L. Xu et al., “End-to-end knowledge-routed relational dialogue system for automatic diagnosis,” in Proc. AAAI Conf. Artif. Intell., vol. 33, no. 01, pp. 7346–7353, Jul. 2019.

[3] A. Fansi Tchango et al. , “Towards trustworthy automatic diagnosis systems by emulating doctors' reasoning with deep reinforcement learning,” Advances in Neural Information Processing Systems, vol. 35, pp. 24502–24515, 2022.

[4] Y. S. Peng, K. F. Tang, H. T. Lin, and E. Chang, “Refuel: Exploring sparse features in deep reinforcement learning for fast disease diagnosis,” in Advances in Neural Information Processing Systems, vol. 31, 2018.

[5] H.-C. Kao, K.-F. Tang, and E. Chang, “Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning,” in Proc. AAAI Conf. Artif. Intell., pp. 2305–2313, 2018.

[6] C. Zhong et al., “Hierarchical reinforcement learning for automatic disease diagnosis,” Bioinformatics, vol. 38, no. 16, pp. 3995–4001, 2022.

[7] A. Al-Hanouf and N. Al-Twairesh, “Building an Arabic flight booking dialogue system using a hybrid rule-based and data driven approach,” IEEE Access, vol. 9, pp. 7043–7053, 2021.

[8] Z. Borhanifard, H. Basafa, S. Z. Razavi, and H. Faili, “Persian Language Understanding in Task-Oriented Dialogue System for Online Shoping,” in Proc. 11th Int. Conf. Inf. Knowl. Technol. (IKT), pp. 79–84, Dec. 2020.

[9] X. Zhao, L. Chen and H. Chen, "A Weighted Heterogeneous Graph-Based Dialog System," in IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 8, pp. 5212-5217, Aug. 2023.

[10] T. Liu et al., “Multichannel flexible pulse perception array for intelligent disease diagnosis system,” ACS Nano, vol. 17, no. 6, pp. 5673–5685, 2023.

[11] L. Ma and T. Yang, “Construction and evaluation of intelligent medical diagnosis model based on integrated deep neural network,” Comput. Intell. Neurosci., 2021.

[12] K. A. Tran et al., “Deep learning in cancer diagnosis, prognosis and treatment selection,” Genome Med., vol. 13, pp. 1–17, 2021.

[13] A. Coronato, M. Naeem, G. De Pietro, and G. Paragliola, “Reinforcement learning for intelligent healthcare applications: A survey,” Artif. Intell. Med., vol. 109, p. 101964, 2020.

[14] Z. Wei et al., “Task-oriented dialogue system for automatic diagnosis,” in Proc. 56th Annu. Meet. Assoc. Comput. Linguist., 2018, pp. 201–207.

[15] C. Yu, J. Liu, S. Nemati, and G. Yin, “Reinforcement learning in healthcare: A survey,” ACM Comput. Surv. (CSUR), vol. 55, no. 1, pp. 1–36, 2021.

[16] Y. E. Chen, K. F. Tang, Y. S. Peng, and E. Y. Chang, “Effective medical test suggestions using deep reinforcement learning,” arXiv preprint arXiv:1905.12916, 2019. [Unpublished].

[17] A. Tiwari, S. Saha, and P. Bhattacharyya, “A knowledge infused context driven dialogue agent for disease diagnosis using hierarchical reinforcement learning,” Knowledge-Based Systems, vol. 242, p. 108292, 2022.

[18] L. Yan, Y. Guan, H. Wang, Y. Lin and J. Jiang, "Efficient Evidence-Based Dialogue System for Medical Diagnosis," 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Istanbul, Turkiye, 2023, pp. 3406-3413.

[19] L. Yan et al., "EIRAD: An evidence-based dialogue system with highly interpretable reasoning path for automatic diagnosis," IEEE J. Biomed. Health Inform., vol. 28, no. 10, pp. 6141–6154, Oct. 2024.

[20] R.S. Sutton, and A.G. Barto, “Reinforcement Learning: An Introduction,” MIT Press, 2018.

[21] L. Baird, “Residual algorithms: Reinforcement learning with function approximation,” in Machine learning proceedings 1995, Morgan Kaufmann, pp. 30–37, 1995.

[22] G. Tesauro, “Temporal difference learning and TD-Gammon, Commun,” ACM, vol. 38, pp. 58–68, 1995.

[23] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, “Policy gradient methods for reinforcement learning with function approximation,” in Advances in neural information processing systems 12, pp. 1057–1063, 1999.

[24] J. Li et al, “Adversarial learning for neural dialogue generation,” in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2157-2169, 2017.

[25] J. S. Park et al., “Generative agents: Interactive simulacra of human behavior,” in Proc. 36th Annu. ACM Symp. User Interface Softw. Technol., Oct. 2023, pp. 1–22.

BlockMed: AI Driven HL7-FHIR Translation with Blockchain-Based Security

Yonis Gulzar[1]*, Faheem Ahmad Reegu[2]*, Abdoh Jabbari[3], Rahul Ganpatrao Sonkamble[4], Mohammad Shuaib Mir[5], Arjumand Bano Soomro[6]

Department of Management Information Systems, College of Business Administration, King Faisal University, Al-Ahsa 31982, Saudi Arabia[1, 5, 6]

Department of Electrical and Electronics Engineering, College of Engineering and Computer Science, Jazan University, Jazan 45142, Saudi Arabia[2, 3]

Pimpri Chinchwad University, Pune, Maharashtra 412106, India[4]

***Abstract*—Blockchain is a peer-to-peer (P2P) network that distributes information and protects data integrity, security, and privacy. Constant simplification is required for information exchange. This comprehensive assessment seamlessly integrates Electronic Health Record (EHRs) with blockchain technology. EHRs are represented with different standards mainly HL7 and FHIR. EHR should be interpreted to both parties after exchange. Such interpretation after exchange may face few interoperable challenges. To overcome EHR interoperability difficulties, 18 blockchain-based alternatives were examined. Despite their promise, these systems have a variety of drawbacks, including reliability, privacy, data integrity, and collaborative sharing. Six phases make up the systematic review: research, investigation, article curation, keyword abstraction, data distillation, and project trajectory monitoring. In total, 18 seminal articles on EHR interoperability and Blockchain integration were identified. Many unique interoperability methods are proposed for Blockchain-integrated EHR systems in these contributions. Several Blockchain applications, standards, and issues associated with EHR interoperability are described and analyzed. Implemented and proposed blockchain-based EHR frameworks are numerous. The security aspects have been covered, but standards compliance and interoperability requirements are lacking. Research in this area is needed. This research study has analyzed the different national and international EHR standards. This paper describes the current state of EHRs, including blockchain-based implementations, along with the interoperability issues between existing blockchain-based EHR frameworks. The research has proposed novel BlockMed framework which is interoperable for the HL7 and FHIR EHR standards. BlockMed framework is evaluated with Data Accuracy, Mapping Quality, Response Time, Latency, Interoperability Coverage, AI Model Efficiency, Consent and Security Management, Cross-Chain Support, Patient and Provider Satisfaction.**

***Keywords*—Blockchain; health care; electronic health records (EHRs); interoperability; and healthcare system**

I. INTRODUCTION

Blockchain technology, characterized by its decentralized nature, fast transaction processing, high security, and privacy features, can significantly shift the traditional healthcare system. Significantly, it becomes a crucial facilitator in upholding the confidentiality and security of patient data. The potential of this technology to significantly transform the transmission and storage of patient's electronic health records is emphasized by its implementation of advanced security measures for the secure transfer of medical data within the healthcare industry, utilizing a robust stochastic. Blockchain technology enhances the level of robustness and security in the field of electronic health records EHRs. The potential impact of this technology on the healthcare industry is a shift towards prioritizing the needs and preferences of patients, as well as enhancing the security, transparency, and compatibility of health data [1]. The potential for a significant shift in health information exchange (HIE) is evident through enhanced efficiency and security in electronic health records EHRs. These repositories contain detailed patient information, including diagnostic and treatment procedures. It is crucial to acknowledge that patient data holds significant value within the healthcare domain, as it serves as essential material for examining healthcare. Electronic health records EHRs serve as stores for highly sensitive medical data, which justifies their classification as repositories for valuable patient insights. The widespread availability of healthcare data plays a crucial role in advancing national healthcare initiatives and improving quality-of-service delivery [2]. The EHR ecosystem is the culmination of a comprehensive and meticulously organized collection of patient health data distributed throughout many healthcare institutions and governmental health authorities [3].

EHRs are digital databases that store a person's whole health history. Medical clinics, institutes, and experts collect and manage this database. Electronic health records and EHRs face many challenges with semantic interpretability. Electronic healthcare systems are prone to cyberattacks, making security a major issue. Cyber-attack victims in healthcare systems make up one-third of all documented cases. Given these conditions, Blockchain technology's ability to establish shared trust and disseminate data could improve collaborative healthcare decisions in telemedicine and precision medicine [4]. As shown by the "Anthem breach attack," which compromised 80 million people's data on February 4, 2015, 88% of attacks target healthcare systems. Rapid growth in electronic health records and EHR databases creates another issue. Patient data, X-ray images, and computed tomography scans comprise the enormous corpus, which demands lots of storage. The average healthcare facility had 665 terabytes of storage in 2015, but by 2020, it had 25,000 petabytes. This storage effort focuses on unstructured medical images. Healthcare systems' variability in database management systems, architectural configurations, and

*Corresponding Author, Email ID: ygulzar@kfu.edu.sa; freegu@jazanu.edu.sa

data infrastructures presents another problem. Thus, inter-provider health data transmission must prioritize data integrity and uniformity [5]. Diversity makes transmitting correct, standardized information difficult and hinders its efficient deployment in relevant circumstances. The study covers the following:

The value of blockchain-based interoperability for EHRs.

Potential snags in integrating blockchain technology into EHRs systems.

The necessity of a blockchain-based system that can communicate with other networks.

Brief overview of the needs, issues, and potential solutions for interoperable EHR.

This article is divided into five sections. Section II reviews the research that specifically relates to blockchain's potential application in healthcare. The systematic review's research methods and procedures are outlined in Section III, and the answers to the review's research questions are presented in Section IV. In Section V of the paper, the findings and implications are provided.

II. Related Work

A. Blockchain

Blockchain, a decentralized system, has revolutionized Internet information sharing. This technology, created for financial applications, eliminates the need for transaction middlemen, including trustworthy third parties in government and business. TTP trustworthiness and authenticity may be damaged by malfunctions or security breaches, threatening the transactional framework. Blockchain is a peer-to-peer database with nodes, contracts, and blocks. In blockchain technology, data is kept in blocks. These blocks store and record numerous data kinds. However, nodes help blockchain network participants communicate and engage. Nodes relay data, transactions, and other information between places. Every node in this framework includes a block of localized data. After a contract is signed, transactions are authenticated [6] [7].

B. Blockchain Technology and Electronic Health Records

Resilience of blockchain-based applications has offered adequate support for the medical sector and medical infrastructure. The literature study presents several tools for interoperability aware EHR maintenance. Because EHRs can communicate, healthcare providers can change patient records using distributed ledger technology. Blockchain-based electronic health records are a cutting-edge method of tracking patients' medical histories and appointments [6].

By handling health insurance claims and refreshing authentications in the payments system, interoperable EHRs streamline the insurance system. Integrating blockchain technology into EHRs would help medical researchers and developers of diagnostic tools and drugs make more accurate statistical estimates from patient data. The significance of Blockchain is demonstrated through analyses of related research, practical examples, and novel approaches to securing data [8].

Removing the middleman from a distributed healthcare system causes a significant disruption in the current healthcare models. Data integrity, security, and privacy can be achieved using blockchain in EHR-based healthcare systems, which are difficult challenges in conventional healthcare systems [9].

Samala et al. explores the integration of blockchain technology in healthcare, focusing on data security, EHR privacy, and patient ownership. The authors argue that traditional EHR systems often lack patient-controlled data access, which blockchain can address by providing decentralized and secure data management. However, the study also notes challenges in scalability and the need for standardized protocols to ensure interoperability across different healthcare systems. A blockchain-based framework utilizing smart contracts to enhance the security and interoperability of EHRs[10]. The framework aims to provide secure data sharing among healthcare providers while maintaining patient privacy. Despite its potential, the study acknowledges challenges related to the integration of existing EHR systems and the computational overhead associated with smart contract execution. De Novi et al. discussed the transformative potential of blockchain technology in healthcare, particularly in enhancing patient identity management and public health data interoperability[11]. The authors highlight the synergy between blockchain, artificial intelligence, and digital twins in creating a more secure and interoperable healthcare ecosystem. However, they also point out the need for regulatory frameworks and the challenges of integrating blockchain with existing healthcare infrastructures[12]. Agbo et al. examines various blockchain-based EHR management systems, evaluating their effectiveness in ensuring data security and interoperability. The study identifies common challenges such as scalability issues, high energy consumption, and the complexity of integrating blockchain with current EHR systems [13]. The authors suggest that future research should focus on developing lightweight blockchain solutions and establishing universal interoperability. Researchers introduced MedBlock, a blockchain-based framework designed to enhance the privacy and interoperability of EHRs. The framework employs a permissioned blockchain and hybrid on-chain/off-chain storage to balance transparency with confidentiality[14]. The study demonstrates MedBlock's ability to achieve high transaction throughput with low latency, though it acknowledges challenges related to cross-blockchain interoperability and the integration with existing EHR systems. Ouaguid et al. analyze various approaches to integrating blockchain into the e-healthcare ecosystem, focusing on data management, security, scalability, and interoperability. The study highlights the advantages of blockchain in providing secure and decentralized data management but also points out limitations such as the incomplete representation of major stakeholders in the blockchain network and the lack of regulatory flexibility to ensure legal interoperability by country. Bathula et al. explores the convergence of blockchain and artificial intelligence in healthcare, addressing challenges in securing EHRs, ensuring data privacy, and facilitating secure data transmission [15]. The study provides a comprehensive analysis of the adoption of these technologies within healthcare, spotlighting their role in fortifying security and transparency. However, it also discusses challenges like data security, privacy, and decentralized computing, forming a robust tripod.

Agbeyangi et al. investigates the implementation of blockchain technology, specifically Hyperledger Fabric, for EHR management at Frere Hospital in South Africa. The study examines the benefits and challenges of integrating blockchain into healthcare information systems, highlighting the role of blockchain in transforming healthcare[16]. The findings underscore the transformative potential of blockchain technology in healthcare settings, fostering trust, security, and efficiency in the management of sensitive patient data. Guo et al. presents hybrid blockchain-edge architecture for managing EHRs with attribute-based cryptographic mechanisms. The architecture introduces a novel attribute-based signature aggregation scheme and multi-authority attribute-based encryption integrated with Paillier homomorphic encryption to protect patients' anonymity and safeguard their EHRs. The study shows that the performance meets real-world scenarios' requirements while safeguarding EHR and is robust against unauthorized retrievals. Building upon these existing studies, the following section details our systematic research approach to analyzing interoperability in blockchain-based EHR systems [17] [18].

III. Research Methodology

The systematic review follows a structured approach consisting of six key stages to ensure a thorough analysis of blockchain interoperability in Electronic Health Records (EHRs). These stages include formulating a query, conducting research, selecting relevant articles, developing a keyword list, extracting data, and mapping. The process was designed to rigorously evaluate existing literature, categorize solutions, and identify challenges in blockchain-based EHR interoperability.

A. Research Question Formulation

The research process began with the formulation of a focused research question aimed at identifying obstacles and solutions for achieving blockchain interoperability in EHR systems. This research question was critical for guiding the literature search and ensuring relevance, as it enabled a targeted approach in understanding the issues associated with blockchain applications in healthcare and evaluating potential solutions. To ensure rigor and relevance, articles were selected based on predefined inclusion and exclusion criteria:

1) Inclusion Criteria

a) Publication date: Only articles published after 2019 were included to reflect the latest advancements in blockchain technology for healthcare.

b) Relevance: Articles specifically addressing blockchain in healthcare with an emphasis on EHR interoperability, security, or data sharing frameworks were prioritized.

c) Language and accessibility: Only English-language articles available through academic databases were considered to maintain consistency and accessibility.

d) Peer-reviewed sources: Preference was given to studies from reputable, peer-reviewed journals and conferences, including IEEE, PubMed, and ScienceDirect, to ensure quality.

2) Exclusion Criteria

a) Lack of focus: Articles that discussed blockchain technology broadly without a focus on healthcare or EHR interoperability were excluded.

b) Inadequate data: Studies with limited sample sizes or weak methodological frameworks were omitted to maintain high research quality.

c) Outdated or duplicate research: Articles presenting redundant information or findings duplicated in more recent studies were excluded to avoid redundancy.

Each selected article was evaluated based on these criteria, allowing for a curated collection of relevant and high-quality studies. The graphical representation of inclusion and exclusion criteria has been presented in Fig. 1.

3) Quality assessment

3) Quality assessment: A systematic quality assessment was performed in the selected articles to ensure reliability. Articles were evaluated based on several metrics:

a) Methodological rigor: Each article's methodology was reviewed for clarity and robustness, with particular attention to research design, data collection methods, and analytical techniques.

b) Data Sources and sample size: Studies that utilized reliable data sources and larger sample sizes were given preference, as these factors increase the generalizability and credibility of findings.

c) Credibility of findings: Each study's conclusions were analyzed for coherence with existing literature and evaluated for clarity, consistency, and validity, ensuring that all sources provided well-substantiated insights.

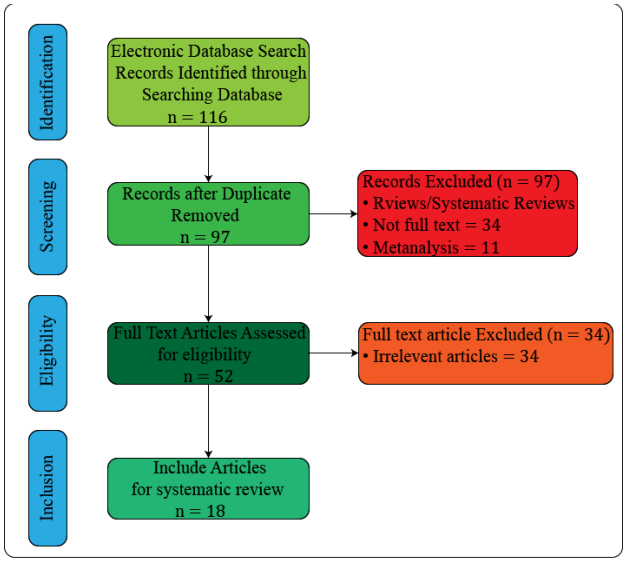


Fig. 1. Flowchart illustrating the criteria used to choose articles for systematic reviews.

4) Data extraction and mapping procedures

4) Data extraction and mapping procedures: Data extraction involved keyword-based searches for terms such as "blockchain," "EHR interoperability," "data privacy," and "healthcare security" across databases like IEEE, PubMed, and ScienceDirect. Using Atlas.ti software, the extracted data was mapped and organized, facilitating the thematic categorization

of research insights. This software-assisted approach streamlined the identification of common challenges, proposed solutions, and emerging trends in blockchain-based EHR interoperability. Atlas.ti further enhanced the efficiency of the data mapping process, allowing for a detailed comparison of various frameworks and solutions.

5) PRISMA flow diagram: To provide transparency in the article selection process, a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram was created. This diagram visualized each stage of the selection process, beginning with the total number of articles identified through database searches and progressing through the screening, inclusion, and exclusion phases. The flow diagram detailed the number of articles removed at each stage, along with reasons for exclusion (e.g., lack of relevance or inadequate data). By mapping the selection process, the PRISMA flow diagram enhanced transparency and ensured a systematic approach in curating the final articles included in the review.

6) Specifications and benchmarks: The review identified specific standards and benchmarks from the literature, highlighting the key interoperability requirements for blockchain-based EHR systems. Table I provides a summary of these benchmarks, including standards such as FHIR for data interchange, HL7 for cross-border data execution, and HIPAA for data privacy and security. These benchmarks were instrumental in categorizing the technical requirements necessary for a robust blockchain-based EHR system.

TABLE I. STANDARDS FOR DATA EXCHANGE

Ref	*Specifications*	*Benchmarks*
[19]	FHIR	Data Interchange
[13]	HL7	Transboundary execution
[20]	HITECH	Metamorphosis
[21]	PHR	Data transmission through APIs
[22]	Open EHRs	Protocol conformity
[23]	DICOM	Safety measures
[24]	SNOMED CT	Harmonious operability
[25]	CEN/ISO EN13606	Confidentiality and safeguarding
	HIPAA	Privacy and data uniformity

Bibliometric analysis was employed to deepen the understanding of blockchain-based EHR interoperability research. By using VOSviewer software, the frequency of author citations, keyword occurrences, and term co-occurrences within the literature was analyzed. The bibliometric analysis also identified collaborative relationships between authors and institutions from different countries, illustrating the global interest in blockchain technology for healthcare interoperability. This systematic methodology provides a solid foundation for evaluating the literature on blockchain-based EHR interoperability. By following a structured approach for article selection, quality assessment, and data mapping, this review ensures that only high-quality, relevant studies inform the analysis. The use of bibliometric analysis and standardized benchmarks further supports a comprehensive understanding of the current challenges and potential solutions in the field.

IV. RESULTS AND DISCUSSION

This section presents the systematic review's results. 116 research articles were pulled for the following search technique from multiple sources. After removing the duplicate records using the first selection procedure, 97 articles were chosen. The removed articles had nothing to do with how EHRs and blockchain's interacted. To streamline the selection process, the chosen articles underwent additional scrutiny. Implementing exclusion criteria resulted in removing 45 articles from the original 97 items. The papers eliminated during screening lacked full text and were therefore ineligible for the meta-analysis. 52 additional publications were checked to see if they qualified for the systematic review. 34 irrelevant items were eliminated from the list of chosen articles throughout the eligibility selection process. To complete the systematic review of Blockchain-based Electronic Health Records 18 publications must meet all the criteria.

These chosen articles are utilized to delineate the requirements and obstacles associated with achieving interoperable EHRs utilizing blockchain technology. The study's inquiries are outlined as follows:

Q1. How can Blockchain-based electronic health records (EHRs) be made interoperable?

Q2. In a Blockchain context, what are the interoperable standards for EHRs look like?

Q3. How does the blockchain-based framework enable EHR exchange between different hospitals using different EHR standards?

A. Q1. How can Blockchain-Based Electronic Health Records (EHRs) be made Interoperable?

The comprehensive literature analysis in this study found three degrees of blockchain-based EHR interoperability criteria. Protocols, standards, and the exchange and management of data across platforms are essential in the technology industry. Patient data privacy and security must be built into the legal and organizational interoperability norms for blockchain-based electronic health records (EHRs). The economic model and partnerships between public and private healthcare organizations are also considered.

The standards specify the variety of protocols for transmitting messages and health data. A solid business plan and knowledge-sharing system are needed to implement the approach at the organizational level. Data standards that ensure data integrity and adaptability are needed to share medical information efficiently. As shown in Table II, interoperability among EHR systems built on blockchain is essential for seamless data exchange. Blockchain-based EHRs offer advantages in filing exact health insurance claims due to their uniform implementation. Data mapping standards must be established and followed to efficiently move data among entities with different ownerships. Federal organizations can more efficiently manipulate health service provider data using logical models and computer language features. Blockchain systems need semantic consistency. Standardized coding procedures improve EHR security. Betek et al. identified blockchain-based EHR needs. These needs included reliable data gathering and

effective EHR-medical researcher information sharing. These requirements are aimed at improving healthcare system administration and reliability. Electronic Health Records (EHRs) compliance with Blockchain technology requires a framework for secure and structured encrypted communication among various systems. Additionally, stakeholders must be able to decrypt these messages. This can be done by setting message standards, values, and technological databases according to norms. Increased collaboration between manufacturers, corporations, researchers, and medical institutions can help meet privacy and security standards. A complete structure of legislation and norms enabling healthcare professionals and patients to communicate data is necessary to protect patient privacy. Blockchain technology secures insurance and incentives for public and private health service organizations, protecting data. To efficiently implement and use EHRs. When systems and organizations share health information, legal frameworks are needed to maintain data integrity. Systems and organizations must build legal frameworks to share information securely.

1) Semantic and technological requirements for blockchain-based EHR interoperability: Interoperable Blockchain-based Electronic Health Record (EHR) systems require specific semantic standards to ensure seamless communication across healthcare networks. Key semantic demands include common practices and methods for data exchange, ensuring data integrity remains unquestionable, and a dictionary of standardized data and communication protocols. Additionally, guidelines for the structured collection and exchange of information are crucial to maintain uniformity across systems. These semantic standards enable various EHR systems to "speak the same language," thus supporting accurate, meaningful data transfer and interpretation [26].

From a technological perspective, interoperable Blockchain-based EHRs must meet specific technical prerequisites to function effectively. These include standards for plug-and-play interoperability of services, permitted types of information and data formats, and data encoding specifications to secure both the production and transmission of data. Furthermore, protocols for safe data transmission are essential to protect sensitive health information during exchanges. Adhering to these technological standards ensures that Blockchain-based EHRs can securely and efficiently exchange data within an interoperable framework [27].

TABLE II. NEED FOR INTEROPERABILITY AMONGST EHR SYSTEMS BUILT ON THE BLOCKCHAIN

Requirements		
Conceptual (syntactic and semantic)	**Technological**	**Organizational and (Legal)**
Agreed vocabulary for messages and clinical documents. Common terminologies and information models for advanced messages	Signal, protocol, and technological plug-and-play compatibility	Fundamentals of doing business collaboration between companies to facilitate the exchange of information
For data accuracy, use standard terms	The seamless exchange of health data is essential for providing adequate healthcare	
To assess the discrepancies, data element mapping to the common terminology	Logical models developed without regard to platform or programming language limitations	Federally mandated program data reporting burden reduced
Integrity of meaning	Coding standard technical concerns need to be resolved	Protecting the confidentiality of medical records
Obtaining a shared dataset Collect the doctors' agreement on the dataset	Strong technical standards for sharing health information throughout institutions	An effective healthcare informatics group capable of handling all tasks Professionals in the healthcare industry coming together to reach an agreement on a particular project
Ability of structured message transmission between two or more systems The capacity to comprehend and use a sent message Creating a well-chosen vocabulary	Define data items, rules, values, and formats Agree on technical data models for database management systems	The collaboration of informaticists, vendors of EHRs, and clinicians in the industry
Information from the exchange Identification of healthcare professionals and patients	Reporting clinical data securely and in a timely manner	Partnerships between public-private entities and government incentives have been adopted more widely
There should be no ambiguity in the data for transmitting systems	Multiple systems exchange data to take action based on what they've learned	ensuring that organizations function under various legal regimes

2) Organizational and legal prerequisites for interoperable blockchain-based EHRs: The successful implementation of interoperable Blockchain-based EHRs also requires certain organizational and legal prerequisites. Collaboration between EHR vendors and healthcare providers is vital, as it enables the development of shared business models that facilitate information exchange. Furthermore, organizations must engage experts who specialize in maintaining the privacy of shared information, as privacy and security are central to handling health data on a Blockchain. Access to data related to insurance and incentive programs is also essential, as it helps create a supportive environment for interoperability while ensuring compliance with relevant policies and regulations. These organizational and legal prerequisites form the foundation for a sustainable and secure interoperable EHR ecosystem [28].

B. Q.2 In a Blockchain Context, What are the Interoperable Standards for EHRs look like?

Table III details the requirements for interoperability between various Blockchain-based EHR applications. The study's authors studied different methods for EHR sharing, security, and interoperability. Several norms are being implemented to make it easier for healthcare providers to employ solutions built on the Blockchain [29][30].

TABLE III. AN EHR INTEROPERABILITY STANDARD BASED ON BLOCKCHAIN TECHNOLOGY

Block-Chain Based Standard Available	Description
FHIR	Data attributes are contained in an HL7-based resource. Adherence to FHIR for Information Exchange standards
HL7	A developing standard based on FHIR. Robust operation on mobile devices.
HITECH	MIPS, HER certification, interoperability, and healthcare system transformation
PHR	HL7 was utilized for tethering and quick data exchange. Interchangeability of PHR and EHIR via APIs.
Open EHR	EHR is developed using open-source components. Clinical deployments that validate EHR standards.
DICOM	Secure transmission of health records and medical images. APIs for integrating various health systems.
SNOMED CT	EHRs' current clinical procedures. Consistency, interoperability, and accuracy.
CEN/ISO EN13606	Semantic guidelines for the exchange of EHR information. Standards of privacy and security for interface access.
HIPAA	Security requirements for patient data privacy. EHR interoperable system with confidence

When thinking about harmonization, it is essential to find the standard that is most compatible with other standards. This is demonstrated by Tables III and IV, further substantiating the superior interoperability qualities of HL7 and FHIR.

TABLE IV. CHARACTERISTICS OF BLOCKCHAIN-DRIVEN EHR STANDARDS

Properties	CEN-13606	CEN	Open-EHRs	HL7	HITECH	DICOM	FHIR
The Better-Workflow	Y	M	Y	Y	Y	M	Y
Reduced-Ambiguity	M	Y	Y	Y	M	Y	Y
Better Quality-of-Care	Y	M	M	Y	Y	Y	Y
In terms of Reliability	M	Y	Y	Y	M	M	Y
The Information-security	M	M	Y	Y	Y	M	Y
Security and the Privacy	M	Y	M	Y	Y	Y	Y

*M-Moderate, Y- Yes

C. Q3. How does the Blockchain-Based Framework Enable EHR Exchange Between Different Hospitals Using Different EHR Standards?

1) Interoperable blockchain-based EHR framework: BCIF-EHR: The BCIF-EHR framework is designed to enable seamless interoperability between healthcare providers, utilizing blockchain technology, HL7 and FHIR standards, and AI-driven data mapping to securely share EHR between hospitals. Below is a breakdown of each component in this framework and how they interact. The block diagram of EHR and standards in presented in Fig. 2.

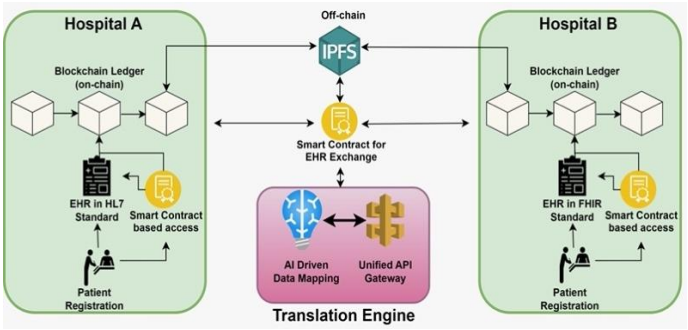


Fig. 2. Blockchain-based interoperable EHR system architecture (BCIF-EHR).

a) Patient registration: When a patient whose EHR system uses the HL7 standard, registers at Hospital A, their information (such as demographics and IDs) is captured in an

HL7 message format. This data flows through the AI-driven data mapping layer in the unified API gateway, which transforms the HL7 message into FHIR resources. This translation ensures that the patient's data can be compatible across different EHR systems.

b) Consent management using blockchain technology: Once the patient agrees to allow other hospitals to access their EHR, the blockchain-based consent management system records this consent immutably. Only authorized entities can access the data, with all data exchanges facilitated securely via HL7 and FHIR protocols.

c) Data exchange request from another hospital: When Hospital B requests access to the patient's medical history from Hospital A (using FHIR-based EHR), the unified API gateway handles the request. The gateway converts the request into an HL7-compliant query, and the AI-driven data mapping layer matches the necessary data fields, ensuring the request aligns with the HL7 structure.

d) Translation engine: The translation engine is an important component who is responsible for providing the interoperability between HL7 and FHIR. Interoperability between these standards will be by AI driven data mapping.

e) Interoperability platform for data processing and storage: After the request, Hospital A's HL7-based EHR sends the relevant data to a interoperability platform. This platform processes the data, ensuring it adheres to both HL7 and FHIR standards, stores it securely, and makes it accessible as FHIR resources.

f) Decentralized IPFS storage: For larger datasets, such as genomic data or medical images, decentralized storage through IPFS (Interplanetary File System) is used. Hospital B can retrieve this data through hash pointers in the FHIR resource, which link to the data stored in IPFS.

g) Data validation and smart contracts for EHR exchange: Before data is transmitted to Hospital B, a data validation process checks compliance with HL7 and FHIR standards. Smart contracts enforce the terms of data exchange, including patient consent, ensuring the conditions are met before the EHR data is shared.

h) Cross-chain support for multi-standard systems: If the patient visits a new healthcare provider that uses blockchain-enabled EHR, cross-chain support enables seamless data transfer across different blockchain networks. This component ensures interoperability across varied healthcare standards.

i) Patient data retrieval and update: When Hospital B receives the patient's data, it can retrieve and integrate it into its FHIR-based EHR system. Any updates made by Hospital B, such as test results, are converted back into HL7 format for Hospital A, ensuring both systems remain synchronized.

j) Secure logging and audit trail: Every transaction and data exchange is securely logged on the blockchain, creating an audit trail that medical professionals can monitor. This ensures data transparency, security, and traceability, reinforcing the trustworthiness of data sharing. Pseudo code of the Algorithm for Patient Registration (HL7 to FHIR) is presented below:

Pseudo code of PatientRegistrationHL7 to FHIR(HL7Message) Algorithm

Input:	***HL7Message***
Output:	***FHIRResource***
Step 1:	HL7Message ← CapturePatientRegistrationDetails()
Step 2:	APIService ← UnifiedAPIGateway(HL7Message)
Step 3:	MappingLayer ← AIDrivenDataMapping(APIService) HL7Segments ← ExtractSegments(HL7Message, ["PID", "OBX", "PV1"])
Step 4:	FHIRMapping ← MapHL7toFHIR(HL7Segments)
Step 5:	FHIRData ← TransformHL7toFHIR(FHIRMapping) FHIRResource ← CreateFHIRResource(FHIRData,
Step 6:	resourceType="Patient") isValid ← ValidateFHIRResource(FHIRResource) If isValid == False then RaiseError("FHIR resource validation failed")
Step 7:	Exit StoreFHIRResource(FHIRResource, InteroperabilityPlatform) ShareFHIRResource(FHIRResource, AuthorizedEntities) Return FHIRResource
	End Algorithm

D. Performance Evaluation Measures

1) Data accuracy and mapping quality: Data accuracy and mapping quality are fundamental for preserving the integrity of health information as it is transferred between healthcare systems, specifically when converting data between HL7 and FHIR standards. High data accuracy ensures that the information remains correct, up-to-date, and relevant, which is essential for providing reliable patient care and enabling informed decisions by healthcare professionals. Mapping quality assesses the effectiveness of data translation between these standards, maintaining platform integrity and reducing errors. The AI-driven data mapping enhances this quality by minimizing inaccuracies and data loss. As illustrated in Fig. 3, the mapping accuracy rate stands at an impressive 99.5%, with only a marginal inaccuracy of 0.2%, reflecting the robustness of data translation across these standards. Furthermore, the data loss rate remains low, averaging around 1.2%, which may be attributed to issues like unsupported fields or technical challenges that occasionally lead to minor data losses. Consistence in data translation is also high, with a consistency rate of 98.9% across multiple translation cycles, underscoring the system's ability to maintain stable data integrity throughout the mapping process.

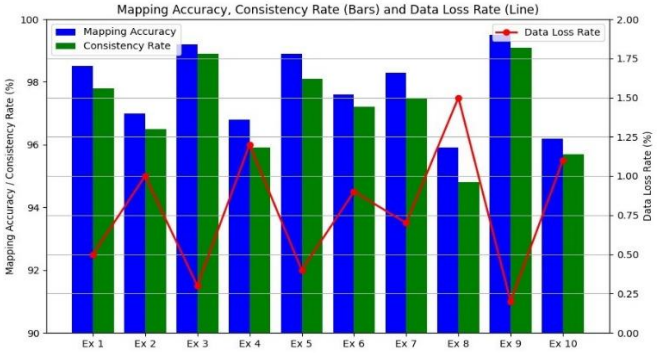


Fig. 3. Comparison of mapping accuracy, consistency rate, and data loss.

2) Response time and latency: Response time, particularly in healthcare environments, is critical for real-time access to

patient data. This measure is evaluated through both average latency and peak load latency, as shown in Fig. 4. Average latency is consistently around 150 milliseconds, which supports the need for rapid data access and retrieval, essential in emergency scenarios where every second counts. Under high transaction loads, the peak load latency reaches about 325 milliseconds, demonstrating the system's ability to perform efficiently even under heavy demand. This stability in latency ensures that the system remains responsive and reliable, which is vital for healthcare professionals who rely on timely data for patient care.

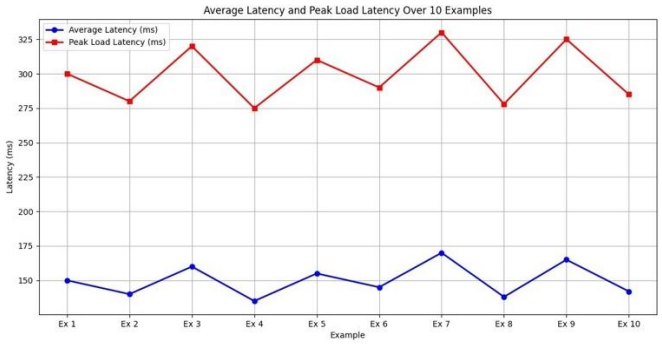


Fig. 4. Average latency and peak load latency across 10 samples.

3) Interoperability coverage: Interoperability coverage, depicted in Fig. 5, assesses the system's ability to handle various HL7 and FHIR standards, such as HL7 v2, v3, and FHIR versions like DSTU2, STU3, and R4. Effective interoperability is essential for consistent communication across diverse healthcare providers and systems, supporting seamless information exchange. The system achieves a high standard compliance rate, with close to 96% of HL7 and FHIR features supported. This high level of compliance ensures that the system adheres to established healthcare data standards, which is critical for effective data interoperability. Additionally, the system maintains a cross-version compatibility rate above 92%, indicating its flexibility to handle both older and newer versions of healthcare standards, a necessary feature for consistent data exchange across different platforms and systems.

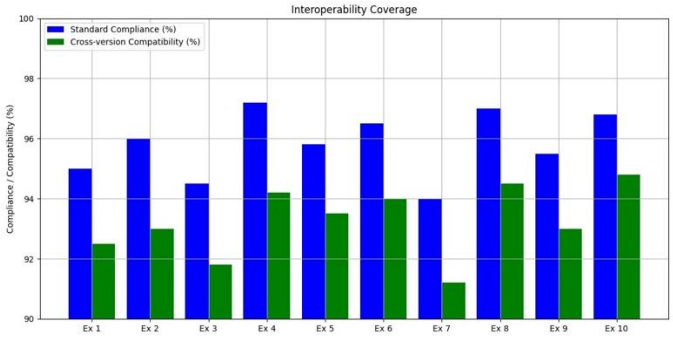


Fig. 5. Interoperabil*ity coverage*.

E. AI Model Efficiency

AI model efficiency plays a crucial role in system performance, affecting mapping speed, training time, and error rates. Fig. 6 provides insights into these metrics, where the AI model's mapping speed ranges from 330 to 370 milliseconds, enabling timely data processing in real-time healthcare environments. Training times for the AI model vary between 24 and 29 seconds, which supports quick model updates and adaptation to evolving data formats. The model error rate is low, around 2%, indicating high accuracy in data mapping with minimal errors. This low error rate is essential for maintaining data integrity across standards, ensuring that the AI-driven mapping layer performs reliably in a healthcare setting.

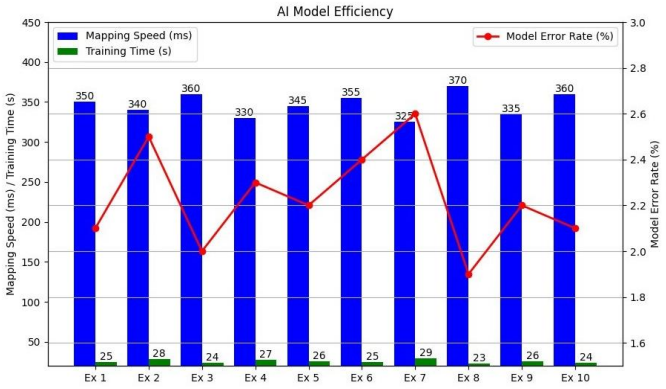


Fig. 6. AI model mapping speed and training time.

F. Consent and Security Management

Consent and security management are essential for safeguarding patient data during exchange.

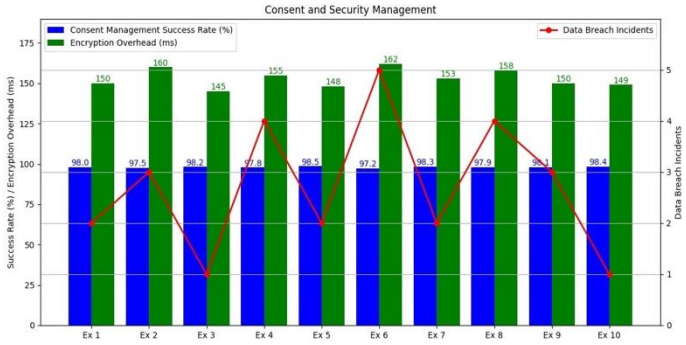


Fig. 7. Consent management success rate and encryption overhead.

Fig. 7 illustrates key metrics related to consent management success rate, encryption overhead, and data breach incidents. Consent management is consistently high, with success rates around 98%, ensuring that patient consent preferences are respected, and only authorized personnel can access sensitive data. Encryption overhead, which varies from 145 to 162 milliseconds, balances data security without causing significant delays, maintaining a secure and efficient data exchange process. The frequency of data breach incidents is minimal, indicating strong access control measures and robust security protocols that protect patient data and support compliance with privacy regulations.

G. Cross-Chain Support Evaluation

Cross-chain support evaluation, depicted in Fig. 8, enables seamless data exchanges across blockchain networks, which is essential for interoperability in multi-network environments.

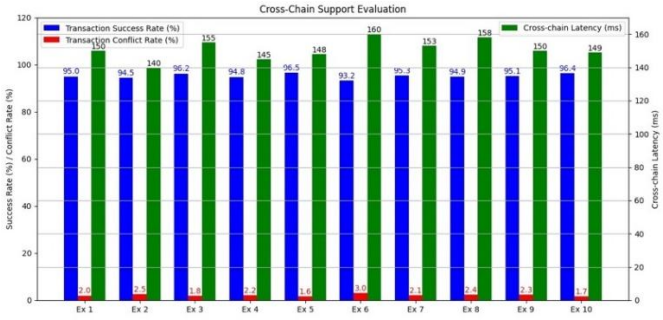


Fig. 8. Transaction success rate, conflict rate, and latency.

The system demonstrates a high cross-chain transaction success rate of approximately 95%, ensuring reliable interactions across different blockchain systems. The transaction conflict rate remains low, under 3%, indicating smooth data flow with minimal errors in cross-chain communication. Cross-chain latency averages around 150 milliseconds, allowing quick data access across blockchain networks and minimizing delays for healthcare providers requiring patient information from various sources.

H. Patient and Provider Satisfaction

Patient and provider satisfaction metrics, shown in Figure 9, provide insights into the system's usability, efficiency, and reliability from the perspective of end-users. High satisfaction scores indicate positive feedback from both patients and providers, reinforcing the system's effectiveness in real-world healthcare settings. Patient satisfaction scores range between 4 and 5 on a 5-point scale, reflecting a favorable view of the system's accessibility, data privacy, and ease of data retrieval. Provider satisfaction scores range from 3.5 to 4.5, suggesting that healthcare providers find the system beneficial for workflow efficiency, data accuracy, and accessibility, which enhances their productivity and decision-making capabilities.

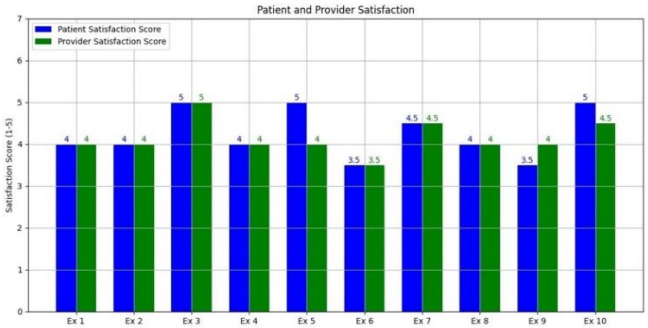


Fig. 9. Patient and service provider success score.

V. Conclusion and Future Work

This study demonstrates the potential of blockchain technology to address critical challenges in EHR interoperability by enhancing data security, privacy, and accessibility. The proposed blockchain-based framework BlockMed provides secure EHR exchange with translation of the standards i.e HL7 and FHIR using AI module. It is a decentralized approach to EHR management, ensuring compliance with contemporary standards and regulatory requirements like HIPAA while giving patients control over their own data. The integration of smart contracts further enhances the system by enforcing data sharing rules and maintaining accessibility without compromising privacy and data integrity. The BlockMed is proven to be secure and efficient after evaluation with the metrics i.e. Data Accuracy, Mapping Quality, Response Time, Latency, Interoperability Coverage, AI Model Efficiency, Consent and Security Management, Cross-Chain Support, Patient and Provider Satisfaction.

Despite these advancements, the real-world implementation of interoperable blockchain-based EHR systems remains limited, with few existing solutions that offer seamless data exchange across diverse healthcare platforms. This research contributes a foundational framework that can be expanded to develop scalable, interoperable EHR systems that meet the evolving needs of the healthcare industry. Future work should focus on real-world applications, addressing scalability issues, refining cross-chain support, and improving system performance under heavy data loads.

Overall, the study presents a promising path forward in the healthcare sector, leveraging blockchain to ensure secure, efficient, and interoperable EHR systems that can evolve with technological and regulatory developments. Further research and collaboration with healthcare providers and policymakers will be essential to fully realize the benefits of blockchain-based EHR interoperability. This study contributes a framework for blockchain-based EHR interoperability that adheres to HIPAA and HL7 standards, facilitating secure, cross-institutional patient data exchange. Future research should explore AI-driven data mapping to enhance translation accuracy between HL7 and FHIR standards and investigate cross-chain solutions to support data portability.

Acknowledgment

This work was supported by the Deanship of Scientific Research, the Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia under project KFU250655.

References

[1] F. A. Reegu, W. A. Bhat, A. Ahmad, and M. Z. Alam, "A review of importance of blockchain in IOT security," in AIP Conference Proceedings, 2023, vol. 2587, no. 1.

[2] A. A. Dar, F. A. Reegu, and G. Hussain, "Comprehensive Analysis of Enterprise Blockchain: Hyperledger Fabric/Corda/Quorom: Three Different Distributed Leger Technologies for Business BT - Mobile Radio Communications and 5G Networks," 2024, pp. 383–395.

[3] F. A. Reegu, S. Ayoub, A. A. Dar, G. Hussain, Y. Gulzar, and U. Fatima, "Building Trust: IoT Security and Blockchain Integration," in 2024 11th International Conference on Computing for Sustainable Global Development (INDIACom), 2024, pp. 1429–1434, doi: 10.23919/INDIACom61295.2024.10499070.

[4] A. A. Dar, F. A. Reegu, S. Ahmed, and G. Hussain, "Blockchain Technology and Artificial Intelligence based Integrated Framework for Sustainable Supply Chain Management System," in 2024 11th International Conference on Computing for Sustainable Global Development (INDIACom), 2024, pp. 1392–1397, doi: 10.23919/INDIACom61295.2024.10498149.

[5] A. A. Dar, F. A. Reegu, S. Ahmed, and G. Hussain, "Strategic Security Audit Protocol: Safeguarding Smart Home IoT Devices against Vulnerabilities," in 2024 11th International Conference on Computing for Sustainable Global Development (INDIACom), 2024, pp. 1386–1391, doi: 10.23919/INDIACom61295.2024.10498906.

[6] M. Z. Alam, F. Reegu, A. A. Dar, and W. A. Bhat, "Recent Privacy and Security Issues in Internet of Things Network Layer: A Systematic Review," Int. Conf. Sustain. Comput. Data Commun. Syst. ICSCDS 2022 - Proc., no. October, pp. 1025–1031, 2022, doi: 10.1109/ICSCDS53736.2022.9760927.

[7] A. A. Dar, M. Z. Alam, A. Ahmad, F. A. Reegu, and S. A. Rahin, "Blockchain Framework for Secure COVID-19 Pandemic Data Handling and Protection," Comput. Intell. Neurosci., vol. 2022, p. 7025485, 2022, doi: 10.1155/2022/7025485.

[8] X. Zhou, J. Liu, Q. Wu, and Z. Zhang, "Privacy Preservation for Outsourced Medical Data with Flexible Access Control," IEEE Access, vol. 6, pp. 14827–14841, 2018, doi: 10.1109/ACCESS.2018.2810243.

[9] M. English, S. Auer, and J. Domingue, "Block Chain Technologies & The Semantic Web : A Framework for Symbiotic Development," Comput. Sci. Conf. Univ. Bonn Students, pp. 47–61, 2016, doi: 10.1111/j.1364-3703.2010.00667.x.

[10] M. Al-Shabi and A. Al-Qarafi, "Improving blockchain security for the internet of things: challenges and solutions," Int. J. Electr. Comput. Eng., vol. 12, no. 5, pp. 5619–5629, 2022, doi: 10.11591/ijece.v12i5.pp5619-5629.

[11] M. Samaniego and R. Deters, "Blockchain as a Service for IoT," Proc. - 2016 IEEE Int. Conf. Internet Things; IEEE Green Comput. Commun. IEEE Cyber, Phys. Soc. Comput. IEEE Smart Data, iThings-GreenCom-CPSCom-Smart Data 2016, no. January 2020, pp. 433–436, 2017, doi: 10.1109/iThings-GreenCom-CPSCom-SmartData.2016.102.

[12] T. Alam, "Blockchain-Based Internet of Things: Review, Current Trends, Applications, and Future Challenges," Computers, vol. 12, no. 1, 2023, doi: 10.3390/computers12010006.

[13] G. Carter, H. Shahriar, and S. Sneha, "Blockchain-based interoperable electronic health record sharing framework," Proc. - Int. Comput. Softw. Appl. Conf., vol. 2, pp. 452–457, 2019, doi: 10.1109/COMPSAC.2019.10248.

[14] E. Lee, Y. Yoon, G. M. Lee, and T. W. Um, "Blockchain-based perfect sharing project platform based on the proof of atomicity consensus algorithm," Teh. Vjesn., vol. 27, no. 4, pp. 1244–1253, 2020, doi: 10.17559/TV-20200218052217.

[15] S. M. H. Bamakan, A. Motavali, and A. Babaei Bondarti, "A survey of blockchain consensus algorithms performance evaluation criteria," Expert Syst. Appl., vol. 154, 2020, doi: 10.1016/j.eswa.2020.113385.

[16] F. Bizzaro, M. Conti, and M. S. Pini, "Proof of Evolution: Leveraging blockchain mining for a cooperative execution of Genetic Algorithms," Proc. - 2020 IEEE Int. Conf. Blockchain, Blockchain 2020, pp. 450–455, 2020, doi: 10.1109/Blockchain50366.2020.00065.

[17] M. Du, Q. Chen, and X. Ma, "MBFT: A New Consensus Algorithm for Consortium Blockchain," IEEE Access, vol. 8, pp. 87665–87675, 2020, doi: 10.1109/ACCESS.2020.2993759.

[18] X. Fu, H. Wang, and P. Shi, "A survey of Blockchain consensus algorithms: mechanism, design and applications," Sci. China Inf. Sci., vol. 64, no. 2, pp. 1–15, 2021, doi: 10.1007/s11432-019-2790-1.

[19] F. M. Bublitz et al., "Disruptive technologies for environment and health research: An overview of artificial intelligence, blockchain, and internet of things," Int. J. Environ. Res. Public Health, vol. 16, no. 20, pp. 1–24, 2019, doi: 10.3390/ijerph16203847.

[20] G. G. Dagher, J. Mohler, M. Milojkovic, and P. B. Marella, "Ancile: Privacy-preserving Framework for Access Control and Interoperability of Electronic Health Records Using Blockchain Technology," 2017.

[21] C. Mcfarlane, M. Beer, J. Brown, and N. Prendergast, "Patientory: A Healthcare Peer-to-Peer EMR Storage Network v1.0," 2017.

[22] [A. Hossain, R. Quaresma, and H. Rahman, "Investigating factors influencing the physicians' adoption of electronic health record (EHR) in healthcare system of Bangladesh: An empirical study," Int. J. Inf. Manage., vol. 44, no. September 2018, pp. 76–87, 2019, doi: 10.1016/j.ijinfomgt.2018.09.016.

[23] F. A. Reegu et al., "Systematic Assessment of the Interoperability Requirements and Challenges of Secure Blockchain-Based Electronic Health Records," Secur. Commun. Networks, vol. 2022, 2022, doi: 10.1155/2022/1953723.

[24] R. Saripalle, C. Runyan, and M. Russell, "Using HL7 FHIR to achieve interoperability in patient health record," J. Biomed. Inform., vol. 94, p. 103188, Jun. 2019, doi: 10.1016/j.jbi.2019.103188.

[25] M. Farhadi, H. Haddad, and H. Shahriar, "Compliance Checking of Open Source EHR Applications for HIPAA and ONC Security and Privacy Requirements," pp. 704–713, 2019, doi: 10.1109/COMPSAC.2019.00106.

[26] Reegu Faheem, Zada Khan Wazir, Mohd Daud Salwani, Arshad Quratulain, and Armi Nasrullah, "A Reliable Public Safety Framework for Industrial Internet of Things (IIoT)," Proceeding - 2020 Int. Conf. Radar, Antenna, Microwave, Electron. Telecommun. ICRAMET 2020, pp. 189–193, Nov. 2020, doi: 10.1109/ICRAMET51080.2020.9298690.

[27] T. R. Vance and A. Vance, "Cybersecurity in the Blockchain Era," pp. 107–112, 2019.

[28] S. Niu, L. Chen, J. Wang, and F. Yu, "Electronic Health Record Sharing Scheme With Searchable Attribute-Based Encryption on Blockchain," IEEE Access, vol. 8, pp. 7195–7204, 2020, doi: 10.1109/ACCESS.2019.2959044.

[29] N. Andola, Raghav, S. Prakash, S. Venkatesan, and S. Verma, "SHEMB: A secure approach for healthcare management system using blockchain," 2019 IEEE Conf. Inf. Commun. Technol. CICT 2019, 2019, doi: 10.1109/CICT48419.2019.9066237.

[30] S. Ayoub, Y. Gulzar, F. A. Reegu, and S. Turaev, Generating Image Captions Using Bahdanau Attention Mechanism and Transfer Learning, Symmetry (Basel)., vol. 14, no. 12, 2022, doi: 10.3390/sym14122681.

Improving Air Quality Prediction Models for Banting: A Performance Evaluation of Lasso, mRMR, and ReliefF

Siti Khadijah Arafin[1], Suvodeep Mazumdar[2], Nurain Ibrahim[1, 3]*
School of Mathematical Sciences, College of Computing, Informatics and Mathematics,
Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia[1]
Information School, University of Sheffield, The Wave, 2 Whitham Road, Sheffield S10 2AH, United Kingdom[2]
Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Kompleks Al-Khawarizmi,
Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia[3]

***Abstract*—This study explores the effectiveness of various feature selection methods in forecasting next-day PM2.5 levels in Banting, Malaysia. The accurate prediction of PM2.5 concentrations is crucial for public health, enabling authorities to take timely actions to mitigate exposure to harmful pollutants. This study compares three feature selection methods: Lasso, mRMR, and ReliefF using a dataset consisting of 43,824 data points collected from Banting air quality monitoring stations (CA22B). The dataset includes ten variables, including pollutant concentrations such as O3, CO, NO2, SO2, PM10, and PM2.5, along with meteorological parameters such as temperature, humidity, wind direction and wind speed. The results revealed that Lasso outperformed both mRMR and ReliefF in terms of various performance metrics, including accuracy, sensitivity, precision, F1 score, and AUROC. Lasso demonstrated superior ability to handle multicollinearity, significantly improving the interpretability of the model by retaining only the most important variables. This suggests that the effectiveness of feature selection methods is highly dependent on the characteristics of the dataset, such as correlations among features. Thus, the top eight features to predict PM2.5 levels in Banting selected by Lasso method are relative humidity, PM2.5, wind direction, ambient temperature, PM10, NO2, wind speed, and O3. The findings from this study contribute to the growing body of knowledge on air quality prediction models, highlighting the importance of selecting the appropriate feature selection method to achieve the best model performance. Future research should explore the application of Lasso method in other geographical regions, including urban, suburban and rural areas, to assess the generalizability of the results.**

***Keywords*—PM2.5 concentration; feature selection; Lasso; mRmR; RBFNN; ReliefF**

I. Introduction

There is an increasing emphasis on air quality research, such as air quality prediction and the health effects of contaminants. This surge in attention is driven by rising pollution levels, growing health concerns, and advancements in technology that make monitoring and prediction more accessible. According to [1], aside from Kuala Selangor, all monitoring stations in the Klang Valley reported poor air quality days, with Banting having the second-highest occurrence at five days, eventhough Banting has less traffic compared to other urban area such as Shah Alam.

Prediction of PM2.5 levels (particulate matter with a diameter of less than 2.5 microns) has attracted considerable interest because of its significant impact on human health and its role as a key indicator of air quality. Previous studies such as in study [2] highlight the association between prenatal and postnatal PM2.5 exposure and a high incidence of tic disorders in children. Additionally, the study in [3] used machine learning techniques to predict PM2.5 concentrations using historical air quality data from Kuala Lumpur, demonstrating that advanced modelling approaches can significantly improve the accuracy of air quality predictions, which is critical for public health advisories and environmental management.

More recently, neural networks (a type of machine learning) have gained prominence in predicting air quality due to their ability to model complex, nonlinear relationships inherent in air quality data. One significant study by [4] on artificial neural networks (ANNs) in Peninsular Malaysia demonstrated their effectiveness in accurately predicting pollutants and the Air Pollutant Index. Radial Basis Function Neural Network (RBFNN), a machine learning method that has a simple architecture consisting of an input, hidden, and output layers, demonstrated faster convergence during training compared to more complex architectures like Multi-Layer Perceptrons (MLP). Moreover, the accuracy of the RBFNN model in predicting the occurrence of haloketones in tap water, as shown in [5], exhibited high performance, effectively capturing the complex relationships between input parameters and the predicted outcomes.

To further improve the accuracy of PM2.5 prediction models, researchers have increasingly focused on feature selection methods. According to a study by [6], the findings indicated that feature selection methods improved prediction accuracy by at least 13.7% compared to models that did not employ feature selection. ReliefF is a filter-based feature selection method that is effectively used to train models for classification due to its ability to identify relevant features by ranking them based on their capacity to distinguish between instances [7]. In addition, the study in [8] noted that many

previous studies address multicollinearity issues in feature selection methods by removing redundant and irrelevant features from high-dimensional data, which can be effective in preventing deterioration in model performance. However, removing redundant features based solely on the correlation between variables might not provide accurate predictions, as the removed variables may have unique characteristics.

Therefore, a feature selection method that focuses on the correlation between variables is needed. For example, the study in [9] discussed using correlation-embedded attention modules to reduce multicollinearity can improve the model performance and interpretability. Minimum Redundancy Maximum Relevance (mRMR) is a filter-based, correlation-based feature selection method that focuses on selecting features that maximize relevance to the target variable while minimizing redundancy among them, thereby extracting the most representative features closely related to the target variable [10]. In addition, Least Absolute Shrinkage and Selection Operator (Lasso) is another feature selection method that effectively addresses multicollinearity by applying regularization to shrink less important feature coefficients to zero, thus enhancing model interpretability and reducing overfitting [11].

There are numerous feature selection techniques, such as Recursive Feature Elimination (RFE) and XGBoost Feature Importance. However, our study focuses on Lasso, mRMR, and ReliefF due to their distinct selection mechanisms and prior applications in air quality prediction research. The research in [12] points out that while RFE can achieve high performance, it suffers from computational inefficiency due to its reliance on iterative model training. This issue is particularly relevant when dealing with large datasets, as the time required for multiple iterations can become prohibitive. Meanwhile, XGBoost Feature Importance ranks features based on their contribution to decision trees but does not explicitly account for multicollinearity. Unlike Lasso, which penalizes correlated predictors to improve interpretability, XGBoost may distribute importance among correlated variables, making it less effective for identifying the most relevant individual predictors in highly correlated datasets.

Additionally, a study by [13] compared the ReliefF, mRMR, and Lasso methods using air quality data from Shah Alam. The study concluded that ReliefF outperformed the other two methods and recommended that future researchers compare these three feature selection methods in other urban air quality datasets. Based on this recommendation, our study applies these methods to air quality data from Banting, providing insights into how these methods perform in a different urban setting with distinct environmental and meteorological conditions. As an industrial area, Shah Alam is influenced by emissions from factories, vehicles, and urban activities, leading to higher concentrations of pollutants such as PM2.5, PM10, SO2, NO2, and CO. In contrast, Banting, an agricultural area, experiences pollution primarily from agriculture activities and occasional biomass burning, which may result in different pollutant levels and patterns compared to Shah Alam.

Therefore, this study will compare the performance of the RBFNN model combined with ReliefF, mRMR, and Lasso feature selection methods to determine the most effective approach to predict the next day concentration of PM2.5 in Banting. The study's findings will benefit policymakers and other relevant parties by providing evidence-based insights into the most effective feature selection methods for improving air quality prediction models.

II. Research Methods

A. Dataset

This study used the Banting air quality dataset provided by the Department of Environment Malaysia (DOE). Ten variables (PM2.5, PM10, SO2, NO2, O3, CO, wind direction, wind speed, relative humidity, and ambient temperature) used as independent variables, were extracted from hourly data spanning four years (2018 to 2022). However, there are missing values in the dataset as shown in Table I, with NO2 having the highest percentage of missing data at 8.68%, while relative humidity is the lowest at 1.39%. In order to address the missing data points, we employed linear interpolation method, as was suggested in study [14].

TABLE I. Percentage of Missing Values

Variable	N	Missing Value
PM2.5	43207	617 (1.41%)
PM10	43075	749 (1.71%)
SO2	40938	2886 (6.59%)
NO2	40021	3803 (8.68%)
O3	41182	2642 (6.03%)
CO	40735	3089 (7.05%)
WD	43135	689 (1.57%)
WS	42927	897 (2.05%)
Humidity	43217	607 (1.39%)
Temperature	42443	1381 (3.15%)
$PM2.5_{D+1}$	43207	617 (1.41%)

The target variable in this study is the PM2.5 levels of the next day. Hence, the hourly dataset was transformed to daily categorical data by averaging values over 24 hours. The PM2.5 breakpoint of air quality categories is shown in Table II based on guidelines by DOE. Furthermore, this study employed classification task, hence the target variable, $PM2.5_{D+1}$ was transformed into binary classification system, not polluted (0) and polluted (1) as suggested in [15]. The "good" and "moderate" category represent not polluted (0) class, while other than that are represent polluted (1) class [15].

TABLE II. Labels for the Respective PM2.5 Breakpoint and AQI Categories

AQI Category	PM2.5 Breakpoints
Good	0.0-12.0
Moderate	12.1-35.4
Unhealthy for Sensitive Groups	35.5-55.4
Unhealthy	55.5-150.4
Very Unhealthy	150.5-250.4
Hazardous	250.5 and above

B. Research Framework

Research framework of this study is shown in Fig. 1. The process begins with data extraction as mentioned previously. Then, followed by extensive data pre-processing steps which include imputation using linear interpolation, converting hourly data into daily averages, binary categorization of PM2.5 levels, min-max normalization, and balancing the dataset with SMOTE technique. Subsequently, the three feature selection methods: ReliefF, mRMR, and Lasso are applied to rank and select the top eight variables most relevant to PM2.5 prediction as suggested in study [13] and study [16]. The selected features are used to train a Radial Basis Function Neural Network (RBFNN). A comparative evaluation of accuracy, specificity, precision, F1 score, and AUROC was finally used to identify the most effective model.

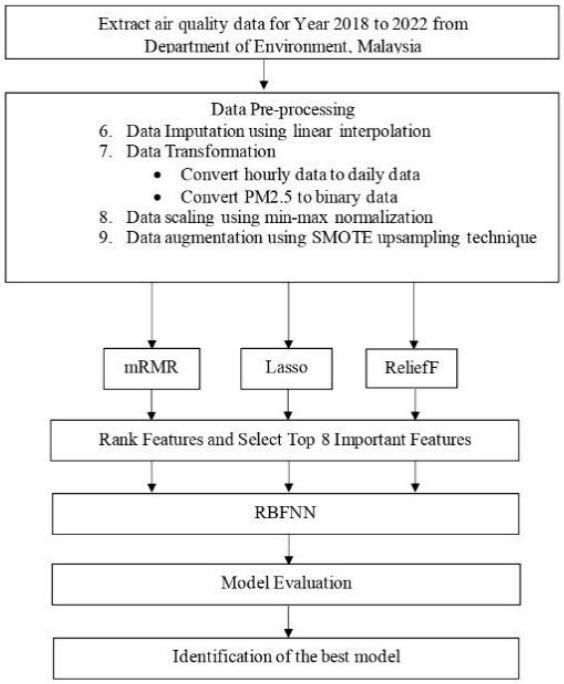


Fig. 1. Research flowchart.

C. Feature Selection Method

1) ReliefF: The ReliefF algorithm is a filter-based feature selection method known for its effectiveness in high-dimensional and noisy datasets [17]. It evaluates feature importance by assessing their ability to distinguish between instances of different classes. Unlike traditional methods relying on statistical correlations, ReliefF employs a distance-based approach, sampling data points and identifying the k-nearest neighbors within the same class (nearest hits) and from other classes (nearest misses). By comparing feature values between these neighbors, ReliefF assigns higher importance to features with substantial variation across classes and minimal variation within the same class.

The weight of a feature W[A] is updated iteratively using the formula in Eq. (1) [18], where H_i is the nearest hit (same class), while M_i is the nearest miss (different class). Moreover, the difference is calculate as shown in Eq. (2), where $diff(A, X, Y)$ are the difference in feature A values between instances X and Y.

$$W[A] = W[A] - \frac{1}{m}\sum_{i=1}^{m}(diff(A, H_i) - diff(A, M_i)) \quad (1)$$

$$diff(A, X, Y) = \begin{cases} 0 \; if \; X[A] = Y[A] \\ 1 \; if \; X[A] \neq Y[A] \end{cases} \quad (2)$$

2) Maximum Relevance Minimum Redundancy (mRMR): Maximum Relevance Minimum Redundancy (mRMR) is a feature selection method that aims to choose the most relevant features while minimizing redundancy among them. It is particularly useful in high-dimensional datasets where feature selection is crucial for improving model performance. The approach maximizes the relevance of selected features to the target variable and minimizes the redundancy between them. The relevance of a feature x_i to the target variable, c are calculated using mutual information, while redundancy between features is determined by the pairwise mutual information between features x_i and x_j. Eq. (3) and Eq. (4) shows the formulas to calculate maximum relevance and minimum redundancy, respectively.

$$maxD(S, c), D = \frac{1}{|s|}\sum_{x_i \in s} I(x_i, c) \quad (3)$$

$$minR(S), R = \frac{1}{|s|^2}\sum_{x_i, x_j \in s} I(x_i; x_j) \quad (4)$$

3) Lasso: The Least Absolute Shrinkage and Selection Operator (Lasso) is a method that helps improve model interpretability and performance by performing feature selection and regularization. It is particularly effective when working with datasets that contain many features, as it can reduce overfitting by penalizing less relevant variables [19]. Lasso works by adding a penalty term to the loss function, specifically the L1 penalty, which forces some of the feature coefficients to become exactly zero [19]. The L1 penalty, controlled by a tuning parameter λ, directly influences how many coefficients are driven to zero, with larger values of λ resulting in more features being eliminated. This regularization technique ensures that the model is both efficient and less likely to overfit the data. The mathematical formulation of Lasso is represented in Eq. (5). Where the first term represents the residual sum of squares, and the second term is the L1 penalty.

$$\hat{\beta} = min\,\beta\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \quad (5)$$

4) Radial basis function neural network: Radial Basis Function Neural Networks (RBFNN) are a specialized type of artificial neural network that utilize radial basis functions for activation. These networks are well-suited for tasks such as function approximation, classification, and regression, owing to

their capacity to model involved nonlinear relationships. RBFNN typically consist of three essential layers: the input layer, the hidden layer, and the output layer, with weights connecting each layer.

The input layer receives the source node, representing the independent variable, and links it to the surrounding network. The hidden layer then applies a nonlinear transformation, mapping the input space to a higher-dimensional hidden space. The output layer generates the final predicted result based on the transformed inputs from the hidden layer. In the hidden layer, each unit corresponds to a transfer function, often a Gaussian function. The radial basis function (RBF), which has a symmetric shape, acts as the transfer function in this case. The number of hidden units is directly related to the number of RBF employed in the network. The Gaussian RBF is mathematically expressed as shown in Eq. (6), meanwhile the output is computed by summing the weighted contributions of the RBF as shown in Eq. (7).

$$\emptyset(x) = \exp(-\frac{||x - c||^2}{2\sigma^2}) \tag{6}$$

$$g(X) = \sum_{j=1}^{k} w_j \emptyset_j(r\|X - C_j\|) \tag{7}$$

where w_j are the weights assigned to each radial basis function, C_j represents the centers of the RBF, and r is a scaling factor. In the output layer, a logistic (sigmoid) activation function is commonly used for binary classification. This function transforms the weighted sum of the hidden layer outputs into a probability between 0 and 1, and is defined as in Eq. (8). Thus, the final output of the RBFNN for binary classification is calculated using Eq. (9), where w_0 is the bias term.

$$\sigma(\mathrm{z}) = \frac{1}{1 + e^{-\mathrm{z}}} \tag{8}$$

$$\hat{y} = \sigma\left(w_0 + \sum_{j=1}^{k} w_j \emptyset_j(r\|X - C_j\|)\right) \tag{9}$$

D. Model Performances

The developed model's performance will be evaluated using six metrics: accuracy, sensitivity, specificity, precision, F1 score, and AUROC. To evaluate the classification performance of the developed model, a confusion matrix is first constructed. The confusion matrix, shown in Table III provides a detailed summary of the model's predictions by categorizing them into four groups: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Meanwhile, the total number of instances is the sum of all four categories: TP, TN, FP, and FN.

TABLE III. CONFUSION MATRIX

	Actual Positive (1)	**Actual Negative (0)**
Predicted Positive (1)	TP	FP
Predictive Negative (0)	FN	TN

Accuracy reflects the percentage of correct predictions and is calculated as shown in Eq. (10). Sensitivity, or the true positive rate, measures the proportion of actual positive cases correctly identified in Eq. (11), while specificity, or the true negative rate, quantifies the proportion of actual negatives accurately calculated using in Eq. (12).

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{10}$$

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{11}$$

$$Specificity = \frac{TN}{(TN + FP)} \tag{12}$$

Precision determines the percentage of correct positive predictions among all predicted positives in Eq. (13). The F1 score, a harmonic mean of precision and sensitivity in Eq. (14), assesses the model's ability to maintain balance between these metrics.

$$Precision = \frac{TP}{(TP + FP)} \tag{13}$$

$$F1\ Score = \frac{2\ x\ Precision\ x\ Sensitivity}{(Precision + Sensitivity)} \tag{14}$$

The Area Under the ROC Curve (AUROC) curve illustrates the relationship between sensitivity and the False Positive Rate, with AUROC measuring the model's ability to differentiate between classes. A higher AUROC value, nearing 1.0, signifies better classification performance. Collectively, high scores across these metrics indicate a reliable and effective model.

III. RESULTS AND DISCUSSION

This section discusses the results of the study. Table IV summarizes the descriptive statistics of the independent variables, revealing a substantial range in standard deviations, from 0.001 to 38.15. This wide variability reflects the diverse scales of the variables, necessitating data normalization. Based on Fig. 2, the histogram reveals PM2.5, PM10, and CO have higher variability, suggesting potential impact from external factors. The distribution of these variables over the four years may have been affected by the COVID-19 period, with changes in human activity significantly impacting pollution levels during this time.

To address this, min-max normalization was employed to transform all variables to a consistent scale, ensuring comparability across features. Min-Max normalization is one of the most commonly used normalization methods in various applications, including air quality datasets [20]. Additionally, Fig. 3 illustrates the distribution of the $PM2.5_{D+1}$ category, highlighting a significant class imbalance in the dataset, where one category is disproportionately represented. Such imbalances can bias predictive models, undermining their reliability and generalizability. To resolve this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was applied, a proven method for addressing class imbalance by generating synthetic samples for underrepresented categories. By balancing the dataset, SMOTE enhances the model's ability to learn from all

categories effectively, thereby improving prediction accuracy and ensuring robust, reliable outcomes.

TABLE IV. DESCRIPTIVE STATISTICS OF BEFORE DATA PRE-PROCESSING

Variable	N	Mean	Median	Std. Dev.	Skewness
PM2.5	1825	21.236	18.648	12.157	3.653
PM10	1825	30.433	28.016	14.137	3.165
SO2	1825	0.001	0.001	0.001	1.377
NO2	1825	0.009	0.009	0.003	0.335
O3	1825	0.019	0.018	0.006	0.881
CO	1825	0.596	0.575	0.201	0.883
WD	1825	163.226	159.162	38.15	0.787
WS	1825	1.026	0.985	0.329	1.967
Humidity	1825	83.366	83.827	6.023	-2.678
Temperature	1825	27.123	27.138	1.14	-0.265

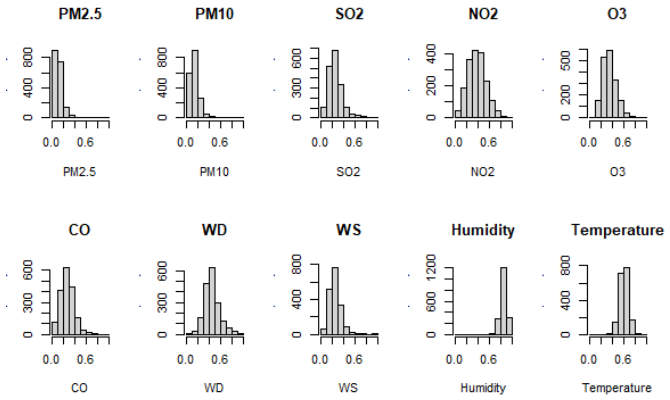


Fig. 2. $PM2.5_{Dt1}$ Distribution of variables.

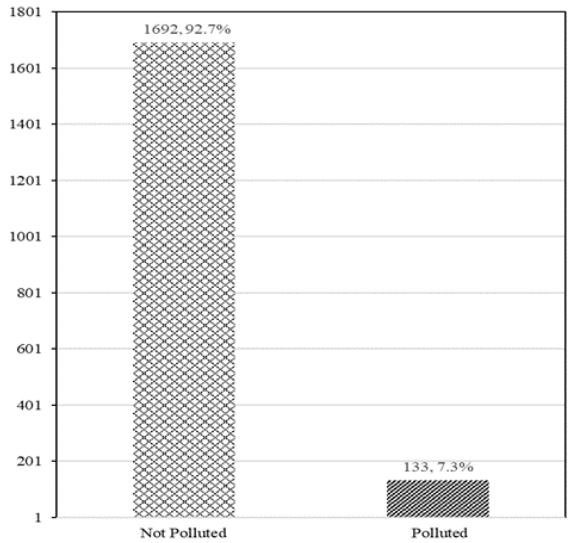


Fig. 3. $PM2.5_{Dt1}$ Distribution of (Before SMOTE).

Table V presents the descriptive statistics after data preprocessing. The application of min-max normalization successfully scaled the data, as evidenced by all mean and median values now falling within the standard range of 0 to 1. The data scaling ensures that the feature scales are consistent, which makes model training more successful. Furthermore, the skewness values have moved closer to zero, indicating a more symmetric and balanced distribution throughout the sample.

TABLE V. DESCRIPTIVE STATISTICS OF AFTER DATA PRE-PROCESSING

Variable	N	Mean	Median	Std. Dev
PM2.5	3288	0.172	0.136	0.143
PM10	3288	0.201	0.167	0.152
SO2	3288	0.278	0.255	0.137
NO2	3288	0.433	0.435	0.17
O3	3288	0.359	0.34	0.123
CO	3288	0.319	0.307	0.139
WD	3288	0.431	0.418	0.122
WS	3288	0.24	0.229	0.094
Humidity	3288	0.832	0.834	0.06
Temperature	3288	0.62	0.626	0.076

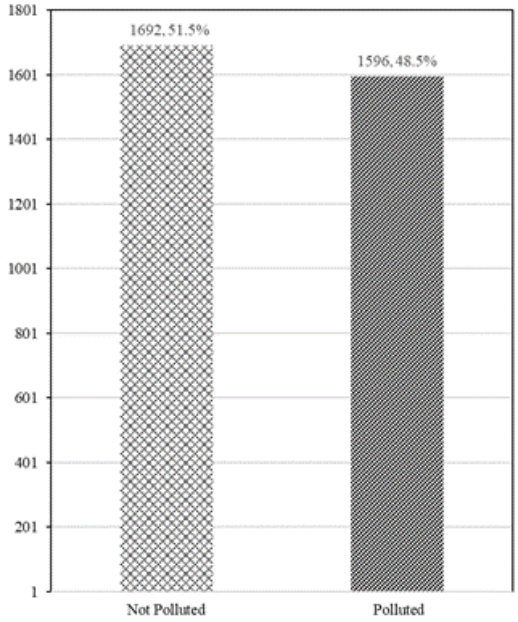


Fig. 4. $PM2.5_{Dt1}$ Distribution of (After SMOTE).

Fig. 4 illustrates the distribution of the $PM2.5_{D+1}$ category following the application of SMOTE for upsampling. This technique effectively addressed the class imbalance, resulting in a nearly equal representation of both groups: 1692 samples (51.5%) classified as "not polluted" and 1596 samples (48.5%) as "polluted." This balance dataset can enhance the reliability of the dataset for subsequent analyses and model predictions [21].

Table VI ranks independent variables according to three feature selection methods: Lasso, mRMR, and ReliefF, for predicting next-day PM2.5 levels. The rankings of independent variables by Lasso, mRMR, and ReliefF reveal both similarities and key differences. Lasso identifies relative humidity as the most important variable, emphasizing its critical role in PM2.5 prediction. In contrast, mRMR assigns the top 1 rank is SO2,

highlighting its unique contribution to the dataset. ReliefF, however, ranks NO2 as the most critical feature, suggesting its strong influence on pollutant dispersion and PM2.5 formation. These differences in top-ranking features underscore the variability in feature importance prioritization across the methods. Furthermore, the least important variables and exclusions also vary. CO is consistently ranked eighth by both Lasso and mRMR. However, ReliefF determine SO2 as the least important feature to predict PM2.5 in Banting [22].

Additionally, certain features were not selected by the three methods. Both particulate matter (PM10 and PM2.5) were excluded by mRMR method. Additionally, Lasso did not select SO2 and O3 as important feature. On other hand, ReliefF exclude both relative humidity and ambient temperature from the model.

TABLE VI. FEATURE RANKING ACROSS LASSO, MRMR, AND RELIEFF METHODS

Variable	Lasso	mRMR	reliefF
PM2.5	2	-	7
PM10	5	-	6
SO2	-	1	8
NO2	6	6	1
O3	-	4	4
CO	8	8	2
WD	3	7	3
WS	7	2	5
Humidity	1	5	-
Temperature	4	3	-

Next, the three feature selection methods will be evaluated using the performance of the RBFNN model to predict $PM2.5_{D+1}$ in Banting is shown in Table VII. According to the table, the Lasso method yields higher values in accuracy (0.771), sensitivity (0.765), precision (0.778), F1 score (0.771), and AUROC (0.769), when compared to the mRMR and ReliefF methods. These findings contradict the results of [13], which concluded that ReliefF outperforms Lasso. The discrepancy arises from the contrasting characteristics of the study areas. Shah Alam, as an industrial area, is heavily influenced by traffic and industrial emissions, resulting in more consistent pollution sources throughout the year. In contrast, Banting, an agricultural region, experiences pollution patterns influenced by seasonal meteorological factors. For instance, according to a study in [23], during the northeast monsoon, precipitation in Banting have influence on air pollution levels. This difference in pollution patterns might explain why Lasso outperformed ReliefF in Banting, as ReliefF did not select two key meteorological factors, humidity and temperature, which are more significant in this region.

However, this study aligns with [23], which found that the Lasso algorithm enhances model performance more effectively than the ReliefF method. According to study [19], Lasso is known for its ability to handle multicollinearity in predictors, reducing the impact of correlated features, and enhancing model interpretability by retaining only significant variables. This explains why Lasso performs better than the ReliefF method. Moreover, the mRMR method also demonstrates better performance compared to ReliefF. This may be attributed to the high correlations among variables in the Banting air quality dataset.

TABLE VII. COMPARISON MODEL PERFORMANCE

Model	Lasso	mRMR	ReliefF
Accuracy	**0.771**	0.725	0.568
Sensitivity	**0.731**	0.623	0.544
Specificity	0.807	**0.819**	0.591
Precision	**0.778**	0.761	0.551
F1 Score	**0.754**	0.685	0.548
AUROC	**0.829**	0.777	0.608

Table VIII shows the confusion matrix provides a detailed breakdown of the Lasso-based RBFNN model's classification performance in predicting PM2.5 levels. In this case, the model correctly classified 231 polluted instances (TP) and 276 non-polluted instances (TN)**,** while misclassifying 66 non-polluted instances as polluted (FP) and 85 polluted instances as non-polluted (FN)**.** The total number of instances used for evaluation is 658, calculated as the sum of TP, TN, FP, and FN.

TABLE VIII. CONFUSION MATRIX

	Actual Positive (1)	Actual Negative (0)
Predicted Positive (1)	231	66
Predictive Negative (0)	85	276

IV. CONCLUSION

In conclusion, this study highlights the significant role of feature selection methods in enhancing the predictive performance of air quality models in Banting, Malaysia. Among the three methods evaluated, Lasso emerged as the most effective for predicting next-day PM2.5 levels, consistently outperforming both mRMR and ReliefF in key performance metrics such as accuracy, sensitivity, precision, F1 score, and AUROC. Thus, Thus, the top eight features to predict PM2.5 levels in Banting selected by Lasso method is relative humidity, PM2.5, wind direction, ambient temperature, PM10, NO2, wind speed, and O3. While previous studies have recognized the strengths of ReliefF in detecting relevant features, this research reinforces the advantages of Lasso, particularly in its ability to improve model performance by addressing feature redundancy and focusing on the most impactful variables.

The superior performance of Lasso can be attributed to its ability to handle multicollinearity, reduce the impact of correlated features, and enhance model interpretability by retaining only the most significant predictors. Given the success of Lasso, future research should further explore its application in different settings, including suburban and rural areas, to evaluate its generalizability across diverse environments, especially dataset that contains high correlation between variables.

This study uses data from 2018 to 2022, a period that includes the COVID-19 years. During this time, air quality is likely to have been positively affected due to lower traffic and congestion, which may have influenced the results. Future studies might compare the results by differentiating between the COVID-19 and non-COVID-19 years to better understand how these factors impact air quality predictions. The results from this study are specific to the Banting dataset and may not be directly applicable to other countries due to regional differences in air quality patterns. This study benefits researchers, policymakers, public health authorities, and technology developers by improving air quality prediction models. Policymakers can use these predictions to implement real-time air quality monitoring systems, establish dynamic traffic control measures to reduce vehicular emissions during high-pollution periods, and enforce stricter industrial regulations to limit pollutant discharge. Additionally, predictions can guide the development of targeted public health advisories, such as issuing alerts for vulnerable populations and adjusting outdoor activity recommendations during hazardous air quality events. This study also supports NGOs in advocating for better air quality regulations and raising public awareness. Moreover, future researchers can apply the methods explored in this study to other areas, such as water quality prediction or environmental monitoring in different ecosystems. By adapting the feature selection techniques to new domains, researchers can validate the approach's versatility and effectiveness in improving prediction models across various environmental factors.

ACKNOWLEDGMENT

The authors would like to acknowledge the Ministry of Higher Education (MOHE) for support this work under the Fundamental Research Grant Scheme (FRGS) (FRGS/1/2023/STG06/UITM/02/8). This research was financially supported by Institute of Postgraduate Studies Universiti Teknologi MARA (UiTM).

REFERENCES

[1] Ministry of Environment and Water Malaysia. (2021). Laporan Kualiti Alam Sekeliling 2021 [Environmental Quality Report 2021]. Department of Environment, Malaysia. https://enviro2.doe.gov.my/ekmc/wp-content/uploads/2022/10/Laporan-Kualiti-Alam-Sekeliling-2021.pdf

[2] Chang, Y., Jung, C., Chang, Y., Chuang, B., Chen, M., & Hwang, B. (2022). Prenatal and postnatal exposure to PM2.5 and the risk of tic disorders. Paediatric and Perinatal Epidemiology, 37(3), 191–200. https://doi.org/10.1111/ppe.12943

[3] Zaini, N., Ean, L. W., Ahmed, A. N., Malek, M. A., & Chow, M. F. (2022). PM2.5 forecasting for an urban area based on deep learning and decomposition method. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-21769-1

[4] Shafi, M. S. M., & Juahir, H. (2024). Forecasting Air Quality in Peninsular Malaysia: Unveiling the Power of Artificial Neural Networks.

[5] Deng, Y., Zhou, X., Shen, J., Xiao, G., Hong, H., Lin, H., ... & Liao, B. Q. (2021). New methods based on back propagation (BP) and radial basis function (RBF) artificial neural networks (ANNs) for predicting the occurrence of haloketones in tap water. Science of The Total Environment, 772, 145534.

[6] Nguyen, M. H., Nguyen, P. L., Nguyen, K., Le, V. A., Nguyen, T., & Ji, Y. (2021). PM2.5 Prediction Using Genetic Algorithm-Based Feature Selection and Encoder-Decoder Model. IEEE Access, 9, 57338–57350. https://doi.org/10.1109/access.2021.3072280

[7] Wu, Y., Liu, G., Li, Y., & Jiang, M. (2021). Filter - based feature ranking technique for target recognition by radar infrared combined sensors. IET Radar Sonar & Navigation, 16(1), 182 - 192. https://doi.org/10.1049/rsn2.12175

[8] Hikichi, S., Sugimoto, M., & Tomita, M. (2020). Correlation-centred variable selection of a gene expression signature to predict breast cancer metastasis. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-64870-z

[9] Chan, J. Y., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z., Lin, J., & Chen, Y. (2022). A Correlation-Embedded Attention Module to Mitigate Multicollinearity: an algorithmic trading application. Mathematics, 10(8), 1231. https://doi.org/10.3390/math10081231

[10] Huo, X., Su, H., Yang, P., Jia, C., Liu, Y., Wang, J., ... & Li, J. (2024). Research of Short-Term Wind Power Generation Forecasting Based on mRMR-PSO-LSTM Algorithm. Electronics, 13(13), 2469.

[11] Yan, Q., Wang, R., Dong, Y., Lv, X., Tang, X., Li, X., & Niu, Y. (2022). Logistic LASSO Regression for Dietary Intakes and Obesity: NHANES (2007-2016).

[12] Abdelwahed, N. M., El-Tawel, G. S., & Makhlouf, M. A. (2022). Effective hybrid feature selection using different bootstrap enhances cancers classification performance. BioData Mining, 15(1), 24.

[13] Arafin, S. K., Ul-Saufie, A. Z., Ghani, N. A. M., & Ibrahim, N. (2024). Feature Selection Methods Using RBFNN Based on Enhance Air Quality Prediction: Insights from Shah Alam. International Journal of Advanced Computer Science & Applications, 15(11).

[14] Ul-Saufie, A. Z., Hamzan, N. H., Zahari, Z., Shaziayani, W. N., Noor, N. M., Zainol, M. R. R. M. A., ... & Vizureanu, P. (2022). Improving air pollution prediction modelling using wrapper feature selection. Sustainability, 14(18), 11403.

[15] J. Kalajdjieski et al., "Air Pollution Prediction with Multi-Modal Data and Deep Neural Networks," Remote Sensing, vol. 12, no. 24. 2020, doi: 10.3390/rs12244142.

[16] Arafin, S. K., Ul-Saufie, A. Z., Ghani, N. A. M., & Ibrahim, N. (2024). A Two-Stage Feature Selection Method to Enhance Prediction of Daily PM2. 5 Concentration Air Pollution: 10.32526/ennrj/22/20240049. Environment and Natural Resources Journal, 22(6), 500-509.

[17] Desiani, A., Andriani, Y., Irmeilyana, I., Primartha, R., Arhami, M., Fitrianti, D., & Syafitri, H. N. (2023). The comparison of ReliefF and C.45 for feature selection on heart disease classification using backpropagation. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 17(2). https://doi.org/10.22146/ijccs.82948

[18] Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. Machine learning, 53, 23-69.

[19] Ibrahim, N. B. (2020). Variable selection methods for classification: application to metabolomics data. The University of Liverpool (United Kingdom).

[20] Umar, M. A., Chen, Z., Shuaib, K., & Liu, Y. (2024). Effects of feature selection and normalization on network intrusion detection. Data Science and Management. https://doi.org/10.1016/j.dsm.2024.08.001

[21] Koc, K., Ekmekcioğlu, Ö., & Gurgun, A. P. (2022). Prediction of construction accident outcomes based on an imbalanced dataset through integrated resampling techniques and machine learning methods. Engineering Construction & Architectural Management, 30(9), 4486–4517. https://doi.org/10.1108/ecam-04-2022-0305

[22] Rahman, M. N. A., Ismail, M. S., Wakid, S. A., Syazwan, W. M., Abd Aziz, N. A., Mustafa, M., ... & Yap, C. K. (2023). A Preliminary Checklist of Molluscs in the Kelanang Coast at Banting: An Observational Study. Journal ISSN, 2766, 2276.

[23] Hammad, M. S., Ghoneim, V. F., Mabrouk, M. S., & Al-Atabany, W. I. (2023). A hybrid deep learning approach for COVID-19 detection based on genomic image processing techniques. Scientific Reports, 13(1), 4003.

Lightweight CA-YOLOv7-Based Badminton Stroke Recognition: A Real-Time and Accurate Behavior Analysis Method

Yuchuan Lin[1]*
Public Physical Education Department, Fujian Agriculture and Forestry University, FuZhou FuJian, 35000, China

***Abstract*—With the rapid development of sports technology, accurate and real-time recognition of badminton stroke postures has become essential for athlete training and match analysis. This study presents an improved YOLOv7-based method for badminton stroke posture recognition, addressing limitations in accuracy, real-time performance, and automation. To optimize the model, pruning techniques were applied to the backbone structure, significantly enhancing processing speed for real-time demands. A parameter-free attention module was integrated to improve feature extraction without increasing model complexity. Furthermore, key stroke action nodes were defined, and a joint point matching module was introduced to enhance recognition accuracy. Experimental results show that the improved model achieved a mAP@0.5 of 0.955 and a processing speed of 44 frames per second, demonstrating its capability to deliver precise and efficient badminton stroke recognition. This research provides valuable technical support for coaches and athletes, enabling better analysis and optimization of stroke techniques.**

***Keywords*—Badminton shot; pose recognition; YOLO V7; size adaptive input; model pruning; attention mechanism**

I. Introduction

The core of badminton lies in the precise and skillful execution of movements, such as high shots, smashes, and picks. While the sport itself has a low entry barrier, mastering its complex technical aspects requires systematic training and professional guidance. However, badminton instruction faces significant challenges, including limited teacher resources, the inability to provide personalized guidance to each student, and uneven teaching standards. These factors collectively constrain the efficiency and quality of students' learning of technical movements.

Fortunately, technological advancements offer new possibilities for addressing these challenges [1] [2]. In particular, the development of artificial intelligence, computer vision, and deep learning technologies has begun to play an increasingly important role in sports training and education [3] [4]. By leveraging these technologies, it is possible to develop an intelligent badminton technique action recognition system that can automatically recognize and analyze athletes' movements, provide objective feedback for improvement, and assist both amateur enthusiasts and students in learning badminton techniques independently and efficiently.

Building on these advancements, this article proposes an improved deep learning model—a badminton posture recognition system based on YOLOv7—designed to enhance the quality and efficiency of badminton technique instruction through precise motion capture and analysis. This system aims to improve the recognition accuracy and efficiency of existing models by incorporating data augmentation, anchor box fine-tuning, and the Coordinate Attention (CA) mechanism. The proposed system will serve as a valuable tool for both instructors and students, providing better support for teaching and self-training in badminton. By improving technical proficiency and reducing the risk of injuries caused by incorrect movements, this system seeks to offer scientific and systematic support for learning badminton techniques, ultimately enhancing both the quality and effectiveness of badminton training.

The research will explore the integration of these advanced methods into a cohesive system, aiming to address the current limitations in badminton training and teaching. Through thorough experimentation and analysis, this study seeks to contribute to the development of more effective and accessible tools for badminton instruction, potentially influencing broader applications in other sports as well.

II. Literature Review

This paper will collect existing work on pose recognition in sports to highlight the shortcomings of existing research.

A. Research on Deep Learning in Sports Action Recognition

In the realm of sports, particularly in small ball racket sports like badminton, tennis, and table tennis, motion recognition technology primarily utilizes two approaches: the first involves using the acceleration sensors in wearable devices to collect and classify data for motion detection, while the second applies deep learning technology to extract and learn features from video images for action recognition.

Numerous studies have highlighted the potential of these technologies. For instance, Ang et al. [5] developed a tennis visualization system that organizes and classifies match information, helping users better understand tennis events. Johnson et al. [6] introduced a method that enhances badminton serving accuracy through 3D tracking technology. Building on this, Dierickx et al. [7] further improved the accuracy of trajectory detection. Situmeang et al. [8] combined multiple visual analysis techniques to study athletes' movement characteristics and countermeasures. Jing et al. [9] utilized a support vector machine to recognize common swing actions in badminton game videos, whereas Wang et al. [10]

*Corresponding Author

significantly enhanced action recognition accuracy with a double-layer classifier algorithm.

While deep learning demonstrates high accuracy and comprehensive feature analysis in sports action recognition, it also presents challenges, such as the need for large data sets and the complexity of data collection. In small sample data sets, these techniques are prone to overfitting and require careful processing and optimization in practical applications.

B. Human Posture Recognition Method

The advent of Convolutional Neural Networks (CNNs) has led researchers to adopt deep learning-based techniques for human pose recognition. This approach essentially involves constructing convolutional layers to leverage large datasets, thereby extracting effective feature information to represent the human body. These features are then used in conjunction with efficient classification models for supervised training to achieve accurate predictions. The input data typically consists of an image or a video sequence, and after training the deep learning model, it can identify key body parts in the image, such as the head, arms, and knees.

Currently, there are two main research approaches to deep learning-based human pose recognition: direct regression based on keypoint coordinates and regression based on keypoint heatmaps. In 2014, the release of the MPII dataset [11], which contains approximately 25,000 images and covers over 40,000 human instances, with each instance including 16 keypoints, significantly advanced research in this field. In 2016, Wei et al. designed the CPM deep network [12], which extracts receptive fields of different sizes through repeated convolution operations and combines contextual information from the image to recognize human poses. CPM also introduced the concept of intermediate supervision, effectively addressing the issues of network depth and vanishing gradients, thus improving pose recognition accuracy. Additionally, the Stacked Hourglass Network (SHN) [13] introduced multi-resolution heatmap regression and multi-scale receptive field mechanisms, which enhanced the CNN structure for feature extraction, yielding excellent results. That same year, the release of the MSCOCO dataset [14] further enriched research resources, increasing the number of human keypoints to 18, making it more precise and comprehensive than MPII.

Pose recognition in complex multi-person scenes involves two approaches: top-down and bottom-up. The top-down approach first detects each person in the image through object detection, then performs keypoint recognition on each detected person. This method is straightforward but its performance heavily depends on the accuracy of human detection, and its processing time increases linearly with the number of people. Representative models include HRNet [15] and RMPE [16]. The bottom-up approach, on the other hand, first identifies all keypoints and then connects them according to the human pose model. This method does not slow down with an increase in the number of people, but matching keypoints in complex scenes can be challenging. Representative algorithms include OpenPose [17] and DeepCut [18].

Despite the achievements in theoretical innovation and technological development both domestically and internationally, human pose recognition technology still faces several challenges and difficulties that require further research and optimization.

C. Research Gaps

Despite significant progress in badminton stroke posture recognition technology, several challenges remain in practical applications:

1) The demand for high-precision recognition: Accurate recognition of badminton stroke posture is crucial, as even minor differences in movement can significantly affect the effectiveness of a stroke. However, existing technologies still require improvements in accuracy and robustness, especially in complex environments.

2) Challenges with real-time performance: In actual training or matches, there is a critical need for real-time posture recognition, enabling coaches and athletes to receive immediate feedback and adjust strategies accordingly. Current systems continue to face limitations in processing speed and providing real-time feedback.

3) Issues with automation and adaptability: Most current posture recognition systems require specific setups, such as particular camera angles and lighting conditions, which restricts the system's applicability and flexibility.

The above three points are the key to improving badminton hit recognition and also the focus of this study.

III. Detection Model

A. Framework

Badminton stroke posture recognition technology is primarily utilized in sports training and match analysis. By recognizing athletes' stroke postures, coaches can more effectively guide athletes in adjusting their techniques and optimizing training outcomes. This technology captures detailed aspects of athletes' movements, thereby aiding in the analysis of the precision and efficiency of stroke techniques. Currently, this field predominantly relies on video analysis and sensor technology, combined with machine learning and deep learning methods, to achieve motion capture and data analysis.

To address the limitations of existing methods, a badminton stroke posture recognition method based on an improved YOLO V7 model is proposed. The enhanced YOLO V7 model features a fast detection architecture composed of the Backbone module (lite-Darknet), Bottleneck module, Head module, and threshold matching module, as illustrated in Fig. 1. By introducing a parameter-free attention module, the model's feature extraction capability is enhanced, allowing it to adapt to varying environments and conditions. Additionally, model pruning techniques are employed to improve processing speed, and key nodes are defined, with a joint point matching module integrated to enhance matching accuracy. These advancements significantly increase the automation, real-time performance, and accuracy of posture recognition,

thus better meeting the needs of badminton training and match analysis.

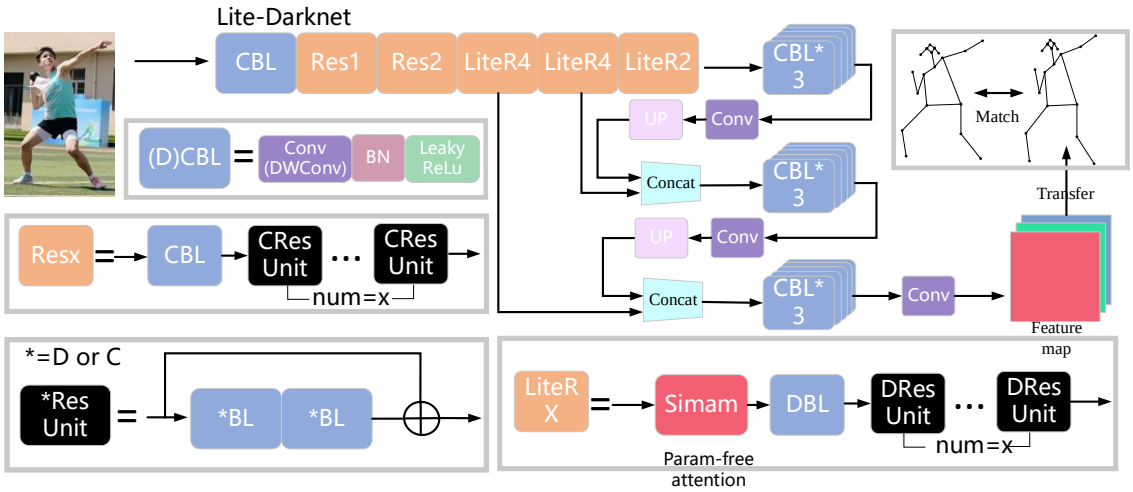


Fig. 1. Model architecture diagram.

To further improve the model's adaptability, the size of the input image is first adjusted to 640×640 pixels through scale normalization. The Backbone module extracts feature from the input images, identifies areas where node features are located, and proceeds to fuse these features across layers via the Bottleneck module. The semantic features and location features are then combined, and the simplified Head layer classifies the image. Finally, the classified features are sent to the node matching module, which accurately determines the connections between human nodes, thereby generating the detection result.

B. Node Feature Definition

Human body nodes need to be set before model training. In this paper, 19 nodes are selected and marked with 1-19 labels respectively, as shown in Fig. 2.

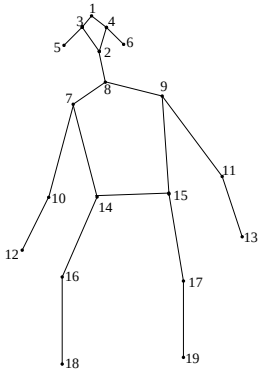


Fig. 2. Human body node diagram.

In Fig. 2, 1 is forehead, 2 is nose, 3 is left eye, 4 is right eye, 5 is left ear, 6 is right ear, 7 is left shoulder, 8 is neck. 9 is the right shoulder, 10 is the left elbow, 11 is the right elbow, 12 is the left wrist, 13 is the right wrist, 14 is the left hip, 15 is the right hip, 16 is the left knee, 17 is the right knee, 18 is the left ankle, 19 is the right ankle. The marked data set of 19 nodes was input into the improved YOLO V7 network for training, and the inference model was obtained. The inference model could detect 19 nodes of the human body in the image and preliminarily match them to get the result of human body pose.

C. Node Matching Template

There may be some errors in the preliminary matching result, for example, the node is matched to the wrong target body when multiple people overlap, so the node matching template needs to be used for accurate matching. According to the actual situation, this paper sets up a single human body posture of the node template, part of the Fig. 3, mainly including human swing, standing, single leg lift, squat and other posture.

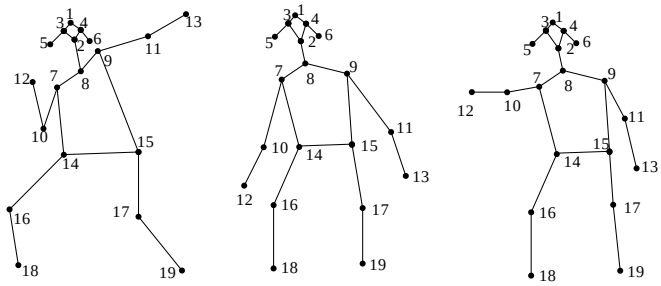


Fig. 3. Single human body posture of the node template.

In the exact matching operation, we first need to establish the threshold node descriptor set of the preliminary matching result graph and the template graph and realize the threshold node accurate matching by comparing the distance of the threshold node descriptor in the two-point sets. The calculation of the node descriptor in the template diagram is shown in Eq. (1):

$$R_i = (r_{i1}, r_{i2}, \ldots, r_{in}) \quad (1)$$

Where, the R_i presentation template key descriptor in figure collection, r_{in} presentation template the key points in the graph. Preliminary matches the key descriptor in figure calculation as shown in Eq. (2):

$$S_i = (s_{i1}, s_{i2}, \dots, s_{in}) \quad (2)$$

Where, the S_i said preliminary joint point descriptor set, match the picture s_{in} said preliminary match key points in the figure. Any two-descriptor similarity measure computation as shown in Eq. (3):

$$d(R_i, S_i) = \sqrt{\sum_{j=1}^{n} \left(r_{ij} - s_{ij}\right)^2} \quad (3)$$

$d(R_i, S_i)$ need satisfaction:

$$\frac{S_j}{S_p} < \delta \quad (4)$$

The S_j for distance R_i point recently, S_p for distance R_i near point, δ for distance threshold, when the threshold value is less than the set value to get the final figure.

D. Size Adaptive Normalization

In order to make the image have a unified scale, it is necessary to carry out size adaptive operation. At the same time, in order to deal with the problem of image feature intensity decline after operation and stabilize the detection accuracy of the model, it is necessary to enhance the image first. When image enhancement is carried out, the standardization process is carried out first. Common standardization methods include linear stretching, mean-variance normalization, histogram equalization, etc. This paper adopts mean-variance normalization method. Variance normalized (Z - score Normalization) is put all the data to the distribution of mean, variance 1 to 0, calculated as shown in Eq. (5):

$$x_{\text{saale}} = \frac{x-\mu}{S} \quad (5)$$

Where, the x_{saale} for the value of the normalized after x to the value of the normalized μ for the average of the image pixels, S as the standard deviation of the image and the standard deviation of calculated as shown in Eq. (6):

$$S = max\left(\sigma, \frac{1.0}{\sqrt{N}}\right) \quad (6)$$

Where σ is the standard variance and N is the number of pixels in the image. Standardized normalized processing after processing, the original data is mapped to get image Z on [0, 1] interval, calculated as shown in Eq. (7):

$$Z = \frac{x_i - min(x)}{max(x) - min(x)} \quad (7)$$

Where, the x_i with the value of image pixels, Max (x) and min (x) respectively, of the maximum and the minimum of image pixels. After normalized image to zoom in, fixed to 640 x 640 pixels, calculation process is as follows: according to the target size and the size of the original image, computing needs zoom ratio, can according to the proportion of long or short while zooming. For example, suppose the target size is 640×640 pixels, the scaling ratio needs to be calculated based on the width and height of the original image, as shown in Eq. (8) and (9):

$$\text{scale}_1 = max\left(\frac{w}{W}, \frac{h}{H}\right) \quad (8)$$

$$\text{scale}_2 = min\left(\frac{w}{W}, \frac{h}{H}\right) \quad (9)$$

Where, the scale_1 to long side scaling, scale_2 short while scaling w for target image (this article is 640 pixels), long W Z long, for the image h for target image of high (this article is 640 pixels), H is the height of the Z image. Will be calculated according to the original image to zoom scaling operation, during operation (such as bilinear interpolation), using interpolation algorithm is used to keep the image quality.

E. Parameterless Attention Mechanism

The core idea of Simam is based on the local self-similarity of images. In an image, there is usually a strong similarity between adjacent pixels, while the similarity between distant pixels is weak. Simam uses this property to generate attention weights by calculating the similarity between each pixel in the feature map and its neighbors. The following energy function is defined for each neuron:

$$e_t(w_t, b_t, y, x_i) = (y_t - \hat{t})^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}(y_o - \hat{x}_i)^2 \quad (10)$$

Where, $\hat{\mathrm{t}} = w_t t + b_t$, $\hat{\mathrm{x}}_i = w_t x_i + b_t$ is the target neurons and other neurons $x\ i$ in figure x characteristics of the single channel of the linear transformation, i is the index on the spatial dimension, T y_t and y_o is the target neurons and other neurons x_i two different values, M is the number of neurons on one channel. w_t and b_t is linear transformation of weights and bias. All values are in the scalar type, when t equals y_o, x_i equals y_t, minimum energy function. To minimize the above formula is equivalent to the training of neurons within the same channel linear separability between t with other neurons. For simplicity, we take binary labels and add regular terms, and the final energy function is defined as follows:

$$e_t(w_t, b_t, y, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1 - (w_t x_i + b_t))^2 + (1 - (w_t t + b_t))^2 + \lambda {w_t}^2 \quad (11)$$

The analytical solution of the above equation is:

$$w_t = -\frac{2(t-\mu_t)}{(t-\mu_t)^2 + 2\sigma_t^2 + 2\lambda} \quad (12)$$

$$b_t = -\frac{w_t(t+\mu_t)}{2} \quad (13)$$

Since all neurons on each channel follow the same distribution, it is possible to first calculate the mean and variance of the input features in the H and W dimensions to avoid double computation:

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t-\hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \quad (14)$$

The whole process can be expressed as:

$$\tilde{X} = sigmoid\left(\frac{1}{E}\right) \odot X \quad (15)$$

By calculating the mean and variance, the weight of each pixel in the feature map is adjusted adaptively, so that important pixels are amplified, and unimportant pixels are suppressed. Since SimAM does not require additional convolutional operations or full connection layers, it has low computational overhead and is suitable for embedding in various convolutional neural networks. In YOLOv7 target

detection algorithm in the whole network structure is divided into three parts, respectively is the network part of feature extraction, feature fusion network part and YOLO head testing part finally. Since the feature extraction network will output feature images of three different scales, three attention mechanisms need to be added, followed by a three-layer feature extraction module LiteR2. Attention mechanism is added in the network to determine the image characteristics of the channel number, the number of channels can be determined by the upper the output of the network structure, network structure exist three output in feature extraction, from top to bottom respectively is 80 x 80 x 512, 1024, 20 x 40 * 40 * 20 x 1024, Therefore, the number of channels corresponding to the three attention mechanism modules is 512, 1024, and 1024 respectively.

IV. Experiment and Verification

In this section, we will conduct ablation experiments based on self-made test data sets to verify the effectiveness of the proposed method.

A. Data Set Construction and Processing

In this study, an enhanced OpenPose model was selected to detect the skeletal key points of badminton technical movements. Given the limitations of existing datasets such as FineGym and UCF101 in supporting badminton-specific scenarios, a new badminton action sample dataset was independently constructed. This dataset specifically focuses on collecting videos of forehand high shots in badminton, which will be used for subsequent feature extraction and classification tasks.

The forehand high shot (using a right-handed racket as an example) is a fundamental and crucial technique in badminton, as it can effectively force the opponent to the back of the court and create an advantageous position. This technique can be systematically broken down into the following four steps, as shown in Fig. 4.

1) Preparation Phase

- Position and Posture: Stand with feet shoulder-width apart, left foot forward, right foot back, and body positioned sideways to the net, forming a stable support base.
- Grip and Arm Position: Hold the racket in the right hand with a moderate grip. Bend the arms naturally, point the racket in the direction of the ball, and slightly raise the left hand to aid in body balance.
- Sight: Focus the eyes on the incoming ball, preparing to execute the stroke.

2) Lead-Up Action

- Racket Head Movement: From the ready position, lift the right elbow upward and pull the racket head back and upward to a position directly above the head, with the racket face oriented forward.
- Body Coordination: Move the body slightly backward in sync with the racket head to increase the power of the shot. Gradually shift the weight from the right foot to the left, maintaining balanced movement.
- Racket Face Adjustment: Rotate the racket face slightly inward when overhead, ensuring readiness to strike the ball at the correct angle.

3) Swinging to the Ball

- Hit Point Selection: Select a hitting point slightly above shoulder height on the right side of the body, allowing better control over the flight and power of the ball.
- Swing: Swing the racket from the bottom of the lead position forward and upward, swiftly and forcefully, ensuring that the racket face is aligned with the ball at the moment of contact.
- Wrist Utilization: At the end of the swing, accelerate the racket head through rapid wrist action to generate more power and control, directing the ball toward the opponent's far court.

4) Follow-Through and Recovery

- Finishing Position: After striking the ball, rotate the body gently to the left and forward, aiding in returning to a stable state and preparing for the next action.
- Weight Adjustment: Quickly redistribute the weight after the shot to be ready to move or return to a defensive position, in anticipation of the opponent's return.
- Racket Face Adjustment: Ensure that the racket face is realigned toward the net after the shot, facilitating preparation for the next stroke.

Following these four detailed steps allows for effective execution of forehand high shots, thereby not only improving the quality of the stroke but also gaining better control over the rhythm and strategic layout of the court.

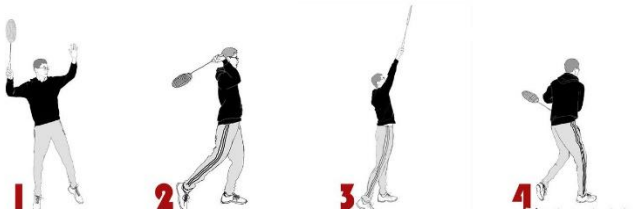


Fig. 4. Forehand high ball technique breakdown diagram.

B. Collection of Experimental Data Sets

In this study, a high-quality badminton stroke pose recognition dataset was constructed by carefully selecting 20 right-handed participants, including 10 males and 10 females, aged between 22 and 27 years. The participant group consisted of half national-level badminton players and half badminton enthusiasts with a certain level of skill. All participants passed a series of motor stability and reliability tests before inclusion, and their heights were controlled between 173 cm and 178 cm to minimize variations in technical movements due to differences in body shape.

Data acquisition was conducted through two methods: field acquisition and network video acquisition. Field data were collected using DJI Action2 and Nikon D70s cameras from May to June 2024 on a standard badminton court, with the ADIBO Smart Badminton server A200 used to ensure consistent service conditions, as shown in Fig. 5. In addition, qualified badminton match videos were screened from online video platforms to enhance the diversity and coverage of the dataset. To ensure data consistency and repeatability, the position of the serve was precisely controlled, and participants were required to hit the ball within a specified area, effectively reducing data bias caused by variations in participant positioning.

As a result of these methods, a comprehensive badminton stroke pose dataset was successfully constructed, including data from both professional players and enthusiasts. The aim is to promote the scientific analysis of badminton technical training and competition by accurately identifying and analyzing stroke techniques. This dataset serves as a valuable resource not only for the technical analysis of badminton but also as an experimental foundation for subsequent research in computer vision and machine learning.

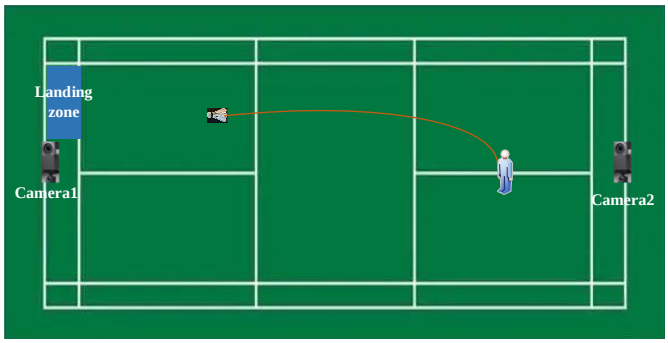


Fig. 5. Data acquisition method.

To further refine the dataset, point coordinates were normalized to reduce the influence of body center variations, and the body was rotated to a fixed angle to minimize the impact of different viewing angles.

The experimental setup is outlined in Table I. The PyTorch deep learning framework was utilized for the experiment. Stochastic Gradient Descent (SGD) was selected as the optimizer, with a batch size of 16. The cross-entropy loss function was employed as the loss function. The learning rate decay strategy implemented a reduction to one-tenth of the original rate at each specified interval. For the JDTD datasets, the initial learning rate was set at 0.1. The first dataset was trained for 50 epochs in total, with the learning rate reduced at the 30th and 40th epochs. The second dataset underwent 65 epochs of training, with the learning rate decaying at the 45th and 50th epochs.

TABLE I. ALGORITHM EXPERIMENT ENVIRONMENT

Environment	Version	Environment	Version
System	Ubuntu20.04	**CPU**	I7-8700k
Graphics card	GTX1080Ti	**Internal memory**	32G
Python	3.7	**PyTorch**	1.2 .0
CUDA	11.1	**CUDNN**	8.0 .4

C. Ablation Experiment

In order to verify the effectiveness of the improved algorithm in this paper, the detection experiment results of YOLO V7 original model, YOLO V7+ adaptive input module, YOLO V7 pruning model + adaptive input module, YOLO V7 model + adaptive input module + node matching module were studied on the same data set, as shown in Table II.

As can be seen from Table II, although the detection speed of the model is not improved after the addition of the adaptive module, the adaptability of the model to images of different sizes is greatly improved, and the detection accuracy is improved. When the pruning model is used, the detection speed and accuracy of the model are greatly improved when the IOU threshold is low, but the detection accuracy is not improved when the IOU threshold is high. After the node matching module is added, the detection speed and accuracy of the model are improved, which proves the effectiveness of the model improvement.

TABLE II. ABLATION RESULTS

Model	mAP@0.5	mAP@0.75	FPS
YOLO V7 original model	0.946	0.81	35
YOLO V7+ adaptive input module	0.947	0.83	35
YOLO V7 Pruning model + adaptive input module	0.952	0.83	39
YOLO V7 model + adaptive input module + node matching module	0.955	0.86	44

Note: mAP@0.5 indicates that the IOU threshold is 0.5, and FPS indicates the number of images processed per second

Fig. 6 illustrates the partial results of badminton stroke posture recognition using the trained model. As shown in the figure, the model demonstrates effective detection for both single-player and multi-player scenarios with minimal occlusion. For images with some occlusion, a node matching template is employed to enhance the model's detection accuracy. This issue arises when multiple players overlap significantly or exhibit similar postures, resulting in the detected nodes being unable to accurately distinguish the target player, thereby leading to detection errors.

The improved model, which incorporates an attention mechanism alongside data augmentation and anchor box adjustment, is based on YOLOv7. This enhanced model is compared with the baseline model to evaluate the performance improvements brought by the addition of the attention mechanism as shown in Table III and Table IV. In terms of posture recognition accuracy, there is noticeable improvement across all phases of the stroke. For instance, the accuracy of detecting the "Swinging to the Ball" phase increased from 86.12% to 88.26%, while the accuracy for the "Follow-Through and Recovery" phase approached 90%. Similarly, recall rates saw significant increases: the "Preparation Phase" improved by 1.73%, the "Lead-Up Action" by 2.06%, the "Swinging to the Ball" by 1.4%, the "Follow-Through and Recovery" by 3%, and overall detection recall exceeded 90% for some phases. These comparisons indicate that the performance of the YOLOv7 model has been significantly enhanced following the integration of the attention mechanism.

The mAP (mean Average Precision) for the model is detailed in Table V.

In this study, the average accuracy of the badminton stroke posture recognition model was improved from an initial 85.25% to 88.50%. This 3.25% increase in accuracy was achieved through a combination of data augmentation, anchor box adjustment, and the incorporation of an attention mechanism.

TABLE III. COMPARISON OF MODEL ACCURACY AFTER CA MECHANISM IS ADDED

Algorithm	Preparation Phase	Lead-Up Action	Swinging to the Ball	Follow-Through and Recovery
YOLOv7+ Data expansion +anchor fine-tuning	86.12%	83.51%	85.52%	88.31%
YOLOv7+ Data expansion +anchor fine-tuning +CA	88.26%	85.58%	86.92%	89.30%

TABLE IV. COMPARISON OF MODEL RECALL RATE AFTER ADDING SIMAM MECHANISM

Algorithm	Preparation Phase	Lead-Up Action	Swinging to the Ball	Follow-Through and Recovery
YOLOv7+ Data expansion +anchor fine-tuning	86.12%	83.51%	85.52%	88.31%
YOLOv7+ Data expansion +anchor fine-tuning +CA	87.85%	85.57%	86.92%	91.30%

TABLE V. COMPARISON OF MODEL RECALL RATE AFTER ADDING SIMAM MECHANISM

Algorithm	mAP
YOLOv7	85.25%
YOLOv7+ Data expansion	87.01%
YOLOv7+ Data expansion +anchors fine tuning	87.50%
YOLOv7+ Data expansion +anchors trimming +CA	88.50%

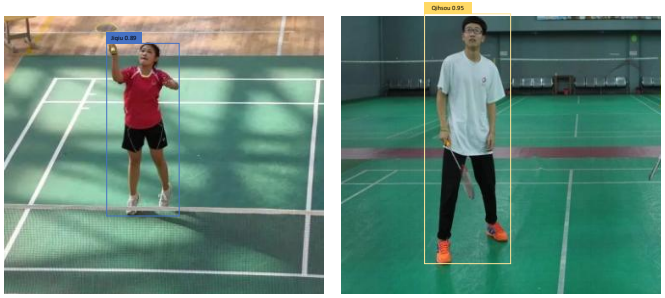


Fig. 6. Identify results after adding the simam attention mechanism.

D. Comparison of Other Algorithms

In the previous discussion, data augmentation, anchor box adjustment, and the integration of an attention mechanism were applied to the original YOLOv7 model. These modifications ultimately demonstrated an improvement in the accuracy of YOLOv7 for badminton stroke posture recognition. The enhanced YOLOv7 algorithm is compared with other target detection algorithms to demonstrate the superiority of the proposed approach.

In the field of target detection, Faster R-CNN, a two-stage detection algorithm, is widely utilized, including in transmission line inspections. Similarly, CenterNet, an anchor-free detection method, has achieved significant results in 3D target detection. Other network structures, such as DenseNet and BiFPN, are also extensively used in target detection. These network structures are applied to badminton stroke posture recognition, and the resulting average accuracies are compared to validate the advantages of the proposed algorithm.

- YOLOv3: Utilizes the Darknet53 network structure, combining residual learning and multi-scale output to optimize deep networks and enhance feature extraction capabilities.
- Faster R-CNN: A typical two-stage target detection algorithm that merges Fast R-CNN and RPN networks to rapidly and accurately generate candidate regions and detect targets.
- CenterNet: An anchor-free detection method that simplifies the detection process and speeds it up by detecting the target center and using heatmap technology.
- MobileNet: A lightweight deep neural network that employs depthwise separable convolution to significantly reduce computation, making it suitable for embedded systems.
- BiFPN: A bidirectional weighted feature pyramid network structure that optimizes feature fusion and improves the performance and efficiency of the detection network.

As shown in the Table VI, the average accuracy of the improved YOLOv7 algorithm presented in this study is 3.6% higher than that of the YOLOv7-BiFPN algorithm, 22.75% higher than that of the MobileNet algorithm, and 4.58% higher than that of the CenterNet algorithm. It also surpasses the Faster R-CNN algorithm by 3.83% and the YOLOv3 algorithm by 7.34%.

TABLE VI. COMPARISON OF MODEL ACCURACY UNDER DIFFERENT ALGORITHMS

Algorithm	mAP
YOLOv3	81.16%
Faster-RCNN	84.67%
CenterNet	83.92%
MobilNet	65.75%
YOLOv7-BiFPN	84.90%
YOLOv7+CA	88.50%

V. CONCLUSION

Badminton, as a fast-paced and technically demanding sport, requires precise analysis of player movements to enhance performance and training outcomes. Recognizing this need, this study proposes a deep learning-based algorithm for badminton player pose recognition. The proposed method addresses the challenges of accuracy and efficiency in pose

recognition by integrating data augmentation, anchor box fine-tuning, and the Coordinate Attention (CA) mechanism into the YOLOv7 algorithm. These improvements significantly enhance the model's recognition capabilities.

Experimental results demonstrate that the average accuracy of the improved YOLOv7 model has increased from 85.25% to 88.5%, outperforming other popular algorithms such as YOLOv3, Faster R-CNN, CenterNet, MobileNet, and YOLOv7-BiFPN. Moreover, the introduction of the CA attention mechanism has yielded particularly noteworthy results in recognizing specific badminton actions, such as smashes and picks, by reducing instances of missed or false detections. Looking ahead, future research will focus on further refining the dataset, exploring additional challenges in motion recognition, and considering the integration of edge computing techniques to optimize real-time processing capabilities.

ACKNOWLEDGMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Liu S, Zheng P, Xia L, et al. A dynamic updating method of digital twin knowledge model based on fused memorizing-forgetting model[J]. Advanced Engineering Informatics, 2023, 57: 102115.

[2] Fu T, Li P, Liu S. An imbalanced small sample slab defect recognition method based on image generation[J]. Journal of Manufacturing Processes, 2024, 118: 376-388.

[3] Fu T, Liu S, Li P. Digital twin-driven smelting process management method for converter steelmaking[J]. Journal of Intelligent Manufacturing, 2024: 1-17.

[4] Fu T, Liu S, Li P. Intelligent smelting process, management system: Efficient and intelligent management strategy by incorporating large language model[J]. Frontiers of Engineering Management, 2024: 1-17.

[5] Polk Tom,Yang Jing,Hu Yueqi,et al.TenniVis:Visualization for Tennis Match Analysi s[J].IEEE transactions on visualization and computer graphics.2014,20(12):2339-2348.

[6] Shayne Vial,Jodie Cochrane,Anthony J,et al.Using the trajectory of the shuttlecock as a measure of performance accuracy in the badminton short serve[J].International Jou rnal of Sports Science&Coaching,2019,14(1):91-96.

[7] Dierickx T.Badminton Game Analysis from Video Sequences[D].Ghent:University of Ghent, 2015 .

[8] WeiTa C,Sitmeang S.Badminton Video Analysis based on Spatitemporal and Stroke Features[J].International Conference on Multimedia Retrieval,2017,12(1):121-122.

[9] Heemskerk C, David L, Steve S, et al. The effect of physical education lesson intensity and cognitive demand on subsequent learning behaviour[J]. Journal of science and medicine in sport,2020,23(6):586-590.

[10] Ben AT, Elleuch I, Guermazi R. Student Behavior Recognition in Classroom using De -ep Transfer Learning with VGG-16[J]. Procedia Computer Science,2021,192:951-960.

[11] Andriluka M, Pishchulin L, et al. Human Pose Estimation: New Benchmark and State of the Art Analysis[C]// Computer Vision and Pattern Recognition (CVPR). IEEE, 2014.

[12] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional Pose Machines[C]// Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[13] Newell A, Yang K, Deng J. Stacked Hourglass Networks for Human Pose Estimation[J]. arXiv e-prints, 2016.

[14] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context[J]. 2014.

[15] Sun K, Xiao B, Liu D, et al. Deep High-Resolution Representation Learning for Human Pose Estimation[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[16] Fang H, Xie S, Tai Y, et al. RMPE: Regional Multi-person Pose Estimation[J]. 2016.

[17] Cao Z, Simon T, Wei S E, et al. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.

[18] Rajchl M, Lee M, Oktay O, et al. DeepCut: Object Segmentation from Bounding Box Annotations using Convolutional Neural Networks[J]. IEEE Transactions on Medical Imaging, 2016, 36(2):674-683.

Fuzzy Evaluation of Teaching Quality in "Smart Classroom" with Application of Entropy Weight Coupled TOPSIS

Yajuan SONG*
Zhengzhou Shengda University, Zhengzhou 410198, Henan, China

***Abstract*—This research aims to investigate the scientific assessment methodology for the teaching quality of smart classrooms and to develop a multi-dimensional evaluation system utilizing a combination of the entropy weight technique and the TOPSIS approach. To comprehensively assess the pedagogical proficiency of educators, this paper selects the dimensions of teaching preparation, the process, teaching effect and teaching reflection, and combines the questionnaire survey and statistical data to collect and analyze the data. The research methodology initially standardized the raw data to mitigate discrepancies among various scales; subsequently, the weight method was employed to ascertain the weight of each evaluation index, thereby indicating the significance of the indices through information entropy; ultimately, the TOPSIS method was utilized to evaluate teachers' performance across each dimension and rank them based on their proximity to the optimal and negative ideal solutions, culminating in a comprehensive assessment of teaching quality. The results of the study show that the entropy weight method can effectively determine the weight of each index, and the TOPSIS method provides teachers with a clear ranking of teaching quality by calculating the distance from the ideal solution, helping to identify strengths and weaknesses in teaching. This paper concludes that the evaluation method combining the entropy weight method and TOPSIS method can provide an objective and comprehensive teaching quality assessment for the smart classroom, but there are limitations such as the small sample data size and some teaching dimensions are not adequately covered, etc. Future research can further improve the evaluation system by expanding the sample size and increasing the evaluation dimensions to enhance its applicability and accuracy, so as to provide stronger support for the continuous optimization of the smart classroom.**

***Keywords*—Smart classroom; entropy weight method; TOPSIS method; teaching quality; optimization and improvement**

I. INTRODUCTION

The evolution of information technology has rendered traditional teaching models insufficient for modern educational needs, thereby catalyzing the development of innovative educational frameworks and philosophies [1]. The smart classroom, which integrates advanced information technology with instructional practices, has emerged as a crucial element in contemporary educational reform [2]. Utilizing the Internet, big data, cloud computing, and other advanced technologies, the intelligent classroom offers educators and learners an abundance of resources and tools, thus substantially broadening the reach and capabilities of educational practices [3]. This innovative teaching model not only enhances classroom instruction but also transforms the educational process into one that is more personalized, interactive, and intelligent. Unlike traditional classrooms, smart classrooms can facilitate personalized teaching, independent learning, cooperative learning, and other diverse teaching methods, thereby better accommodating the varied learning needs and interests of students [4]. As smart classrooms become increasingly widespread in schools, research has begun to focus on optimizing their teaching design and strategies to enhance educational effectiveness [5].

As smart classrooms gain popularity, researchers have increasingly focused on assessing their instructional effectiveness [6]. Accurately evaluating the teaching quality of the smart classroom has emerged as a critical problem in contemporary educational research. Unlike the conventional teaching evaluation system, the assessment in a smart classroom considers not only students' academic performance but also their engagement, interest in learning, collaboration, and the feedback provided by teachers [7]. Therefore, how to scientifically formulate the evaluation standards and methods of teaching quality in smart classroom has become a key direction of research [8]. Many scholars have proposed different evaluation models and methods, such as learning outcome evaluation model, teaching quality evaluation model and so on [9] [10]. Through these evaluation models, the teaching effect of smart classroom can be assessed more comprehensively, providing theoretical support and practical guidance for future educational reform.

Evaluating and enhancing the pedagogical standards in "intelligent learning environments" has emerged as a pivotal concern. This investigation employs the entropy weight approach and the TOPSIS methodology as standard evaluation instruments to gauge the educational quality in "intelligent learning environments." The entropy weight approach assigns weights by computing the information entropy of each metric. Information entropy indicates the extent of variability in a metric; a higher entropy value implies greater information richness and thus a higher weight. The entropy weight approach's advantage lies in its ability to automatically distribute weights based on the data's distribution, reducing subjectivity, minimizing human intervention, and making it appropriate for complex and dynamic decision-making contexts [11]. The TOPSIS approach, which stands for Technique for Order Preference by Similarity to Ideal Solution, serves as a ranking mechanism that assesses the advantages and

*Corresponding Author

disadvantages of different options by measuring the distance of each option from both the optimal and the least favorable solutions [12]. This method pinpoints the choice that is closest to the ideal and farthest from the negative ideal by evaluating the proximity of each alternative to the ideal solution [13]. Recognized for its intuitive comprehension and straightforward computation, it is particularly suitable for decision-making scenarios involving multiple evaluation criteria.

The entropy weight approach can objectively ascertain the weight of each evaluation index in assessing the teaching quality of intelligent learning environments, while the TOPSIS methodology can identify the strengths and weaknesses of different implementation plans for intelligent learning environments [14]. By integrating these two methodologies, it is anticipated that the teaching effectiveness will be evaluated in a more comprehensive and scientific manner, ensuring an objective and thorough assessment process [15].

This study aims to investigate the scientific assessment methodology for the teaching quality of smart classrooms and to develop a multi-dimensional evaluation system utilizing a combination of the entropy weight technique and the TOPSIS approach. To comprehensively assess the teaching quality of teachers, this paper selects the dimensions of teaching preparation, teaching process, teaching effect and teaching reflection, and combines the questionnaire survey and statistical data to collect and analyze the data.

This paper is organized as follows: It begins with an introduction to the study's background and objectives, elucidating the research's significance and innovative aspects in Section I. Section II offers a review of the relevant literature and theoretical framework, laying the groundwork for the subsequent analysis. The third section details the research methodology, including data collection and analysis procedures. Section III presents the study's main findings, illustrated through graphical representations and data analysis. Following this, the paper discusses the results, corroborating and expanding upon existing theories while also highlighting the study's limitations and areas for improvement. Lastly, the paper summarizes the key conclusions and offers recommendations for future research directions.

II. Teaching Quality Evaluation Modeling for the "Smart Classroom"

A. Establishment of an Evaluation Indicator System

This paper outlines the selection of indicators for assessing teaching quality in the "smart classroom," focusing on two primary considerations: the significance and feasibility of crucial aspects, leading to the choice of the most representative and quantitative indicators. These indicators can precisely represent the attainment of the objectives and can be optimized to a certain degree. Secondly, the chosen indicators are assessed to ascertain the actual attainment of the objectives and modified as necessary [16]. These metrics can be achieved by implementing an effective data collection and reporting system to deliver prompt feedback during the teaching and learning process. The ongoing assessment and modification guarantee the efficacy of the teaching quality evaluation system. Future enhancements will encompass many elements: refining indicator settings to align with actual requirements, enhancing data collection methods to augment accuracy, and streamlining processes to boost evaluation speed. The evaluation system can adapt more readily to alterations in the educational landscape and technological advancements.

This paper develops a "smart classroom" teaching quality evaluation index system by integrating the aforementioned concepts and pertinent research, with the objective of establishing a scientific foundation for assessing teaching efficacy and facilitating educational change. The data mostly originate from a questionnaire survey conducted across many classes at a school, hence ensuring the representativeness and practicality of the findings.

B. Entropy Weighted TOPSIS Modeling

1) Entropy weight method: The entropy weighting approach can ascertain the weights of indicators for evaluating teaching quality in a smart classroom, facilitating the assessment of the relative significance of various elements in teaching quality [17]. From Fig. 1, in the intelligent classroom, elements such as educators' proficiency, students' achievement, and the impact of the instructional process are critical assessment metrics. The entropy weighting approach ascertains the weight of each indicator by aggregating data and computing the information entropy associated with each indicator. A higher entropy number indicates a greater disparity among the indicators, necessitating a reduced weight; conversely, a lower entropy value signifies a lesser disparity, warranting a proportionally increased weight. The Entropy Weight Method (EWM) functions as a multi-attribute decision-making tool, designed to determine the weights of different indicators. First introduced by American academic Jay Forrester in 1960, the method has since been advanced and perfected by later scholars. It is grounded in the concept of information entropy, which quantifies the degree of variation among indicators [18]. A higher entropy suggests greater heterogeneity among the indicators, thus warranting a lower weight assignment. In contrast, a lower entropy value implies less variation and a correspondingly higher importance of the indicator. Therefore, by calculating the entropy for each indicator, the EWM can methodically assign appropriate weights to them.

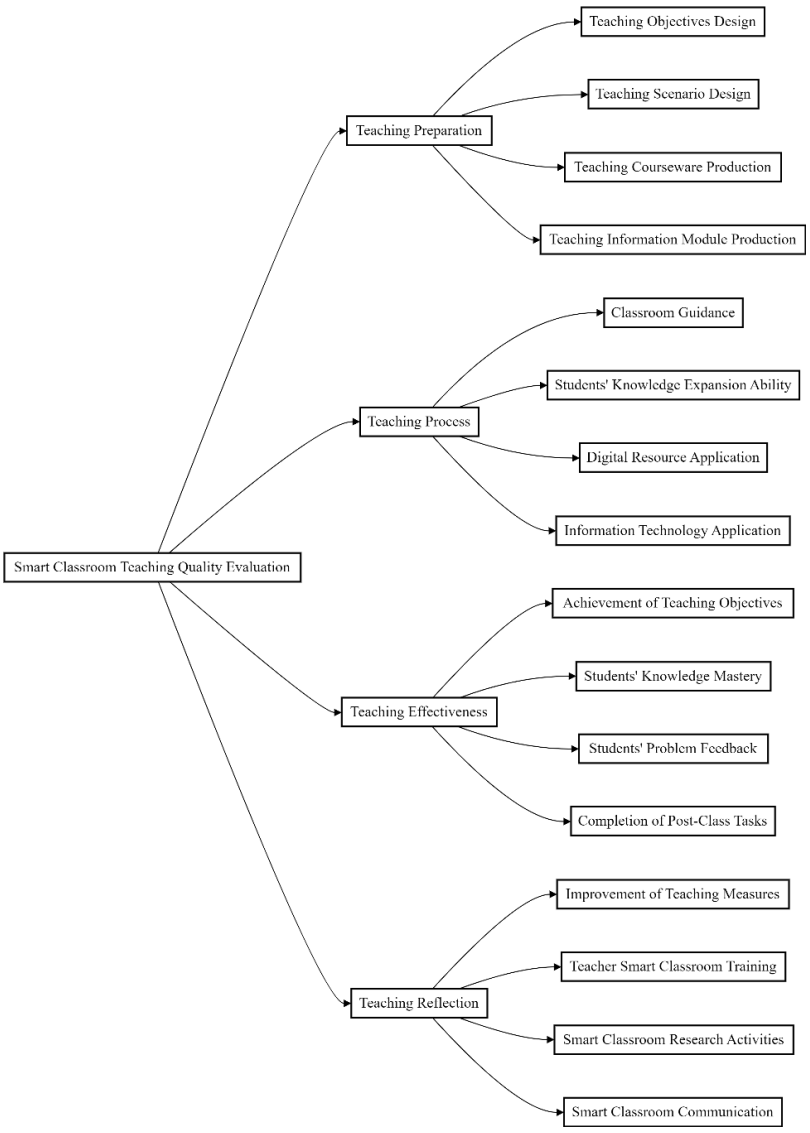


Fig. 1. Teaching quality evaluation index system.

The entropy weight technique offers the benefits of objectively representing the significance of each indication while being straightforward and quick to implement. Nonetheless, the entropy weight technique possesses many disadvantages, including sensitivity to variations in data and stringent standardization prerequisites for the decision matrix. Consequently, this work modifies the standardization approach of the entropy weight method to enhance its applicability and precision [19].

Specifically, the steps of the entropy power method are as follows:

a) Normalization of assessment data: Normalization of sample data involves transforming the original data into a standardized normal distribution characterized by a mean of zero and a standard deviation of one. This process allows for the harmonization of variables that possess disparate scales, units, and ranges into a uniform metric, thereby enhancing the precision of data analysis and the robustness of model development.

In order to maintain the distributional characteristics of the data, reduce the interference of outliers on the results, and facilitate the subsequent application of the model, this paper adopts the following formula for data normalization: normalized over r_{ij}.

Standardized as follows Eq. (1):

$$r_{ij} = \frac{x_{ij} - min(x_j)}{max(x_j) - min(x_j)} \quad (1)$$

Where $max(x_j)$ - Maximum value of sample single indicator data;

$min(x_j)$-Sample single-indicator data minimum.

b) Calculating information entropy E_j

$$E_j = -\frac{1}{lnm}\sum_{i=1}^{m} p_{ij} lnp_{ij} \tag{2}$$

In Eq. (2) and Eq. (3), *m* represents the number of calculation samples, in this paper the calculation sample is 5; p_{ij} computing the median information entropy.

$$p_{ij} = \frac{r_{ij}}{\sum_{j=1}^{n} r_{ij}} \tag{3}$$

c) Calculation of weightsβ_i（ω）, Eq. (4):

$$\beta_i = \frac{1 - E_j}{\sum_{j=1}^{n}\left(1 - E_j\right)} \tag{4}$$

2) TOPSIS model: The TOPSIS technique, a prevalent multi-attribute decision-making approach, evaluates and ranks alternatives by measuring the distance of each criterion from both the ideal and the nadir solutions, thereby gauging their merits and demerits. When applied to the assessment of teaching quality in smart classrooms, each criterion's value is juxtaposed against the best and worst possible outcomes. Scores are computed based on the proximity of each alternative to the ideal and the remoteness from the nadir solution, with rankings established accordingly. The ideal solution embodies the optimal values across all criteria, whereas the nadir solution reflects the least favorable outcomes. Consequently, TOPSIS efficiently discerns the superior teaching quality strategy, namely, the one that most closely approximates the ideal and is maximally distant from the nadir solution.

Since its introduction by Hwang and Yoon in 1981, the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) has gained extensive application across diverse domains, including decision-making, supply chain management, and investment evaluation [20], [21]. The method's fundamental concept involves assessing the strengths and weaknesses of various scenarios by comparing them to two benchmarks: the ideal solution and the negative ideal solution. The ideal solution signifies the optimal outcome that maximizes benefit-oriented metrics or minimizes cost-related ones, whereas the negative ideal solution denotes the least favorable outcome across all metrics. TOPSIS provides an intuitive ranking by quantifying the distance of each alternative from these two reference points.

While TOPSIS excels in comprehensive assessments, it faces a limitation in weight determination, which often depends on the subjective input of experts or decision-makers, potentially compromising the objectivity of the results. To address this, the current study employs the entropy weight method to objectively ascertain the weight of each evaluation criterion. Grounded in information entropy principles, this method automatically computes weights based on the inherent data distribution, thereby minimizing human bias. The integration of the entropy weight method with TOPSIS enhances the precision of teaching quality evaluation in smart classrooms, offering a more robust and objective basis for assessment [22]. As the indicator data have been standardized during the entropy weight calculation, there is no need for re-standardization in the TOPSIS process. The subsequent evaluation steps, leveraging the entropy-determined weights, are outlined as follows Eq. (5) to Eq. (11):

a) Calculate the weighted data matrix

$$e_{ij} = \omega_j r_{ij} \tag{5}$$

b) Calculate the distance between the weighting matrix and the most value

After processing you can form a data matrix

$$R = \left(e_{ij}\right)_{m\times n} \tag{6}$$

Define the maximum value of each indicator, i.e., each column, ase$_j^+$

$$e_j^+ = max\left(e_{1j}\cdots e_{nj}\right) \tag{7}$$

Define the maximum value of each indicator, i.e., each column, ase$_j^-$

$$e_j^- = max\left(e_{1j}\cdots e_{nj}\right) \tag{8}$$

Define the distance of the ith object from the maximum value as d_i^+

$$d_i^+ = \sqrt{\sum_{j=1}^{n}\left(e_j^+ - r_{ij}\right)^2} \tag{9}$$

Define the distance of the ith object from the maximum value as d_i^-

$$d_i^- = \sqrt{\sum_{j=1}^{n}\left(e_j^- - r_{ij}\right)^2} \tag{10}$$

c) Calculation of scores

$$Score_i = \frac{d_i^-}{d_i^- + d_i^+} \tag{11}$$

III. Results

A. Data Sources

Given that the assessment of "intelligent learning environment" teaching quality emphasizes feedback and preparation, the data utilized in this study are primarily derived from statistical information and gathered via questionnaires and other methods [23]. Specifically, in the evaluation process, indicators pertaining to teaching preparation are assessed by specially appointed educators who review the lesson plans of the teachers under scrutiny; the teaching process is evaluated through feedback from both teachers and students; the teaching effectiveness is primarily determined by student feedback data; and the teaching reflection component is scored by relevant

school staff. To effectively conduct the case study, this paper selects the "intelligent learning environment" teaching practices of five teachers as a sample for analysis. The teaching quality evaluation results are presented on a scale of [0, 1], with 1 indicating the highest quality and 0 the lowest. Utilizing this scale, the paper ranks the teaching quality, conducts an in-depth analysis of the strengths and weaknesses, and subsequently proposes recommendations for enhancing teaching methodologies and practices.

In conducting the questionnaire, a scale of 0-10 was used to ensure a more detailed and comprehensive evaluation result, taking into account the subjective perception of the ratings. Eventually, the results obtained from the survey will be presented in the form of a table, as shown in Table I. This method not only helps to more accurately assess the teaching effect of the smart classroom, but also can provide specific data support and decision-making basis for teaching improvement.

TABLE I SURVEY DATA

	Teacher 1	Teacher 2	Teacher 3	Teacher 4	Teachers 5
Instructional Objective Design	7.2	8.1	7.3	8.4	8.2
Teaching Scenario Planning	8.1	8.3	6.4	7.5	8.2
Teaching Material Preparation	6.3	7.4	9.2	8.1	7.3
Digital Module Development	7.1	7.3	8.2	7.4	7.2
Classroom Engagement	8.2	8.4	8.3	7.6	7.5
Knowledge Expansion Skills	8.1	8.3	9.1	7.4	8.3
Digital Resource Utilization	7.2	8.1	8.3	8.2	8.1
Technological Integration	6.2	7.3	6.4	8.2	8.1
Objective Achievement	8.1	6.3	7.4	8.3	7.2
Student Knowledge Acquisition	8.2	7.3	7.4	7.5	7.6
Student Issue Feedback	7.2	7.3	6.4	7.5	6.3
After-School Task Completion	7.3	7.4	6.5	7.6	8.2
Pedagogical Improvement Measures	6.3	7.4	7.5	8.2	7.3
Smart Classroom Training	7.2	8.1	6.4	7.5	7.4
Teaching Research Activities	8.1	7.4	7.5	6.3	7.4
Inter-Classroom Exchange	8.2	7.4	7.5	7.6	6.4

Table I shows the results of the evaluation of the quality of teaching in the "smart classroom" collected through the questionnaire survey, covering the performance of different teachers in various aspects of teaching. Specifically, the data in the table reflect the scores of each indicator, assessing the quality of teaching preparation, teaching process, teaching effectiveness and teaching reflection.

In terms of teaching preparation, ratings were based on the completeness of teachers' lesson plans, the use of teaching resources and the reasonableness of course design. The quality of teachers' preparation directly affects the effectiveness of classroom teaching, so this indicator usually receives a higher rating. Teaching process indicators are based on feedback from teachers and students, examining classroom interaction, application of teaching methods and student participation. The scores for this indicator usually show some fluctuation, reflecting the differences in actual teaching by different teachers.

Teaching effectiveness is assessed primarily through student feedback, measuring student learning outcomes, knowledge acquisition, and classroom satisfaction. Students' subjective evaluation plays an important role in this section, so the scoring of this part is more sensitive and easily influenced by the classroom atmosphere and teaching methods. Finally, the Teaching Reflection section was scored by the school personnel, which mainly assessed the teachers' ability to self-reflect on their own teaching process and their awareness of improvement. The scores of this section reflect the teachers' ability for self-improvement and continuous development in the later stages of teaching.

B. Determination of Indicator Weights Based on Entropy Weighting Method

In this research, the weights of the indicators were ascertained through Eq. (1) to Eq. (3). The data for these indicators were sourced from the survey detailed in Section III, with weights allocated according to empirical data. This approach benefits from the use of statistical data, allowing for an expandable sample size that enhances the objectivity and precision of weight distribution. As data volume grows, indicator weights stabilize, more accurately depicting each indicator's significance in evaluating "smart classroom" teaching quality, thus furnishing robust data support for ensuing evaluations.

The study presents standardized data, intermediate entropy weight method calculations, information entropy values, and final weight outcomes in Tables II, III, and IV, respectively. These tables elucidate the data processing and corresponding values, ensuring the transparency and reproducibility of the weight calculation. They offer a meticulous mathematical foundation for weight determination and systematic data support for subsequent teaching quality assessments and enhancements.

TABLE II STANDARDIZED DATA PROCESSING RESULTS

	Teacher 1	Teacher 2	Teacher 3	Teacher 4	Teachers 5
Instructional Objective Design	0.0020	1.0000	0.0020	1.0000	1.0000
Teaching Scenario Planningdesign	1.0000	1.0000	0.0020	0.5010	1.0000
Teaching Material Preparation	0.0020	0.3347	1.0000	0.6673	0.3347
Digital Module Development	0.0020	0.0020	1.0000	0.0020	0.0020
Classroom Engagement	1.0000	1.0000	1.0000	0.0020	0.0020
Knowledge Expansion Skills	0.5010	0.5010	1.0000	0.0020	0.5010
Digital Resource Utilization	0.0020	1.0000	1.0000	1.0000	1.0000
Technological Integration	0.0020	0.5010	0.0020	1.0000	1.0000
Objective Achievement	1.0000	0.0020	0.5010	1.0000	0.5010
Student Knowledge Acquisition	1.0000	0.0020	0.0020	0.0020	0.0020
Student Issue Feedback	1.0000	1.0000	0.0020	1.0000	0.0020
After-School Task Completion	0.5010	0.5010	0.0020	0.5010	1.0000
Pedagogical Improvement Measures	0.0020	0.5010	0.5010	1.0000	0.5010
Smart Classroom Training	0.5010	1.0000	0.0020	0.5010	0.5010
Teaching Research Activities	1.0000	0.5010	0.5010	0.0020	0.5010
Inter-Classroom Exchange	1.0000	0.5010	0.5010	0.5010	0.0020

TABLE III CALCULATION OF PROCESS VALUES BY ENTROPY WEIGHT METHOD

	Teacher 1	Teacher 2	Teacher 3	Teacher 4	Teachers 5
Instructional Objective Design	0.0010	0.3500	0.0010	0.3500	0.3500
Teaching Scenario Design	0.3000	0.3000	0.0010	0.1500	0.3000
Teaching Material Preparation	0.0010	0.1600	0.4500	0.3000	0.1600
Digital Module Development	0.0025	0.0025	0.9900	0.0025	0.0025
Classroom Engagement	0.3500	0.3500	0.3500	0.0010	0.0010
Knowledge Expansion Skills	0.2200	0.2200	0.4000	0.0010	0.2200
Digital Resource Utilization	0.0010	0.2600	0.2600	0.2600	0.2600
Technological Integration	0.0015	0.2500	0.0015	0.4000	0.4000
Objective Achievement	0.3500	0.0010	0.1700	0.3500	0.1700
Student Knowledge Acquisition	0.9900	0.0025	0.0025	0.0025	0.0025
Student Issue Feedback	0.3500	0.3500	0.0010	0.3500	0.0010
After-School Task Completion	0.2200	0.2200	0.0010	0.2200	0.4000
Pedagogical Improvement Measures	0.0010	0.2200	0.2200	0.4000	0.2200
Smart Classroom Training	0.2200	0.4000	0.0010	0.2200	0.2200
Teaching Research Activities	0.4000	0.2200	0.2200	0.0010	0.2200
Inter-Classroom Exchange	0.4000	0.2200	0.2200	0.2200	0.0010

TABLE IV INFORMATION ENTROPY AND WEIGHT CALCULATION RESULTS

	information entropy	weights
Instructional Objective Design	0.7000	0.0700
Teaching Scenario Design	0.8500	0.0350
Teaching Material Preparation	0.8000	0.0450
Digital Module Development	0.0400	0.2000
Classroom Engagement	0.7000	0.0700
Knowledge Expansion Skills	0.8400	0.0360
Digital Resource Utilization	0.8700	0.0290
Technological Integration	0.6700	0.0720
Objective Achievement	0.8300	0.0360
Student Knowledge Acquisition	0.0400	0.2000
Student Issue Feedback	0.7000	0.0700
After-School Task Completion	0.8400	0.0360
Pedagogical Improvement Measures	0.8400	0.0360
Smart Classroom Training	0.8400	0.0360
Teaching Research Activities	0.8400	0.0360
Inter-Classroom Exchange	0.8400	0.0360

Tables II, III, and IV illustrate the data standardization process, the median values derived from the entropy weight method, the information entropy, and the precise weight outcomes utilized in this study for assessing "smart classroom" teaching quality. A thorough analysis of these tables facilitates a more comprehensive understanding of the weight determination process and its influence on the evaluation outcomes.

Table II presents the standardized raw data results. Standardization aims to convert indicators with varying magnitudes, scales, and units into a uniform standard, ensuring comparability in subsequent analyses. The standardized data, with a mean of 0 and a standard deviation of 1, enable indicators to be compared on an equal footing. This process mitigates data bias and provides a clear, standardized input for the entropy weight method. The standardized data in Table II reveal the distribution of different indicators within the sample, offering a foundation for subsequent weight calculations.

Table III details the intermediate values in the entropy weight method calculation, including each indicator's entropy value, entropy ratio, and corresponding weight coefficients. The entropy value indicates the degree of variation among the data for each indicator; a higher entropy value suggests a more uniform distribution and thus less weight is assigned, while a lower entropy value indicates less variation and more weight is given. These intermediate values in Table III are crucial for the subsequent weight allocation, enabling the entropy weight method to objectively and reasonably assign weights to each indicator. Through these calculations, the relative significance of each indicator within the overall evaluation system can be accurately quantified, establishing a basis for the scientific and objective nature of the assessment results.

Table IV shows the final calculated weights for each indicator, combining the information entropy and the importance of each indicator. These weight values are derived by combining the entropy value of each indicator and its contribution in the overall evaluation. According to the principle of entropy weighting method, the higher weighted indicators indicate that they have more influence on the results in the evaluation of teaching quality, and vice versa, they have less influence. Through Table IV, we can see the differences in the weights of different teaching links (e.g., teaching preparation, teaching process, teaching effect, etc.), which helps us understand the role of each link in teaching quality. For example, if a link has a larger weight, it means that the performance of that link has a stronger impact on the overall assessment of the quality of teaching in the smart classroom. Accordingly, a less weighted link may have a more limited impact on the results in the actual evaluation.

Through the analyses in Tables II, III and IV, it can be seen that this study effectively solves the subjectivity and uncertainty that may exist in the evaluation of teaching quality through the combination of standardization and entropy power method. The standardization process ensures that the indicators are comparable, while the entropy weight method assigns reasonable weights to each indicator through objective data analysis. Ultimately, the calculated weights not only provide a scientific basis for the evaluation of the teaching quality of "Smart Classroom", but also provide a clear direction for subsequent teaching improvement. The data in the table show the relative importance of each teaching aspect in the evaluation of teaching quality, which enables researchers and educators to optimize and adjust the teaching activities in a more targeted way.

IV. Discussion

In this paper, TOPSIS evaluation is performed according to Eq. (4) – Eq. (11).

The weighting matrix is calculated according to Eq. (4) as shown in Table V.

TABLE V TOPSIS Weighting Matrix

	Teacher 1	Teacher 2	Teacher 3	Teacher 4	Teachers 5
Instructional Objective Design	0.0280	0.0320	0.0280	0.0320	0.0320
Teaching Scenario Design	0.0160	0.0160	0.0120	0.0140	0.0160
Teaching Material Preparation	0.0150	0.0180	0.0230	0.0200	0.0180
Digital Module Development	0.0880	0.0880	0.1000	0.0880	0.0880
Classroom Engagement	0.0320	0.0320	0.0320	0.0280	0.0280
Knowledge Expansion Skills	0.0160	0.0160	0.0180	0.0140	0.0160
Digital Resource Utilization	0.0120	0.0140	0.0140	0.0140	0.0140
Technological Integration	0.0280	0.0320	0.0280	0.0360	0.0360
Objective Achievement	0.0180	0.0140	0.0160	0.0180	0.0160
Student Knowledge Acquisition	0.1000	0.0900	0.0900	0.0900	0.0900
Student Issue Feedback	0.0320	0.0320	0.0280	0.0320	0.0280
After-School Task Completion	0.0160	0.0160	0.0140	0.0160	0.0180
Pedagogical Improvement Measures	0.0140	0.0160	0.0160	0.0180	0.0160
Smart Classroom Training	0.0160	0.0180	0.0140	0.0160	0.0160
Teaching Research Activities	0.0180	0.0160	0.0160	0.0140	0.0160
Inter-Classroom Exchange	0.0180	0.0160	0.0160	0.0160	0.0140

Table V demonstrates the results of the teachers' overall quality evaluation through the comprehensive assessment of various indicators. The table scores each teacher's performance in teaching preparation, teaching process, teaching effectiveness and teaching reflection, and finally calculates each teacher's total score. By comparing the scores of different teachers in each dimension, it can be visualized which teacher is more outstanding in terms of teaching quality and comprehensive quality.

If a teacher scores high in several dimensions, especially in teaching effectiveness and teaching reflection, it means that the teacher has strong teaching ability and self-improvement consciousness, and has better comprehensive quality. In addition, the total scores in Table V provide a basis for assessing teachers' comprehensive quality, and teachers with higher scores usually perform better in teaching practice. By analyzing the table, it can provide data support and decision-making reference for subsequent teaching improvement and teacher training.

Calculation of the relevant defined values is shown in Table VI.

The final score was calculated as shown in Table VII.

TABLE VI Calculated Values for Relevant Data

	Teacher 1	Teacher 2	Teacher 3	Teacher 4	Teachers 5
Distance to Ideal Solution (d+)	0.0200	0.0220	0.0200	0.0210	0.0225
Distance to Negative Ideal Solution (d-)	0.0180	0.0130	0.0170	0.0150	0.0135

TABLE VII Evaluation Results

	Appraise value
Teacher 1	0.0200
Teacher 2	0.0220
Teacher 3	0.0200

Teacher 4	0.0210
Teacher 5	0.0225

Table VII shows the ratings of different teachers on each evaluation dimension (e.g., teaching preparation, teaching process, teaching effectiveness, teaching reflection, etc.) and their standardized data. The standardization process ensures that the indicators are compared on the same scale, providing a fair and transparent basis for subsequent weighting calculations. The data allow for the identification of differences in performance across teachers in different aspects of teaching and learning. For example, certain teachers may have higher standardized scores on preparation and teaching process, indicating that they excel in lesson planning and classroom management, while others may have higher scores on teaching effectiveness and reflection, indicating that they are able to effectively promote student learning and self-improvement. The standardized data provide a reliable basis for the weighting calculation that follows.

Table VII further shows the weights of each evaluation dimension calculated based on the entropy weighting method, as well as the scores of each teacher in each dimension and the final weighted total score. This process uses the entropy weighting method to objectively assess the relative importance of each indicator, ensuring that each indicator is assigned a reasonable weight in the final evaluation. By weighting the scores of each teacher, we can visualize the comprehensive quality evaluation results of each teacher. Teachers with higher scores usually perform better in all aspects of teaching quality and their comprehensive quality is more outstanding.

For example, a teacher's high scores on teaching process and teaching effectiveness, and the top composite scores after weighting, indicate that the teacher has strong strengths in the implementation of classroom teaching and its effectiveness. Some teachers, on the other hand, may have scored low on the teaching reflection component, which reflects their deficiencies in self-assessment and improvement. The weighted scores in Table VII not only reveal teachers' strengths and weaknesses, but also provide a valuable basis for educational administrators to use in the direction of teacher training and development.

The comparative analysis of Tables VI and VII enables a more comprehensive assessment of the comprehensive quality of teachers and their teaching performance in the smart classroom. The standardized data provide a guarantee for the objectivity of the indicators, and the introduction of the entropy weighting method ensures the reasonableness and accuracy of the weighting of each dimension in the evaluation process. The final composite scores provide us with the comprehensive performance of teachers in each teaching aspect, thus helping decision makers to formulate more scientific teacher development strategies.

V. Conclusions and Shortcomings

This paper focuses on the teaching quality evaluation of smart classroom, and through constructing a scientific and reasonable evaluation index system and adopting the entropy weight method combined with TOPSIS method, it comprehensively evaluates the performance of teachers in different dimensions. Through the questionnaire survey and statistical data collection, this paper comprehensively considered the key factors of teaching preparation, teaching process, teaching effect and teaching reflection, and sought to present the comprehensive quality of teachers in multiple dimensions and angles. The research method of this paper has strong operability and practicability, and can provide a more objective and precise basis for the assessment of teaching quality in the smart classroom.

Although this paper provides a more comprehensive analysis of the evaluation of the quality of teaching in the smart classroom, there are still some shortcomings. First, the sample data comes from a single source, mainly focusing on five teachers in a particular school, and lacks broader cross-school and cross-region sample data, so the generalizability and representativeness of its conclusions are limited. Second, although the evaluation indicators cover the dimensions of teaching preparation, teaching process, teaching effect and teaching reflection, there is still room for improvement in the setting of specific indicators, and factors such as teachers' ability to educate emotions and innovative teaching methods have not been fully considered. Finally, although the entropy weighting method and TOPSIS method were adopted, subjective factors such as teachers' teaching style and classroom atmosphere were not fully included in the analysis, which may have a certain impact on the results.

Future research can be expanded in the following directions: first, the sample size can be increased to cover teachers from more schools and districts to improve the generalizability of the findings. Second, more aspects about teachers' teaching innovativeness and affective teaching can be introduced into the evaluation indexes to comprehensively assess teachers' teaching quality. Finally, attempts can be made to combine more diversified evaluation methods, such as deep learning and artificial intelligence technology, to further improve the precision and reliability of the evaluation of teaching quality in smart classrooms.

Acknowledgment

This work is supported by Henan Philosophy and Social Sciences Planning Project "Research on the Functional Construction and Implementation of Rural Social Organizations in Rural Revitalization" (Grant No. 2021BSH025)

References

[1] Bruya, B., Ardelt, M. Wisdom can be taught: a proof-of-concept study for fostering wisdom in the classroom[J]. Learning and Instruction, 2018, 58: 106-114.

[2] Bruy, B., Ardelt M. Fostering wisdom in the classroom, part 1: A general theory of wisdom pedagogy[J]. Teaching Philosophy, 2018, 41(3): 239-253.

[3] Zhou, L., Luo, H., Kwong, X. J. Exploration of wisdom teaching mode for

international students under the new crown epidemic - taking university physics experiment as an example[J]. University Physics, 2023, 42(05): 41.

[4] Wu, X. R., Liu, B. Q., Yuan, T. T. Next-generation smart classroom: concept, platform and architecture[J]. China Electrified Education, 2019 (3): 81-88.

[5] Zhou, B. Smart classroom and multimedia network teaching platform application in college physical education teaching[J]. International Journal of Smart Home, 2016, 10(10): 145-156.

[6] Nai, R. The design of smart classroom for modern college English teaching under Internet of Things[J]. Plos one, 2022, 17(2): e0264176.

[7] Zhang, M., Li, X. Design of smart classroom system based on Internet of things technology and smart classroom[J]. Mobile Information Systems, 2021, 2021: 1-9.

[8] Zhang, X. H., Lin, C. College English smart classroom teaching model based on artificial intelligence technology in mobile information systems [J]. Mobile Information Systems 2021 (2021): 1-12.

[9] Saini, M. K., Neeraj, G. How smart are smart classrooms? A review of smart classroom technologies[J]. ACM Computing Surveys (CSUR) 52.6 (2019): 1-28.

[10] Lu, K., Yang, H. H., Shi, Y., Wang, X. Examining the key influencing factors on college students' higher-order thinking skills in the smart classroom environment[J]. International Journal of Educational Technology in Higher Education, 18(2021), 1-13.

[11] Li, L., Wang, Y. C., Ma, C. Z. The Cultivating Strategies of Pre-Service Teachers' Informatization Teaching Ability Oriented to Wisdom Generation [J]. International Journal of Emerging Technologies in Learning (iJET) 16.6 (2021): 57-71.

[12] Tong, Y. P. A Study on the Construction and Evaluation of College English Wisdom Classroom in the "Internet Plus" Era[C]. 8th International Conference on Education, Language, Art and Inter-cultural Communication (ICELAIC 2021). Atlantis Press, 2022.

[13] Ren, Y. F. Research on the Evaluation System of Teaching Effect of Smart Classroom under the Background of Big Data[J]. Journal of Higher Education, 2023, 9(25):91-94. DOI:10.19980/j.CN23-1593/G4.2023.25.023.

[14] Zhan, Z., Wu, Q., Lin, Z., Cai, J. Smart classroom environments affect teacher-student interaction: Evidence from a behavioral sequence analysis[J]. Australasian Journal of Educational Technology, 37(2), 2021 96-109.

[15] Kaur, A., Munish B., Giovanni S. A survey of smart classroom literature[J]. Education Sciences 12.2 (2022): 86.

[16] Zhu, Y. X., Tian, D. Z., Feng, Y. Effectiveness of entropy weight method in decision-making[J]. Mathematical Problems in Engineering, 2020: 1 - 5.

[17] Cong, P. J., Wang, L., Yin, Z. G., Zhang, B., Li, Y. C. Determination of Disease Risk Levels of Earth and Rock Dams Based on Combinatorial Empowerment and Topologizable Theory[J]. Journal of Water Resources and Construction Engineering (02), 2023:36-42.

[18] Wang, T. D., He, Z. R., Shan, X. F., Liu, G., Deng, Q. L., Ren, Z. G. Evaluation and analysis of photovoltaic solar thermal cooling system based on entropy weight-TOPSIS[J]. Journal of Solar Energy (09), 2023: 229-235. doi:10.19912/j.0254-0096.tynxb.2022-0676.

[19] Chakraborty, S. TOPSIS and Modified TOPSIS: A comparative analysis[J]. Decision Analytics Journal 2 (2022): 100021.

[20] Chen, P. Y. Effects of the entropy weight on TOPSIS[J]. Expert Systems with Applications 168, 2021: 114186.

[21] Salih, M. M., Zaidan, B. B., Zaidan, A. A., Ahmed, M. A. Survey on fuzzy TOPSIS state-of-the-art between 2007 and 2017[J]. Computers & Operations Research, 104, 2019:207-227.

[22] Li, Z., Luo, Z., Wang, Y., Fan, G., Zhang, J. Suitability evaluation system for the shallow geothermal energy implementation in region by Entropy Weight Method and TOPSIS method[J]. Renewable Energy Renewable Energy, 184, 2022:564-576.

[23] Wu, H. W., Li, E. Q., Sun, Y. Y., Dong, B. T. Research on the operation safety evaluation of urban rail stations based on the improved TOPSIS method and entropy weight method[J]. Journal of Rail Transport Planning & Management, 20, 2021:100262.

Long-Term Recommendation Model for Online Education Systems: A Deep Reinforcement Learning Approach

Wei Wang*
Xianyang Normal University, Xianyang Shaanxi, 712000 Shaanxi, China

***Abstract*—Intelligent tutoring systems serve as tools capable of providing personalized learning experiences, with their efficacy significantly contingent upon the performance of recommendation models. For long-term instructional plans, these systems necessitate the provision of highly accurate, enduring recommendations. However, numerous existing recommendation models adopt a static perspective, disregarding the sequential decision-making nature of recommendations, rendering them often incapable of adapting to novel contexts. While some recent studies have delved into sequential recommendations, their emphasis predominantly centers on short-term predictions, neglecting the objectives of long-term recommendations. To surmount these challenges, this paper introduces a novel recommendation approach based on deep reinforcement learning. We conceptualize the recommendation process as a Markov Decision Process, employing recurrent neural networks to simulate the interaction between the recommender system and the students. Test results demonstrate that our model not only significantly surpasses traditional Top-N methods in hit rate and NDCG concerning the enhancement of long-term recommendations but also adeptly addresses scenarios involving cold starts. Thus, this model presents a new avenue for enhancing the performance of intelligent tutoring systems.**

***Keywords*—Deep reinforcement learning; long-term recommendation; intelligent tutoring system; Markov Decision Process; recurrent neural network**

I. Introduction

In the current educational landscape, personalized learning is increasingly gaining prominence. Nevertheless, traditional educational approaches often employ static teaching paradigms, overlooking the dynamic and sequential nature of students' learning progress and personalized needs. This oversight may lead to uneven allocation of educational resources, as certain students, regardless of their learning trajectories, might receive similar instructional resources and methods. Additionally, these methods are susceptible to the impact of students' aptitude issues, wherein newly enrolled students may struggle to receive precise personalized recommendations due to a lack of historical learning records [1].

Indeed, education should be construed as a sequential decision-making process, and adaptability is crucial for intelligent tutoring systems, given that students' learning progress and needs invariably evolve over time. To integrate the capability for sequential processing into intelligent tutoring systems, recurrent neural networks (RNNs) have recently been introduced into educational systems. However, the majority of existing sequential learning methods are applicable only to short-term predictions, disregarding predictions for long-term learning. Furthermore, these RNN-based sequence models entirely overlook the interaction between intelligent tutoring systems and students, a pivotal component of interactive reinforcement learning [2].

To address the aforementioned issues, this paper proposes a novel model based on deep reinforcement learning (DRL) for long-term learning prediction. Specifically, the model employs RNN to adaptively evolve the student's learning state to simulate the sequential interaction between the student and the intelligent tutoring system. The model is applicable to cold start scenarios and utilizes an additional gated neural network to balance the influence between the state of the RNN and the historical state derived from the student's learning records [3].

To maximize the expected long-term learning outcomes and optimize model parameters, we present an effective learning approach based on the popular policy gradient algorithm REINFORCE. Extensive experiments conducted on two real-world datasets demonstrate the commendable performance of our proposed model in cold start scenarios, surpassing the current state-of-the-art methods in various learning performance metrics [4].

The remaining part of the paper is organized as follows. Section II provides a literature review on recommender systems in intelligent tutoring systems and deep reinforcement learning, Section III presents the overall framework and detailed mechanisms of the proposed long-term recommendation model based on deep reinforcement learning, Section IV describes the environment and interaction processes within the model, Section V discusses the recommendation agent and its training methodologies, Model training is given in Section VI. Section VII reports on experiments that validate the performance of the proposed model in long-term recommendations, and Section VIII concludes the paper by summarizing the findings and discussing the implications of the deep reinforcement learning approach for intelligent tutoring systems.

II. Related Work

A. Recommender Systems

In Intelligent Tutoring Systems (ITS), the role of recommender systems is to furnish learners with personalized

recommendations, aiding them in selecting suitable learning resources based on their learning progress and comprehension.

Traditional recommender systems, such as those based on Collaborative Filtering (CF) methods [6], Matrix Factorization (MF) methods [5], and neural network-based approaches [7], prove highly effective when recommending static content (e.g., textbooks, videos, exercises). However, these methods often rely on a substantial volume of historical interaction data and frequently grapple with the cold start problem when confronted with new students or content. Moreover, these approaches tend to overlook the dynamism and sequential nature of the learning process, wherein the learner's knowledge state gradually evolves over the course of learning.

In handling sequential data, recommender systems based on Recurrent Neural Networks (RNNs) have made strides. Hidasi et al. [8] employed RNNs for predicting students' learning paths, and other researchers have proposed several RNN-based enhancements [9]. Nevertheless, these methods primarily focus on short-term predictions, neglecting the long-term learning trajectories of students.

In Interactive Recommender Systems (IRS), student feedback is incorporated into the model. Such models can iteratively construct and optimize representations of students and content, thereby holding an advantage in addressing the cold start problem. Some researchers have utilized Multi-Arm Bandit (MAB) methods to build IRS, yet these methods primarily address the exploration-exploitation dilemma and do not explicitly optimize for long-term returns.

B. Deep Reinforcement Learning

Deep Reinforcement Learning (DRL) has made significant breakthroughs in many interactive systems, such as Atari games [10] and Go [11], but its application in recommender systems remains relatively limited. Wang et al. [12] proposed a DRL approach based on the Actor-Critic (AC) framework for page recommendations, a hybrid RL method that integrates value-based and policy-based modules. Li et al. [13] introduced a method based on Deep Q Network (DQN) for keyword prompt recommendations. Through this approach, they achieved effective recommendations in the absence of explicit student feedback. Sewak et al [14] presented a DRL-based interactive recommender system for news recommendations. This method, incorporating a memory network, better handles historical interaction data. Hernandez-Leal et al. [15] proposed a DRL method for personalized recommendations. Their approach learns the latent representation of students through deep neural networks and employs reinforcement learning for recommendation decisions. Ausin et al. [16] introduced a hybrid recommender system combining deep learning and reinforcement learning. Their method dynamically updates after each interaction and can make recommendations without student historical information. Abdelshiheed et al. [17] proposed a DRL-based recommender system to address multi-objective recommendation problems. Their method considers personalized student needs while taking into account business objectives. Koroveshi et al. [18] presented a DRL-based sequential recommendation method that predicts students' future behavior while considering both long-term and short-term interests. Jung et al. [19] introduced a DRL-based interactive recommender system to address cold start problems. Their method effectively recommends in situations where students lack historical interaction data.

The aforementioned research endeavors underscore the potential of deep reinforcement learning [22] in recommender systems, addressing a spectrum of challenges from history-based recommendations to tackling cold start problems and resolving multi-objective recommendation issues.

C. Research Gaps and Motivation

In summary, the primary challenges faced by recommender system research in intelligent tutoring systems include an overreliance on historical interaction data, neglecting the issue of students' long-term learning paths, and the difficulty in handling large action spaces. In the next section, we will introduce how we address these issues by proposing a new DRL-based recommender system framework to provide more effective personalized learning recommendations.

III. Overall Framework

Typically, reinforcement learning-based systems involve interaction between the environment and an intelligent agent [20]. During the training process, the parameters within the intelligent agent are optimized based on rewards generated from the continuous interaction between the environment and the agent [21]. More specifically, this interaction comprises two consecutive steps: 1) the agent selects and executes an action based on the environment's state; 2) the environment responds to the action performed by the agent and returns feedback and a reward.

In the recommender system scenario considered in this study, the environment consists of various students, and the intelligent agent is a recommendation model based on RNN. Actions correspond to generating Top-N recommendation lists for specific students, and feedback indicates whether the student accepts this recommendation list. The entire recommendation process is illustrated in Fig. 1. It can be observed that for each individual student, there is a corresponding recommendation agent. Notably, all recommendation agents share the same network parameters. This allocation strategy for recommendation agents has its advantages as it can prevent mutual interference from different students.

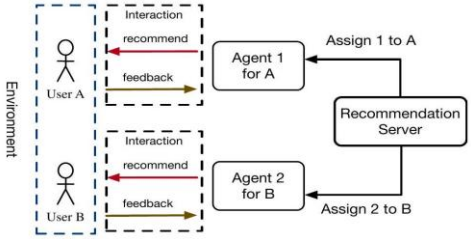


Fig. 1. The entire recommendation process.

IV. Environment and Interaction

The overall environment in this paper is constructed through offline datasets, such as the Secondary_school_curriculum, as online environments are not always feasible. Typically, offline datasets consist of a student-

course rating matrix $\tilde{R} \in \mathbb{R}^{U*M}$, where U represents the students and M denotes the courses. The elements $\tilde{R}_{u,i}$ in $\tilde{R}$ signify the rating given by student u to course i. Specifically, the explicit rating matrix $\tilde{R}$ can be transformed into an implicit feedback matrix F, where the element $F_{u,i}$ indicates whether student u is associated with course i.

$$F_{u,i} = \begin{cases} 1 & \tilde{R}_{u,i} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Based on the implicit feedback matrix F, it is straightforward to derive the set of courses I_u that interest student u. The courses in the set are sorted by timestamp, and for each course $i \in I_u$, the value of $F_{u,i}$ must be 1.

To obtain feedback $f_{u,t}$, the student response function $\mathbb{V}(P^N_{u,t}, I_u)$ used in Eq. (1) can be defined as Eq. (2).

$$f_{u,t} = \mathbb{F}(P^N_{u,t}, I_u) = \begin{cases} 1 & P^N_{u,t} \cap I_u \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Here, $f_{u,t}$ being 1 indicates positive feedback, and 0 denotes negative feedback. $P^N_{u,t} \cap H_u \neq \emptyset$ signifies that $P^N_{u,t}$ successfully includes at least one course liked by the student. In practice, positive feedback can refer to student actions such as clicks or purchases. Similarly, negative feedback refers to students ignoring the recommended list or clicking on courses outside the recommended list. The system automatically provides feedback when the student performs any of these actions, without requiring the student to provide real-time feedback.

During the testing phase, the long-term recommendation performance of a student is the average result obtained over all steps in the corresponding interaction sequence. More intuitively, the overall performance of the recommender system can be evaluated using recall-based metrics, such as hit rate, and precision-based metrics, such as NDCG. These can be calculated using Eq. (3) and Eq. (7), respectively.

$$hit@N = \frac{\Sigma_u \frac{1}{|I_u|} \Sigma_{t=1}^{|I_u|} f_{u,t}}{\#user} \tag{3}$$

.

$$NDCG@N = \frac{\Sigma_u \frac{1}{|I_u|} \Sigma_{t=1}^{|I_u|} \frac{DCG@N(P^N_{u,t})}{iDCG@N}}{\#user}. \tag{4}$$

$$DCG@N(P^N_{u,t}) = \sum_{i=1}^{|K|} \frac{h_i}{\log_2 1 + i}. \tag{5}$$

$$h_i = \begin{cases} 1 & P^N_{u,t,i} \text{ inI} \\ 0 & \text{otherwise.} \end{cases} . \tag{6}$$

$$iDCG@N = DCG@N \binom{N}{u,t} \tag{7}$$

where $^N_{u,t}$ represents the recommended sequence, which includes courses of interest to the student that have not been recommended previously.

V. Recommendation Agent

Within the recommendation agent, the recommendation process is viewed as a Markov Decision Process (MDP), providing a more suitable framework for recommender systems due to its consideration of the long-term impact of each recommendation and the corresponding expected values. Fig. 2 shows the diagram of the model.

Assuming $\pi(i \mid s_{u,t})$ represents the probability of recommending course i given the student's state $s_{u,t}$, the generation function $\mathbb{G}$ can be defined as follows:

$$P^N_{u,t} = \mathbb{G}(s_{u,t}, I) = \text{Top}_{i \in I}\, N(\pi(i \mid s_{u,t}) \times m^t_{u,i}) \tag{8}$$

where, $\pi(i \mid s_{u,t})$ is the recommendation probability of course i at time t, and $m^t_{u,i}$ is the element of course i in the masking vector m^t_u. The value of $m^t_{u,i}$ is 0 or 1, indicating whether the student has previously selected that course.

At time t, the recommendation probability $\pi(i \mid s_{u,t})$ for course i can be obtained as:

$$\pi(i \mid s_{u,t}) = \text{Softmax}(o^i_{u,t}) \tag{9}$$

where, $s_{u,t}$ is a obtained l-dimensional vector (which will be discussed later), and $o^i_{u,t}$ is the i-th element in the vector $o_{u,t}$ defined as:

$$o_{u,t} = \hat{\sigma}(W_s s_{u,t} + b_s). \tag{10}$$

where $\hat{\sigma}$ is the ReLU activation function, $W_s \in \mathbb{R}^{M \times l}$ and $b_s \in \mathbb{R}^M$ are the parameter matrix and bias, respectively. The softmax function is defined as:

$$\text{Softmax}(o^i_{u,t}) = \frac{\exp^{o^i_{u,t}}}{\sum_{k \in I} \exp^{o^k_{u,t}}} \tag{11}$$

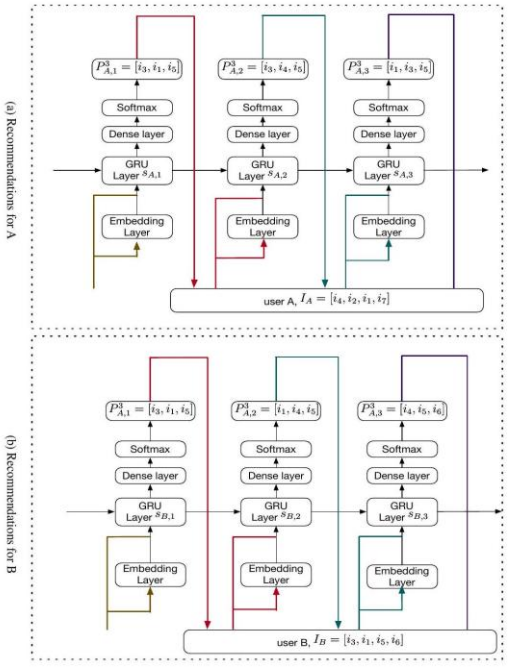


Fig. 2. Model.

In the formula, $s_{u,t}$ is the state of student u at time t, obtained through the following state transition process:

$$s_{u,t} = \mathbb{T}\left(s_{u,t-1}, f_{u,t-1}, P^N_{u,t-1}\right) = S\left(s_{u,t-1}, \hat{a}_{u,t-1}\right) \tag{12}$$

where, $P^N_{u,t-1}$ is the recommended list, $f_{u,t-1}$ is the student's feedback, S is the internal state transition process, and $\hat{a}_{u,t-1}$ is the action used to represent $P^N_{u,t-1}$ and $f_{u,t-1}$ defined as:

$$\hat{a}_{u,t-1} = \underset{a \in A_{u,t-1}}{\operatorname{argmax}}\left(\pi\left(a \mid s_{u,t-1}\right) \times m^{t-1}_{u,i}\right). \tag{13}$$

where, $A_{u,t-1}$ is the auxiliary set for feedback in different situations, defined as:

$$A_{u,t-1} = \begin{cases} P^N_{u,t-1} \cap I_u, & f_{u,t-1} > 0 \\ P^N_{u,t-1}, & \text{otherwise.} \end{cases} \tag{14}$$

The internal state transition process $s_{u,t} = S\left(s_{u,t-1}, \hat{a}_{u,t-1}\right)$ is obtained through an RNN with Gated Recurrent Unit (GRU).

Firstly, $\hat{a}_{u,t-1}$ should be transformed into the input of RNN$x_{u,t}$ at time t. To more effectively incorporate feedback into RNN, the input $x_{u,t}$ is obtained through the following formula:

$$x_{u,t} = E\left(\hat{a}_{u,t-1}\right) = \begin{cases} \hat{e}\left(\hat{a}_{u,t}\right), & f_{u,t} > 0 \\ -\hat{e}\left(\hat{a}_{u,t}\right), & \text{otherwise.} \end{cases} \tag{15}$$

where, $\hat{e}\left(\hat{a}_{u,t}\right) \in \mathbb{R}^{\hat{l}}$ is the l-dimensional embedding of course $\hat{a}_{u,t}$ also a part that needs to be learned in the proposed model. Using E, positive and negative feedback can be clearly distinguished for a given $\hat{a}_{u,t}$.

According to the state transition of GRU, $s_{u,t} = S\left(s_{u,t-1}, \hat{a}_{u,t-1}\right)$ can be obtained as follows:

$$\begin{aligned} s_{u,t} = S\left(s_{u,t-1}, \hat{a}_{u,t-1}\right) &= \left(1 - Z\left(x_{u,t}, s_{u,t-1}\right)\right) \odot s_{u,t-1} \\ &+ Z\left(x_{u,t}, s_{u,t-1}\right) \odot \tilde{S}\left(x_{u,t}, s_{u,t-1}\right). \end{aligned} \tag{16}$$

where, $\odot$ represents the element-wise product, Z is the function of the update gate, $\tilde{S}$ is the function generating candidate states, obtained through the following formulas:

$$Z\left(x_{u,t}, s_{u,t-1}\right) = \sigma\left(W_z x_{u,t} + U_z s_{u,t-1}\right) \tag{17}$$

$$\tilde{S}\left(x_{u,t}, s_{u,t-1}\right) = \tanh\left(W x_{u,t} + U\left(j_{u,t} \odot s_{u,t-1}\right)\right) \tag{18}$$

where, σ is the sigmoid activation function, $W_z \in \mathbb{R}^{l \times \hat{l}}, U_z \in \mathbb{R}^{l \times l}, W \in \mathbb{R}^{l \times \hat{l}}$ and $U \in \mathbb{R}^{l \times l}$ are parameter matrices, and $j_{u,t}$ is the reset gate, obtained through the following formula:

$$j_{u,t} = \sigma\left(W_j x_{u,t} + U_j s_{u,t-1}\right) \tag{19}$$

where, $W_j \in \mathbb{R}^{l \times \hat{l}}$ and $U_j \in \mathbb{R}^{l \times l}$ are parameter matrices.

In the warm-start model, there is an additional component used to merge student historical information.

Assuming $\hat{I}_u$ is the set containing all historical courses related to student u (it should be noted that, $\hat{I}_u \cap I_u = \emptyset$, and if $i \in \hat{I}_u$, then $m^0_{u,i} = 0$), i.e., courses from the student's history cannot be selected during the interaction.

In particular, the latent vector h_u representing student u's historical items is obtained through the following formula:

$$h_u = \tanh\left(\sum_{i \in \hat{I}_u} e(i)\right) \tag{20}$$

where, $e(i) \in \mathbb{R}^{\hat{l}}$ is another l-dimensional embedding defining course i, which also needs to be learned and is different from $\hat{e}_i$.

Additionally, to adaptively balance the effects of h_u and $s_{u,*}$ an External Memory Gated Recurrent Unit (EMGRU) is proposed, with detailed descriptions as follows.

Firstly, the state $s_{u,t}$, is obtained according to the formula, then, a new integrated state $\hat{s}_{u,t}$ used to generate course selection probabilities is obtained through the formula:

$$\hat{s}_{u,t} = d_{u,t} \odot s_{u,t} + \left(1 - d_{u,t}\right) \odot h_u \tag{21}$$

where, $d_{u,t}$ is the balance gate, which can control the impact of static h_u and dynamic $s_{u,t}$.

The balance gate $d_{u,t}$ can be obtained through the following formula:

$$d_{u,t} = \sigma\left(W_d h_u + U_d s_{u,t}\right) \tag{22}$$

where, $W_d \in \mathbb{R}^{l \times \hat{l}}$ and $U_d \in \mathbb{R}^{l \times l}$ are parameter matrices.

It is noteworthy that h_u or $\hat{s}_{u,t}$ does not affect the transition process $\mathbb{T}$ of state $s_{u,t}$ and h_u only affects the generation of course selection probabilities. In other words, the static and dynamic branches are independent of each other. In summary, the specific structure of the EMGRU unit is shown in Fig. 3.

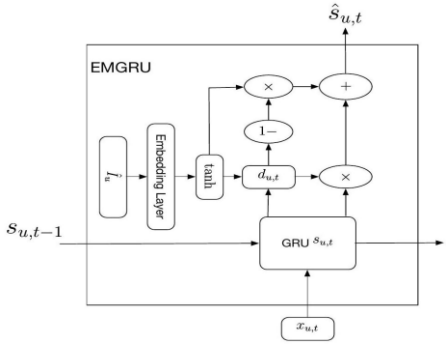


Fig. 3. EMGRU unit.

Finally, the selection probability $\pi\left(i \mid \hat{s}_{u,t}\right)$ is obtained through $\hat{s}_{u,t}$ rather than the state, $s_{u,t}$ and thus the course selection probability is defined as follows:

$$\pi\left(i \mid \hat{s}_{u,t}\right) = \text{Softmax}\left(\hat{o}^i_{u,t}\right) \tag{23}$$

where, $\hat{o}_{u,t} = \hat{\sigma}\left(W_s \hat{s}_{u,t} + b_s\right)$.

The parameter set of the warm-start model is $\hat{\theta} = \tilde{\theta} \cup \{e(*), W_d, U_d\}$. The overall architecture of the warm-start model is shown in Fig. 4.

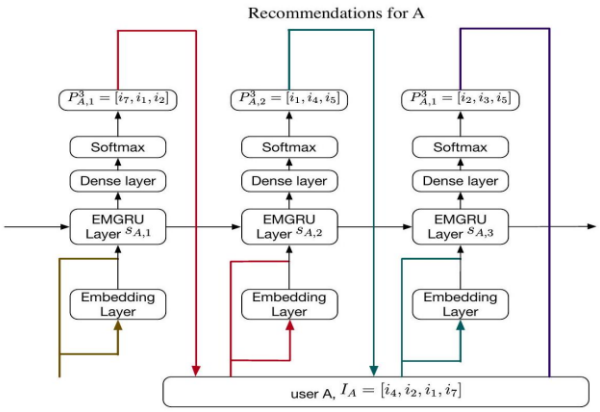


Fig. 4. The overall architecture of the warm-start model.

VI. Model Training

In this section, we will describe how to train the proposed model in the sequential interaction between the recommendation agent and the environment.

A. Reinforcement Learning

The goal of the learning algorithm in this section is to maximize the expected long-term recommendation reward, where θ is learned through the interaction process E_u for each student u. Specifically, E_u represents the complete interaction process obtained by the recommendation agent for student u under the current parameters.

Generally, an E_u interaction process includes the immediate reward $V_{u,t}$ at time t, state $s_{u,t}$, and action $\hat{a}_{u,t}$, defined as:

$$E_u = [s_{u,1}, \hat{a}_{u,1}, f_{u,1}, V_{u,1}, \dots, s_{u,M}, \hat{a}_{u,M}, f_{u,M}, V_{u,M}] \quad (24)$$

where, $s_{u,t}$ is generated by Equation (19) $\hat{a}_{u,t}$ is obtained by Equation (16) and $f_{u,t}$. Thus, $V_{u,t}$ can be computed as:

$$V_{u,t} = \begin{cases} 1.0, & f_{u,t} > 0 \\ -0.2, & \text{otherwise.} \end{cases} \quad (25)$$

where 1.0 and -0.2 are values determined based on experience.

To maximize the expected cost J, each action $\hat{a}_{u,t}$ corresponds not only to an immediate reward $V_{u,t}$, but also to a long-term reward $R_{u,t}$, computed as follows:

$$R_{u,t} = \Sigma_{k=0}^{M-k} \gamma^k V_{u,t+k} \quad (26)$$

where $\gamma \in [0,1]$ is the discount factor. The objective function J is defined as:

$$J = \mathbb{E}_{s_{u,1}, a_{u,1}, \dots}[R_{u,t}] \quad (27)$$

The parameter θ of the recommendation agent can be optimized using the gradient ascent method:

$$\theta = \theta + \eta \nabla_\theta J \quad (28)$$

where η is the learning rate, and the gradient $\nabla_\theta J(\theta)$ is given by:

$$\nabla_\theta J = \sum_{t=1}^{M} \gamma^{t-1} R_{u,t} \nabla_\theta \log \pi(\hat{a}_t \mid s_{u,t}) \quad (29)$$

In the standard practice of applying REINFORCE, the interaction process should be a complete E_u, which means the parameters should be updated after completing the interaction process for student u. However, the lengths of student interactions $I*$ can vary significantly. For example, $|I_A| = 20$, but $|I_B| = 200$. This causes large variances infor different students at the same time t. Moreover, due to the accumulation of excessive negative rewards, long-term interaction processes can hide positive results, making it challenging for the recommendation agent to obtain positive training samples. Therefore, a recommendation agent trained using traditional REINFORCE learning cannot achieve satisfactory recommendation performance.

To address this issue, this paper proposes to divide the original E into I_u/B sub-interaction processes and restart the reward accumulation at the beginning of each sub-interaction process. The learning process for both $\hat{\theta}$ and θ remains the same. The detailed processes of the learning process and the sub-interaction process generation are shown in Algorithm (1) and Algorithm (2), respectively. It should be noted that in the warm-start scenario, to ensure sufficient training data, the total length of the interaction process for each student u in the training phase is equal to $|I_u \cup \hat{I}_u|$. At the same time, the recommendation agent can choose courses from $\hat{I}_u$. However, the length of the interaction process is still equal to $|I_u|$, and during the testing phase, the recommendation agent cannot select courses from $\hat{I}_u$ for each student u.

B. Supervised Learning

Another approach to training the recommendation agent is to optimize it using supervised learning in a short-term prediction scenario and then apply it to a long-term testing environment.

To facilitate the transition of the recommendation agent from short-term to long-term prediction scenarios, the neural network architecture for the recommendation agent under supervised learning and reinforcement learning should be consistent. The only difference is that explicit labels need to be provided for supervised learning, and these labels are the actual course selections made by students at each time step.

Let I_u denote the actual sequence of courses chosen by student u over time. To maximize the accuracy of short-term predictions, this paper uses cross-entropy $\hat{J}$, defined as the cost function for supervised learning:

$$\hat{J} = -\Sigma_{i=1}^{B} \log\left(\pi(I_{u,i} \mid s_{u,i})\right) \quad (30)$$

where θ can be updated as follows:

$$\theta = \theta - \eta \nabla_\theta \hat{J} \quad (31)$$

Compared to the reinforcement learning approach, the supervised learning method is closer to traditional session-

based RNNs, where each recommended step has a specific corresponding label for the training signal.

Furthermore, before applying reinforcement learning, fine-tuning the recommendation agent through supervised learning can be performed. This can help the reinforcement learning-based recommendation agent start from a relatively good policy rather than a random one, thus accelerating the convergence speed of reinforcement learning.

VII. EXPERIMENTS

In this section, a substantial number of experiments were conducted to demonstrate the advantages of the proposed method in long-term recommendations and showcase the effectiveness of the core components of the proposed model.

A. Evaluation Datasets and Experimental Settings

Two offline real-world benchmarks, Secondary_school_curriculum 100K and Secondary_school_curriculum 1M, were used to evaluate the proposed model. Secondary_school_curriculum 100K contains 100,000 rating records about 943 students and 1682 courses. Secondary_school_curriculum 1M includes one million rating records about 6040 students and 3900 courses.

The proposed model was evaluated using the previously mentioned interaction environment to assess the performance of the proposed method and other methods. In the experiments, a 1-layer RNN and GRU with a hidden layer size of 100, matching the embedding size, are used. The size of B is set to 20, and γ is set to 0.9. For experiments on the 100K and 1M datasets, the proposed model is optimized using Adam with a learning rate of 0.005.

B. Benchmark

Pop: This algorithm always recommends the most popular items in the training set. While simple, it often serves as a powerful baseline.

Linear-UCB (L-UCB): A linear bandit algorithm, a widely used and mature multi-armed bandit algorithm. In this study, a context-independent bandit algorithm was used since content information was not considered. Course embeddings were obtained through matrix factorization (MF) methods.

ε-greedy: Similar to Linear UCB, but the balance between exploration and exploitation is adjusted by tuning the ε parameter. Course embeddings were also obtained through MF.

DQN (Deep Q-Network): A deep reinforcement learning algorithm based on value functions.

SARSA (State-Action-Reward-State-Action): A classic policy-based reinforcement learning algorithm commonly used as a benchmark.

Actor-Critic: A deep reinforcement learning algorithm that combines value function and policy methods. The architecture proposed in was used for comparison.

PPO (Proximal Policy Optimization): An advanced reinforcement learning algorithm that improves training stability by adding an "agent" constraint during policy updates.

Among these algorithms, Pop, Linear-UCB, and ε-greedy are traditional recommender systems or multi-armed bandit algorithms, while DQN, SARSA, Actor-Critic, and PPO are classical or advanced algorithms in the field of reinforcement learning.

C. Warm-start Model Comparative Experiment

In the experiment, a comparative study of warm-start models was conducted on the 100K, and the results are shown in Table I. In this table, $p = 10\%$ means that 10% of courses for each student u are retained as the warm-start historical set $\hat{I}_u$, and five different p values are considered in the experiment. Specifically, the historical data $\hat{I}_u$ for each student u in the test set was also used to train static models like BPR and NeuCF because these static models need to obtain corresponding student representations and are evaluated without using $\hat{I}_u$ during testing. However, the historical data of test students was not used to train other models.

TABLE I. NDCG@10 COMPARISON OF THE WARM-START MODELS ON 100K DATASET

	p=10%	p=30%	p=50%	p=70%	p=90%
Pop	2.90%	2.16%	1.65%	1.31%	1.48%
BPR	3.41%	3.25%	2.89%	2.74%	3.23%
NeuCF	3.51%	3.33%	3.02%	2.86%	3.35%
sRNN	8.54%	6.92%	5.45%	4.05%	3.66%
sl-cold	8.97%	6.76%	5.47%	3.96%	3.40%
rl-cold	14.93%	12.11%	8.88%	1.24%	3.69%
sl+rl-cold	15.54%	12.53%	9.38%	6.32%	3.71%
sl-warm	8.65%	6.79%	5.37%	4.73%	4.42%
rl-warm	13.81%	11.53%	8.37%	6.76%	4.44%
sl+rl-warm	14.34%	12.07%	9.64%	7.90%	6.18%

The proposed method's large-scale warm-start model significantly outperforms baseline models. With an increase in p, the warm-start model demonstrates significant improvements in HR and NDCG. These results suggest that incorporating historical data can enhance the model's performance if the historical data is sufficiently rich. It is noted that experiments with larger p values are more challenging than those with smaller p values because the total number of $I_u \cup \hat{I}_u$ is fixed. Specifically, larger p values result in a smaller correct candidate set I_u and a shorter reasoning process. Therefore, the proposed model's performance is relatively better at smaller p values. Additionally, training the recommendation agent through supervised learning generally improves HR and NDCG performance. Furthermore, this is particularly effective for the larger 1M dataset. Without supervised pre-training, the rl-warm model cannot even outperform sRNN at p=90%. Conversely, the gap between sRNN and sl+rl-warm is significant.

D. Long-Term Prediction Performance Comparison

To validate the performance of the proposed model in long-term recommendations, Fig. 5 presents the recommendation results of various comparative methods at different stages. The experimental results are obtained from selected students in the

respective test sets. Specifically, the size of the selected student set I_u is above average because students with longer histories can more clearly reveal the performance of long-term recommendations. Additionally, the results for student u are divided into five different stages: [0%,20%), [20%,40%), [40%,60%), [60%,80%), and [80%,100%], where $[s\%, e\%)$ denotes the steps between $(s\% \times I_u)$ and $(e\% \times I_u)$ during the entire interaction period. Furthermore, e is used to represent the range [s%,e%) and provides the average results within that range.

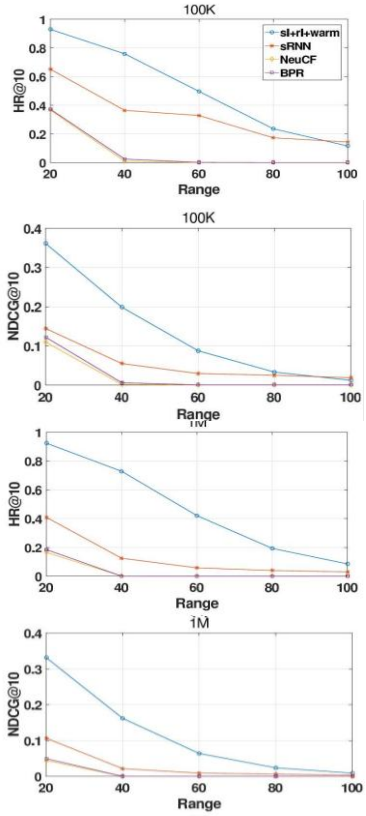


Fig. 5. The Recommendation performance at different stages. (a) HR@10 on 100 K; (b) NDCG@10 on 100 K; (c) HR@10 on 1M; (d)NDCG@10 on 1M.

As shown in Fig. 5, static models like NeuCF and BPR still achieve good hit rates in the first range [0%,20%). However, due to the almost unchanged recommendation results obtained by static methods, their performance sharply drops to 0% in the subsequent stages. Consequently, their overall HR and NDCG results are poor, as shown in Table II. On the other hand, sequential methods like sl+rl-warm and sRNN can achieve hits in all ranges, adapting their respective recommendation results. Compared to sRNN, the proposed sl+rl-warm model significantly surpasses sRNN in the ranges [0%, 20%) to [60%, 80%) because the proposed model already obtains a sufficient hit rate in the early stages. Since I_u has a fixed size, fewer courses can be hit in the last range [80%, 100%]. Therefore, the performance of sl+rl-warm is almost equivalent to that of RNN in the last range [80%, 100%]. All these results indicate that the proposed method can effectively adopt recommendation transfer in long-term recommendations.

TABLE II. TAB.6 HR@10 COMPARISON OF THE WARM-START MODELS ON 1M DATASET

	p=10%	p=30%	p=50%	p=70%	p=90%
Pop	6.83%	5.73%	5.04%	4.52%	4.10%
BPR	4.73%	4.73%	4.95%	4.99%	5.15%
NeuCF	6.27%	6.56%	6.65%	6.99%	7.78%
sRNN	16.55%	17.16%	17.27%	17.67%	18.52%
sl-cold	31.31%	26.99%	22.92%	17.82%	9.63%
rl-cold	41.65%	36.93%	32.07%	26.07%	14.73%
sl+rl-cold	45.15%	40.98%	33.24%	27.28%	15.15%
sl-warm	23.39%	23.23%	22.07%	16.89%	8.59%
rl-warm	42.23%	35.32%	29.53%	26.51%	16.09%
sl+rl-warm	45.06%	43.88%	37.14%	31.58%	21.20%

E. Impact of EMGRU

EMGRU is crucial in the warm-start model as it can adaptively adjust the dynamic RNN state and static historical representation to generate $\hat{s}_{u,t}$. To study the impact of this adaptive balance, the experiment considered four different settings combining RNN states and historical states: 1) using only RNN states (s); 2) using only historical representations (h); 3) the combination of historical representations and RNN states ($s + h$); 4) the combination of historical representations and RNN states with a balanced gate ($s + h\mathrm{w}/$).

As shown in the Table III, the method with only h performs poorly because h_u is unchangeable throughout all interaction processes, and the fixed h_u cannot generate different recommendation results at different times. On the other hand, when the minimum setting is $p = 10\%$, the method with only s surpasses the combination methods $s + h$ and $s + h\mathrm{w}/$ gate in terms of HR because the data volume of h_u is not sufficient. However, as p increases, the combination methods can outperform the method with only s. Moreover, adapting the combination of h and s with a gate generally has a better effect than mixing h and s with an equal constant. These results indicate that the effects between $s_{u,t}$ and h_u are non-fixed and should be adjusted according to the current situation.

TABLE III. NDCG@10 COMPARISON OF THE WARM-START MODELS ON 1M DATASET

	p=10%	p=30%	p=50%	p=70%	p=90%
Pop	2.03%	1.50%	1.18%	0.92%	0.81%
BPR	1.14%	1.05%	0.96%	0.84%	0.92%
NeuCF	1.74%	1.62%	1.37%	1.26%	1.53%
sRNN	3.26%	3.55%	3.37%	3.30%	3.79%
sl-cold	5.63%	4.37%	3.37%	2.53%	1.31%
rl-cold	8.74%	7.09%	5.41%	3.92%	2.28%
sl+rl-cold	11.30%	9.31%	6.78%	4.96%	2.50%
sl-warm	3.97%	3.51%	2.98%	2.08%	1.11%
rl-warm	9.02%	7.47%	5.44%	4.19%	2.71%
sl+rl-warm	11.62%	9.91%	7.02%	5.09%	3.61%

F. Convergence Analysis

To demonstrate the effectiveness of the settings in the REINFORCE algorithm, Fig. 6 illustrates the performance curves of different optimization algorithms in the warm-start (p=50%) scenario on the Secondary_school_curriculum 100K dataset.

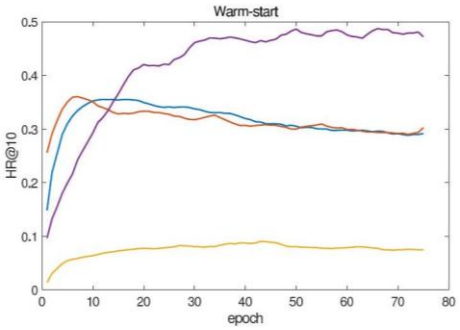


Fig. 6. The Learning curves of different algorithms in the warm-start.

Clearly, the performance of supervised learning methods with complete interaction processes (blue line) or the corresponding separated sub-interaction processes (red line) is similar, indicating that dividing the entire interaction process into several sub-interaction processes does not improve the accuracy of recommendations. Similarly, the performance of the basic REINFORCE algorithm (yellow line) is inferior and even significantly different from methods based on supervised learning. However, the performance of the special REINFORCE method (purple line) shows a significant improvement, as separating and restarting the reward accumulation for overly long interaction processes can obtain more useful self-generated training labels for reinforcement learning. Although this approach is simple, it is highly effective.

G. Recommendation Behavior Analysis

To analyze the recommendation behavior of the recommendation agent, the experiment provides the average results of the relevant popularity and Hit@10 for all test students on the Secondary_school_curriculum 100K dataset at each step t in the warm-start scenario, as shown in Fig. 7.

Thus, the recommendation agent in this study gradually evolves from widespread recommendations to personalized recommendations and can accurately hit courses within the specified range.

H. Dynamic Recommendation Analysis

To evaluate the effectiveness of the dynamic recommendation process, the experiment randomly selected several cases and provided the recommendation results of the proposed model on the dataset. For each case, a sequence of sequentially recommended courses is displayed for a specific student. It is important to note that, for ease of presentation, only one course is displayed in the Top-10 ranking list, i.e. $\hat{a}u, t$. Specifically, for each case from (a) to (c), the results from step 0 to step 14 are always shown. On the other hand, for each case from (d) to (e), the results of a randomly selected consecutive 15 steps within the specified range are displayed.

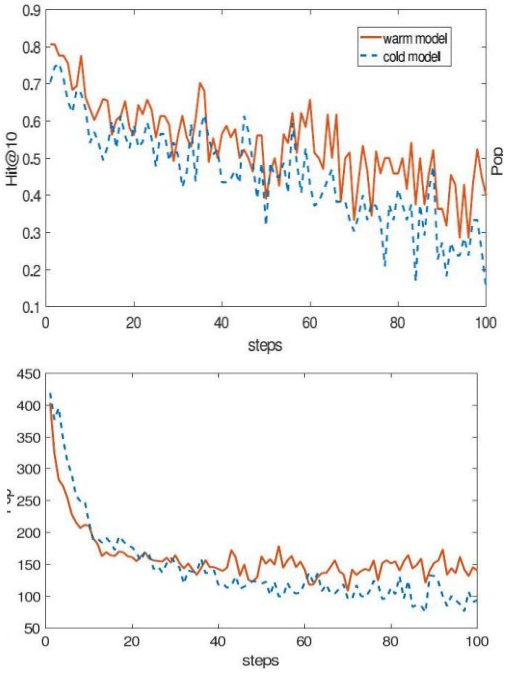


Fig. 7. Characteristics at different time steps in the warm-start scenario. (a) Hit@10; (b) The average results of the relevant popularity.

It can be observed that the proposed method can adaptively adjust recommendations based on past unsuccessful experiences. For example, in case (a), the first two recommendations are incorrect, but the subsequent four recommendations are correct. These results indicate that the proposed method can effectively change recommended courses based on previous feedback from students. Generally, the proposed model can dynamically update student states and modify recommendation results based on corresponding feedback.

In contrast, static methods cannot automatically adjust recommendation results and only make positive predictions at the beginning by correctly predicting recommendations, but they always make negative predictions. Therefore, compared to static methods, the recommendation approach in the proposed model is more effective.

VIII. Conclusion

In this paper, a new Top-N deep reinforcement learning recommender system is proposed to address the problem of long-term recommendations. In the proposed model, the recommendation process is considered as a Markov decision process. Thus, an RNN is used to simulate the sequential interaction between the agent (recommender system) and the environment (students). Moreover, the proposed model can be applied to warm-start scenarios. Additionally, the proposed model does not depend on any content information but only relies on the interaction between the environment and the agent, meaning it can effectively be applied in environments without sufficient content information. Experimental results show that, compared to traditional Top-N recommendation methods, the proposed method has better recommendation performance.

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g.” Avoid the stilted expression, “One of us (R. B. G.) thanks . . .” Instead, try “R. B. G. thanks.”

REFERENCES

[1] Shen X, Liu S, Zhang C, et al. Intelligent material distribution and optimization in the assembly process of large offshore crane lifting equipment[J]. Computers & Industrial Engineering, 2021, 159: 107496.

[2] Xu M, Liu S, Shen H, et al. Process-oriented unstable state monitoring and strategy recommendation for burr suppression of weak rigid drilling system driven by digital twin[J]. The International Journal of Advanced Manufacturing Technology, 2022: 1-17.

[3] Fu T, Liu S, Li P. Intelligent smelting process, management system: Efficient and intelligent management strategy by incorporating large language model[J]. Frontiers of Engineering Management, 2024, 11(3): 396-412.

[4] Zheng H, Liu S, Zhang H, et al. Visual-triggered contextual guidance for lithium battery disassembly: A multi-modal event knowledge graph approach[J]. Journal of Engineering Design, 2024: 1-26.

[5] Nwana, H. S. (1990). Intelligent tutoring systems: An overview. Artificial Intelligence Review, 4(4), 251-277.

[6] Yazdani, M. (1986). Intelligent tutoring systems: An overview. Expert Systems, 3(3), 154-163.

[7] Alhabbash, M. I., Mahdi, A. O., & Naser, S. S. A. (2016). An intelligent tutoring system for teaching English grammar tenses.

[8] Sarrafzadeh, A., Alexander, S., Dadgostar, F., et al. (2008). “How do you know that I don’t understand?” A look at the future of intelligent tutoring systems. Computers in Human Behavior, 24(4), 1342-1363.

[9] Hamed, M. A., & Naser, S. S. A. (2017). An intelligent tutoring system for teaching the seven characteristics of living things.

[10] Al-Bastami, B. G., & Naser, S. S. A. (2017). Design and development of an intelligent tutoring system for C#.

[11] Garnier, P., Viquerat, J., Rabault, J., et al. (2021). A review on deep reinforcement learning for fluid mechanics. Computers & Fluids, 225, 104973.

[12] Wang, H., Liu, N., Zhang, Y., et al. (2020). Deep reinforcement learning: A survey. Frontiers of Information Technology & Electronic Engineering, 21(12), 1726-1744.

[13] Li, Y. (2017). Deep reinforcement learning: An overview. arXiv preprint arXiv:1701.07274.

[14] Sewak, M. (2019). Deep reinforcement learning. Singapore: Springer Singapore.

[15] Hernandez-Leal, P., Kartal, B., & Taylor, M. E. (2019). A survey and critique of multiagent deep reinforcement learning. Autonomous Agents and Multi-Agent Systems, 33(6), 750-797.

[16] Ausin, M. S. (2019). Leveraging deep reinforcement learning for pedagogical policy induction in an intelligent tutoring system. In Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019).

[17] Abdelshiheed, M., Hostetter, J. W., Barnes, T., et al. (2023). Leveraging deep reinforcement learning for metacognitive interventions across intelligent tutoring systems. In International Conference on Artificial Intelligence in Education (pp. 291-303). Cham: Springer Nature Switzerland.

[18] Koroveshi, J., & Ktona, A. (2021). Training an intelligent tutoring system using reinforcement learning. International Journal of Computer Science and Information Security (IJCSIS), 19(3).

[19] Jung, G. (2023). Exploring batch deep reinforcement learning and multi-task learning across intelligent tutoring systems: Lessons learned.

[20] Paduraru, C., Paduraru, M., & Iordache, S. (2022). Using deep reinforcement learning to build intelligent tutoring systems.

[21] Milani, S., Fan, Z., Gulati, S., et al. (2020). Intelligent tutoring strategies for students with autism spectrum disorder: A reinforcement learning approach. In The 2020 CMU Symposium on Artificial Intelligence and Social Good.

[22] Subramanian, J., & Mostow, J. (2021). Deep reinforcement learning to simulate, train, and evaluate instructional sequencing policies. Spotlight presentation at the Reinforcement Learning for Education workshop at the Educational Data Mining 2021 conference.

Advanced Football Match Winning Probability Prediction: A CNN-BiLSTM_Att Model with Player Compatibility and Dynamic Lineup Analysis

Tao Quan*, Yingling Luo
College of Physical Education and Health Science, Chongqing Normal University, Chongqing, 401331, China

***Abstract*—In recent years, with the continuous expansion of the football market, the prediction of football match-winning probabilities has become increasingly important, attracting numerous professionals and institutions to engage in the field of football big data analysis. Pre-match data analysis is crucial for predicting match outcomes and formulating tactical strategies, and all top-level football events rely on professional data analysis teams to help teams gain an advantage. To improve the accuracy of football match winning probability predictions, this study has taken a series of measures: using the Word2Vec model to construct feature vectors to parse the compatibility between players; developing a winning probability prediction model based on LSTM to capture the dynamic changes in team lineups; designing an improved BILSTM_Att winning probability prediction model, which distinguishes the different impacts of players on match outcomes through an attention mechanism; and proposing a CNN-BILSTM_Att winning probability prediction model that combines the local feature extraction capability of CNN with the time series analysis of BILSTM. These research efforts provide more refined data support for football coaching teams and analysts. For the general audience, these in-depth analyses can help them understand the tactical layouts and match developments on the field more deeply, thereby enhancing their viewing experience and understanding of the matches.**

Keywords*—Football big data; match prediction; feature vector; tactical understanding; match analysis

I. INTRODUCTION

Over the past few decades, football has emerged as one of the most popular sports globally, generating significant interest among fans while also attracting the attention of sports scientists, data analysts, and technical researchers [1] [2]. With advancements in technology and the rise of big data, data analysis of football matches has become an expanding research area within competitive sports [3] [4]. Accurate predictions of match winning probabilities provide a scientific basis for team tactics and enhance fans' understanding of the games. Consequently, developing a high-accuracy football match winning probability prediction model holds substantial value for both tactical analysis and audience engagement.

In this context, research on football match analysis technology has progressed. However, the inherent complexity of football matches, characterized by uncertainty and randomness, presents considerable challenges in predicting match outcomes. This study aims to explore and propose innovative data analysis methods to enhance the accuracy of football match-winning probability predictions and facilitate a deeper understanding of tactical layouts and match developments.

Initially, the focus is placed on analyzing player compatibility, as relationships among individual players influence overall team performance. Natural language processing technologies, such as Word2Vec, are employed to generate player feature vectors, which capture potential interactions at the data level and provide foundational features for subsequent winning probability predictions.

Secondly, to account for dynamic factors in football matches, a model based on Long Short-Term Memory (LSTM) networks is introduced to handle time-series data, capturing changes in team lineups and player performance over time. The memory capabilities of LSTM enable the model to learn critical moments during matches and assess the impact of tactical adjustments on outcomes.

Thirdly, to more precisely evaluate the influence of different players on match outcomes, an attention mechanism is integrated into the Bidirectional LSTM (BILSTM) model, resulting in the BILSTM_Att model. The attention mechanism supplies additional information regarding player importance, thereby enhancing prediction accuracy and personalization.

Finally, the study combines the feature extraction capabilities of Convolutional Neural Networks (CNN) with the time-series analysis strengths of BILSTM to propose the CNN-BILSTM_Att model. The CNN layer extracts spatial features of team lineups, while the BILSTM layer processes the evolution of these features over time, with the attention mechanism facilitating the combination of features for more accurate winning probability predictions.

The remainder of the paper is organized as follows: Section I introduces the background and significance of football match winning probability prediction; Section II reviews the literature on prediction methods in competitive sports, developments in natural language processing, and the current state of football match outcome prediction research; Section III discusses data preprocessing, word vector generation, and the development of the CNN-BILSTM_Att winning probability prediction model; Section IV presents a case study to validate the proposed method using a custom experimental dataset; and Section V concludes the paper by summarizing the findings and discussing potential areas for future research.

*Corresponding author

II. Literature Review

This paper will collect works related to existing winning probability predictions to highlight the shortcomings of current research.

A. Analysis Methods of Winning Probability Prediction in Competitive Sports

In the field of winning probability prediction research for competitive sports, the progress of deep learning has become a driving force for transformation [5] [6]. The target detection algorithms based on deep convolutional neural networks, as studied by Mukherjee et al. [7], have now become powerful tools for analyzing players and game dynamics. These algorithms learn complex features from massive amounts of data, helping us understand game development and have demonstrated high accuracy on multiple standard datasets.

To meet the needs of real-time analysis in fast-paced sports, researchers like McGarry et al. [8] have been exploring methods to accelerate target detection algorithms, successfully reducing response times by optimizing network architecture and computational efficiency. This is particularly crucial for winning probability prediction, as instant game information can have a significant impact on the outcome.

However, a major challenge in winning probability prediction is the detection of small targets, such as tracking the position of a soccer ball in a football match or monitoring the movements of distant players in a basketball game. Hodge et al. [9] proposed a series of solutions to this challenge, including multi-scale detection methods and the integration of contextual information, which have been proven to improve the recognition rate of small targets.

The use of ensemble learning has provided another breakthrough for winning probability prediction. The work of Chakraborty et al. [10] shows that by combining the results of multiple prediction models, the accuracy and robustness of predictions can be effectively improved. In sports analysis, ensemble methods can synthesize different data sources and algorithms, such as player statistics, team dynamics, historical performance, etc., to form a more comprehensive prediction of winning probabilities.

These technological advances have provided sports analysts, coaches, and betting companies with unprecedented data insights, helping them make more informed and strategic decisions [11]. They also provide sports fans with a richer viewing experience, such as increasing the interactivity and participation of matches through real-time data analysis [12]. As research continues to deepen, we can expect future winning probability prediction methods to become more refined and personalized, possibly using machine learning models to predict individual player performances or simulate match outcomes more accurately. These research outcomes indicate that target detection algorithms are developing in the direction of being more accurate, faster, and more robust.

B. The Current State of Natural Language Processing Development

Natural Language Processing (NLP) has become a core branch in the field of artificial intelligence, and its current development status highlights significant progress in understanding and generating human language. Particularly, transformative models such as BERT, proposed by Devlin et al., have revolutionized the ability to deeply comprehend semantic meaning in text by introducing bidirectional training mechanisms. These models have also had a revolutionary impact on a variety of NLP tasks including text classification, question-answering systems, and machine translation [13]. In the specific application scenario of sports competition prediction, the advancement of NLP technology is equally noteworthy.

By utilizing advanced text analysis technologies, researchers have been able to extract key information from live sports commentary, aiding in predicting the pace and outcome of games [14]. Furthermore, the application of sentiment analysis technology has made it possible to identify emotions from fans' social media activity, thereby analyzing the correlation between these emotions and game outcomes. Going further, NLP-based methods are being used to understand the patterns of language use among athletes, with the aim of predicting their performance and mental state, and even potential injury risks [15].

Modern NLP technology is also playing a role in automating the parsing of athletes' injury reports, reducing the workload of analysts in data collection and processing [16]. Additionally, semantic analysis of historical game reports can reveal long-term performance trends of teams and players, providing data support for establishing more accurate prediction models [17].

With the continual evolution of pre-trained models, such as the GPT series' capabilities in generating text, NLP has been used to automatically generate sports news articles, pre-game analyses, and even to simulate interviews with coaches and players, bringing new possibilities to media creation and fan interaction [18]. At the same time, these technologies are also helping to create more intelligent virtual assistants that can provide real-time game updates, statistical data interpretation, and personalized sports consumption advice.

In summary, the application of NLP in the field of sports competition prediction indicates that this technology has not only made significant progress in understanding complex text and language patterns but is also bringing profound changes to sports analysis, fan experience, and media reporting. With further research and the refinement of technology, NLP is expected to have an even greater impact in the field of sports in the future.

C. Current Status of Research on Football Match Outcome Prediction

Research on football match outcome prediction involves predicting results by analyzing team and player performance data, match conditions, and other relevant factors. In recent years, this research field has benefited from the rapid development of big data analytics, machine learning, and artificial intelligence technologies, transitioning from traditional statistical methods to more complex and refined predictive models.

With the aid of deep learning technologies, football match outcome prediction models can more effectively process and analyze complex datasets, learn features extracted from historical data, and predict future match outcomes [19]. Some research has adopted structures similar to deep convolutional neural networks, which identify and utilize potential factors affecting match outcomes by learning from a large volume of historical match data [20]. To meet the requirements of real-time prediction, researchers are also striving to improve the computational efficiency of models by optimizing network structures and designing more efficient algorithms, thus reducing the time required for prediction [21]. This enables models to provide predictions rapidly without sacrificing accuracy, satisfying the real-time needs of scenarios such as live match analysis and online betting [22]. In dealing with uncertainties and complex scenarios, such as player absences and weather changes, current research is attempting to incorporate more contextual information and a deeper understanding of the match environment to enhance the robustness of models in the face of varied real game situations [23]. At the same time, football match outcome prediction is focusing on the integrated analysis of multidimensional data, such as players' spatial positioning and team tactical changes. By integrating these multidimensional pieces of information, prediction models can analyze match situations more comprehensively, thus providing more precise predictions [24].

Overall, research on football match outcome prediction has made significant progress in prediction accuracy, processing speed, and adaptation to complex match environments through the adoption of advanced data analysis techniques and algorithmic models. As research continues to deepen and technology continues to develop, this field is expected to achieve higher levels of predictive performance, providing stronger support for football match analysis and related areas.

D. Research Gaps and Future Research Directions

In this research field, we can identify some potential research gaps that may provide directions for future work:

1) Quantification of player psychological states: Current models primarily analyze player performance and compatibility through statistics and machine learning techniques, but players' psychological states also have a significant impact on match outcomes. Integrating players' mentality, stress responses, and other psychological factors into models to predict match outcomes more comprehensively is a research gap that needs to be filled.

2) Comprehensive analysis of environmental factors in the stadium: Although research has considered player compatibility and dynamic performance, factors related to the stadium and environment (such as climate, ground conditions, fan support) are equally important to player performance and match outcomes. Future research could consider incorporating these environmental factors into models to enhance the accuracy and practicality of predictions.

3) In-depth integration of opponent analysis: The tactics of the opposing team, the state of their players, and team compatibility also affect match outcomes, but current models may not fully consider this aspect. Developing models that can analyze the characteristics of both teams and predict the outcomes of their interactions will be a valuable research direction.

The above issues are the key to improving prediction accuracy, and they are also the content of this article's research

III. Data Preprocessing and Word Vectors

A. Collection of Match Data

To ensure the rationality and usability of data, this paper has crawled the historical match records of several strong teams through Whoscored, and obtained the player line-ups used in the matches from these records. The flow diagram of the crawling process is shown in Fig. 1.

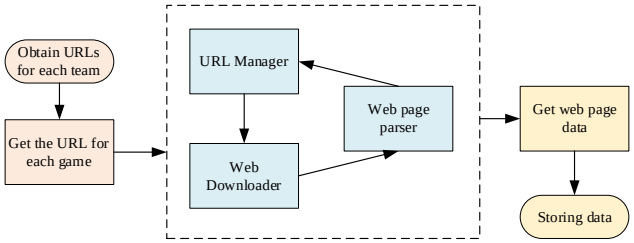


Fig. 1. Match data crawling process.

In the upcoming football events, every coach and fan tries to predict the outcomes of matches by analyzing the line-ups of the teams. Although the performance of the players and the real-time progress of the match are unpredictable and non-generalizable factors, by constructing a reliable pre-match line-up analysis model, we can provide valuable strategic support for the coaching team and also enhance the match experience for fans.

For this analysis, we have already obtained a large amount of line-up data from the database, covering detailed line-up information for 42,596 football matches. To make these data useful, they must first undergo thorough preprocessing. We will filter out information that has potential influence on match results and integrate it to form an efficient match line-up database.

Considering practicality and universality, our model will focus on the following key points:

- Team Strength Balance: Assessing the overall strength of both teams, including the technical statistical data and historical performance of the players.
- Tactical Adaptability: Analyzing the tactical flexibility of each team and how to make corresponding tactical adjustments based on the opponent's line-up.
- Player Condition: Evaluating the current state of key players, including injury conditions and athletic form.
- Historical Match Records: The historical head-to-head records of the two teams are also a factor that cannot be ignored, as it often reveals psychological advantages and disadvantages.

Through the comprehensive analysis of these dimensions, our model aims to provide deeper insights for the upcoming matches, thereby helping coaching teams develop more accurate tactical layouts and bringing richer viewing perspectives to fans.

As the date of the match approaches, we will continue to monitor changes in relevant variables and update our analysis model in a timely manner, ensuring that all predictions are based on the most recent and comprehensive data. We look forward to verifying the accuracy of our model and hope that it will become a powerful tool for predicting football match outcomes.

B. Data Processing

The football match data obtained through web crawling cannot be used directly for lineup analysis; it requires preprocessing of the raw dataset. The main steps of data preprocessing in this article include: data cleaning, data reduction, feature construction, and data annotation.

Since every football player has different positions and technical characteristics, these attributes create special effects that establish interconnections among players. In the match lineup dataset, players are not selected randomly; based on the analysis of the relationships between players, the coaching team selects a set of lineups that can cooperate with each other. In professional matches, the data analysis team will choose a lineup that has strong synergy and can effectively counter the opponent's players.

The interrelationships among players can be summarized as compatibility relationships and counter relationships, both of which jointly influence the outcome of a match. As shown in Fig. 2, the player word vectors in this paper consist of two parts: the first part is the word vectors with compatibility relationships generated by the Word2Vec model, and the second part is the extended vectors generated by the counter relationship algorithm. The composition of the player word vectors is illustrated in the Fig. 2.

Fig. 2. Composition of word vectors.

1) Acquisition of training corpus for word vectors: This article has crawled the home team lineup selection data from 54,652 professional football matches through Wan Plus eSports, using this data to train the Word2Vec model, and these data exclude the previously obtained 42,596 matches. In professional football matches, the lineups chosen after analysis by the coaching team have good compatibility, and the relationship among players within the lineup is strong. Therefore, choosing the lineup of one side in a professional match as the corpus, the Word2Vec model can generate player word vectors with relatively clear relevance. Some data from the corpus are shown in the table.

In this table, NO. represents the sequence number of the match, and P1, P2, P3, P4, P5 represent five different positions, namely [goalkeeper, defender, midfielder, support, forward]. The feature values of these five characteristics indicate the players selected for the corresponding positions. The CBOW model in Word2Vec can predict a word based on the context of that word in the text. During the training process, word vectors are generated as a byproduct, converting text content into a form of a numerical matrix. While generating word vectors, the CBOW model can also calculate the relevance of these vectors. Therefore, based on this feature of the CBOW model, the names representing players are treated as input text information, and the word vectors generated by the CBOW model become the player word vectors. The five different players chosen by the home team in the corpus can thus be seen as a sentence composed of five different words, and the degree of relevance between these words is the degree of compatibility of the players. The structure of the model used to train the word vectors is shown in the Fig. 3.

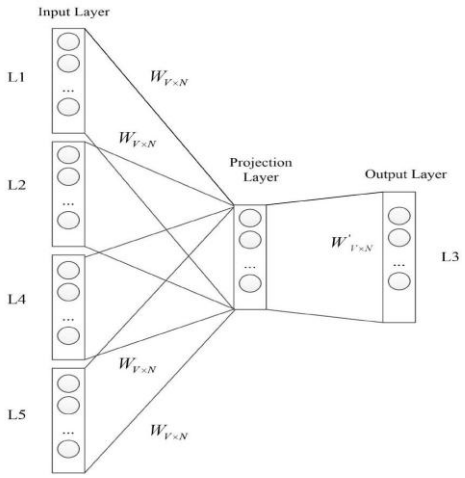


Fig. 3. The model structure for training word vectors.

In this model, the word vector of the target vocabulary is calculated based on the word vectors of the context before and after the target word. In this article, taking the name of the midfielder in the corpus as the target vocabulary (the midfield position usually corresponds to the center of the football formation, similar to the mid-lane position L3 in League of Legends), the preceding vocabulary of the midfielder is the defender (the defender position corresponds to the feature values of features L1 and L2), and the following context is the forward and support (the forward and support positions correspond to the feature values of features L4 and L5). Assuming that the dimension of the word vector is N and the length of the vocabulary is V, the target function of the training model is shown as follows.

$$L = \sum_{L \in V} \log(P(L3 \mid L1, L2, L4, L5))$$

Input layer: First, each player name is encoded using one-hot encoding, for example, player A's one-hot encoding is [1,0,0,0,…]. For convenience of calculation, x_{i-2}、x_{i-1}、x_i、x_{i+1}、x_{i+2} are used to represent the one-hot encodings of the names of feature players P1, P2, P3, P4, P5, which correspond to the positions [goalkeeper, defender, midfielder, support, forward]. Then all one-hot encodings are multiplied by the input weight matrix, as shown below.

$$I_1 = W_{V\times N} \times x_{i-1}$$
$$I_2 = W_{V\times N} \times x_{i-2}$$
$$I_4 = W_{V\times N} \times x_{i+1}$$
$$I_5 = W_{V\times N} \times x_{i+2}$$

Hidden layer: Average all word vectors to get the hidden vector, as shown in equation

$$\hat{I} = \frac{I_1 + I_2 + I_4 + I_5}{4}$$

Output layer: Multiply the hidden vector of the hidden layer by the output weight, as shown in equation.

$$z = W'_{V\times N}\hat{I}$$

Then use a softmax classifier to obtain the probability that each hero in the hero library is predicted as the current result, as shown in equation.

$$\hat{y} = \text{softmax}(z)$$

In this model training process, the weight from the input layer to the hidden layer is the word vector. The model can calculate the similarity probability between players, and the higher the similarity probability between two players, the stronger their relevance, indicating a strong pairing relationship between them. For example, if a midfielder (such as Messi) often appears with a forward player (such as Suarez), their word vector distance will be very small after generating the word vectors.

2) Acquisition of word vectors: This article uses the Gensim toolkit to import the Word2Vec model. Gensim is an open-source Python toolkit that supports multiple topic modeling algorithms such as TF-IDF, LDA, and Word2Vec. During the process of training word vectors with the Word2Vec model, the model's parameters are set as follows:

The sg parameter is the option to select the training mode. When sg=0, the model uses the CBOW (Continuous Bag of Words) model to train word vectors; when sg=1, the model uses the Skip-Gram model to train word vectors. This article uses the CBOW model to train player word vectors;

The size parameter indicates the dimension of the word vectors produced. In small datasets, this is usually set between 100-200. After multiple experiments, this article selects size=100;

The window parameter indicates the maximum possible distance between the current word and the predicted word. The larger the window parameter is set, the more predicted words need to be enumerated. Since there are five players input at a time in the training corpus, this parameter is set to 5;

The min_count parameter indicates that if a word appears in the entire dataset less than the set value, the word will be directly ignored. Since there is enough competition data in this article and each player has many match records, this parameter is set to 0.

The following are some player vectors after training through the Word2Vec model, for ease of representation, the player names in this article are represented as Player(name):

Player("Messi") = [0.25237, -0.10362, 0.75654, …]

Player("Cristiano Ronaldo") = [0.46895, 0.55863, -0.05965, …]

Player("Neymar") = [-0.75294, -0.63594, 0Sure!]

Compatibility can be calculated through cosine similarity, which reflects the degree of approximation of players in the semantic space. To some extent, this can be mapped to their ability to cooperate in actual games. Based on actual game experience, in football matches, among the players with the highest match compatibility with a key player (such as Messi), most are midfielders or wing-backs who can provide support and assists. These players are able to pass the ball well to Messi and create opportunities for him to score. When Messi receives effective support and has enough space to shoot, his team is more likely to win the game. Therefore, in professional games, Messi appears more frequently with these supporting players, indicating a strong compatibility between them.

Through the analysis of Messi's word vector compatibility, it can be seen that the Word2Vec model used in this paper produces word vectors that well reflect the compatibility between players in the team lineup. Players who often appear with Messi in games have a high degree of compatibility; those who have difficulty appearing with Messi have a lower degree of compatibility. This kind of analysis and discovery provides deep insights into the interaction between players in football matches and team cooperation strategies.

For example, if Messi and a specific midfielder often appear together in actual games and usually can cooperate to lead the team to victory, we can expect that in the word vector space produced by Word2Vec, Messi and this midfielder will have a higher degree of similarity or proximity. Such information is very valuable for the formulation of football strategies, especially when selecting player lineups and devising game tactics.

3) Expansion of word vectors: In football match prediction, not only is there a chemical reaction between teams, but also a suppression relationship. To better reflect the characteristics of the team, this paper uses the suppression relationship to generate a suppression vector to expand the team vector based on statistical data. Thus, the final team vector includes both the chemical reaction between teams and their suppression relationships, making the information contained in the team vector more comprehensive.

The suppression relationship between teams is generally determined based on the experience of fans and professional analysts in actual games. For example, most analyses might find that Team_A is limited by Team_B's tactics when evaluating the match between the two. This kind of suppression relationship is difficult to represent with specific numbers or vectors. This paper studies the win rate data between teams and finds that the suppression relationship can be reflected in the team's win rate.

On football data platforms, such as Opta, WhoScored, or Transfermarkt, one can obtain the overall win rate of each team. It can be seen that in a certain season, Arsenal's overall win rate is 60%, while Chelsea's overall win rate is 55%. The total win rate of these two teams is not much different, and usually, the probability of victory for either team in their matches is close to 50%. However, in actual games, Arsenal may find it difficult to compete with Chelsea, and by analyzing the data, it can be found that Arsenal's actual win rate against Chelsea may only be 40%. Therefore, by comparing the expected win rate and the actual head-to-head win rate of the teams, one can determine whether there is a suppression relationship.

Through the analysis of different teams' win rates and head-to-head win rates, the suppression relationship can be quantified through the following calculation process as shown in Fig. 4, the algorithm flow is as follows:

- Collect historical head-to-head data for Team_A and Team_B, including win rates and head-to-head win rates.
- Calculate the average win rate of the two teams as a reference value for the expected win rate.
- Use the difference between the actual head-to-head win rate and the expected win rate to assess the strength of the suppression relationship.
- Convert this difference into a suppression vector and combine it with the team vector.
- Use the expanded team vector for more accurate match predictions.

By this method, we can more comprehensively understand the competitive relationship between teams and consider more variables when predicting future game outcomes.

First, calculate the global win rate p_i for the teams in the dataset using the following Formula:

$$p_i = \frac{N_i}{M_i}$$

In this formula, i represents any team, N_i represents the number of matches won by the team, and M_i represents the total number of matches the team has participated in. Similarly, the global win rate p_j for the team j that team i is facing can be calculated; By using the global win rates of both teams, the expected win probability when the teams face each other can be calculated using the following formula:

$$E_{ij} = \frac{p_i - p_i \times p_j}{p_i + p_j - 2 \times p_i \times p_j}$$

In this formula, E_{ij}represents the expected win rate when team i faces team j;

Calculate the actual win rate $p'_{i,j}$ of the teams' encounters using the existing match dataset.

$$p'_{i,j} = \frac{n_{ij}}{m_{ij}}$$

In this formula, n_{ij} represents the number of matches won by team i against team j, and m_{ij} represents the total number of matches played between team i and team j;

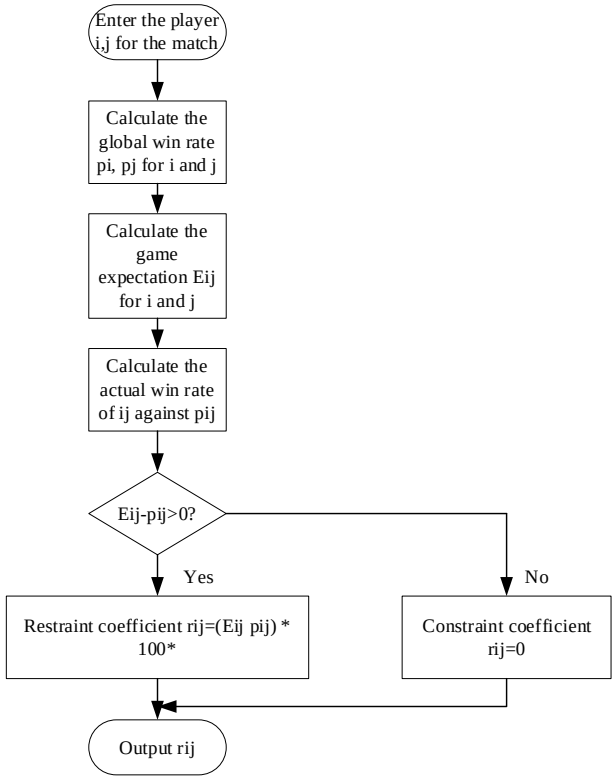


Fig. 4. Suppression relationship algorithm flow.

The specific algorithm process is as follows:

By comparing the expected win rate of the encounter between team i and team j with the actual win rate, if the actual win rate is less than the expected win rate, that is, $E_{ij} - p'_{i,j} > 0$,, then it can be considered that team j has a suppressing effect on team i;

After determining the suppression relationship between teams, further quantify the suppression relationship to be input into the prediction model. The win rate difference can be used as the basis for the suppression relationship coefficient. The calculation formula for the suppression relationship coefficient is as follows:

$$r_{ij} = \left(E_{ij} - p'_{i,j}\right) \times 100\%$$

In this formula, r_{ij} represents the coefficient of the suppression relationship between teams. The difference between the expected win rate and the actual win rate represents the suppression relationship index. The larger the value of r_{ij}, the stronger the suppressive effect of team j on team i.

Calculate the suppression relationship coefficient for each team and its opposing teams one by one. If there is no suppression relationship, the suppression relationship coefficient is set to 0. In this way, the suppression relationship vector for each team can be obtained.

In order to facilitate the generation of suppression relationship vectors between teams, an algorithm for generating suppression relationship vectors has been constructed. The logical structure of this algorithm is visualized as follows:

- Collect and process match data.
- Calculate the global win rate of the teams.
- Calculate the expected win rate of the encounters.
- Calculate the actual win rate of the encounters.
- Assess and calculate the suppression relationship coefficient.
- Generate the suppression relationship vector for the teams.

Through these steps, a structured dataset can be obtained, which is used to improve the accuracy of football match outcome predictions.

To facilitate the generation of expanded vectors for hero suppression relationships, this paper has constructed an algorithm for generating expanded vectors. Fig. 5 shows the logical structure of this algorithm.

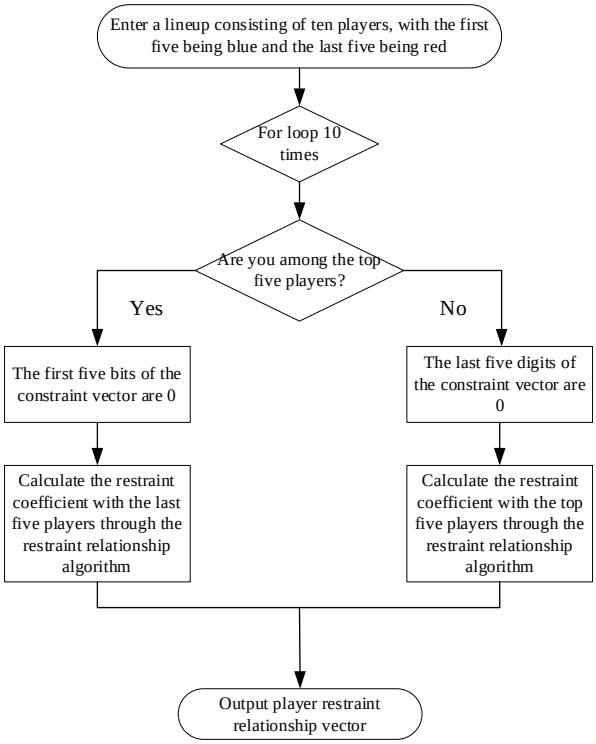


Fig. 5. The logical structure of the expanded vector generation algorithm.

C. CNN-Enhanced CNN-BILSTM_Att Win Probability Prediction Model

In the field of football match prediction, due to the multitude of factors involved in a game and the complex relationships between them, reliance solely on traditional statistical data and historical records for predictions is limited. For example, the dimension of "head-to-head history" is often affected by various factors, and its degree of influence on the outcome of the match is not easily quantifiable. Current football match prediction methods vary, including those based on odds, subjective predictions based on historical experience, etc., but there is a lack of a unified data mining-based quantitative analysis process to reveal the intrinsic connections between features, which poses obstacles to subsequent match analysis and prediction.

To address this issue, the BILSTM_Att prediction model adopts a Bi-directional Long Short-Term Memory network combined with an attention mechanism to effectively extract the characteristic information of team line-ups. However, the BILSTM_Att model may not fully capture local information when processing input features, and the model fitting process can be slow due to the large number of weight parameters in the hidden layers. To overcome these shortcomings, we propose an improved prediction model: CNN-BILSTM_Att.

This model incorporates the advantages of Convolutional Neural Networks (CNN), which can quickly process data and reduce the dimensionality of features through their pooling layers, thereby effectively reducing the number of weight parameters in the prediction model. This CNN-optimized model not only retains the sensitivity of the BILSTM_Att model to time-series data but also introduces the ability of CNN to extract local features in image and sequence processing, enhancing the model's capability to capture local contextual information.

The improved CNN-BILSTM_Att prediction model can analyze the counter-relations between teams and the dynamics of matches more accurately, thereby improving the accuracy of match outcome predictions. The final model structure and core process can be visually represented, as shown in Fig. 5-8. This will help researchers and analysts better understand the workings of the model and apply it to subsequent match prediction work.

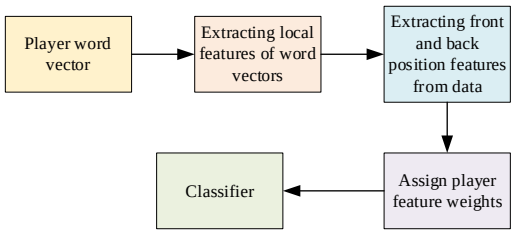


Fig. 6. Core process of the prediction model optimized with CNN.

The architecture of the improved model achieves an effective combination of feature extraction and time series analysis, with the following detailed structure:

1) Convolutional Neural Network (CNN) layer: The main task of this layer is to process the input lineup data and extract local features from it. In the context of football data analysis, this is akin to the ability of CNN to extract local pixel features in image processing tasks. In the scenario of football match prediction, the CNN can extract key local patterns and features from player data, positional information, and other relevant statistics, such as the relationships between players and the team's formation structure.

2) Bi-directional Long Short-Term Memory (Bi-directional LSTM) with Attention Mechanism (Att) layer: Data processed by the CNN layer is passed on to the BILSTM_Att layer. The BILSTM can capture the forward and backward dependencies in time-series data, while the attention mechanism helps the model focus on those pieces of information that are most critical to the final prediction. In football match prediction, this means the model can analyze not only the individual characteristics of players but also understand the dynamics of interactions between players and their changing impacts over time.

3) Softmax classifier: Finally, the feature vector output from the BILSTM_Att layer is passed to the softmax classifier. This classifier predicts the possible outcomes of the match, such as home win, draw, or away win, based on the extracted features.

This improved model—referred to as the CNN-BILSTM_Att model—leverages CNN's ability to extract highly relevant local features, and BILSTM's strength in processing time-series data. In this way, the model not only understands the current state of each team's lineup but also considers the potential impact of the evolution of team lineups over time on the match outcome.

The model's architecture can be represented by the following structure diagram:

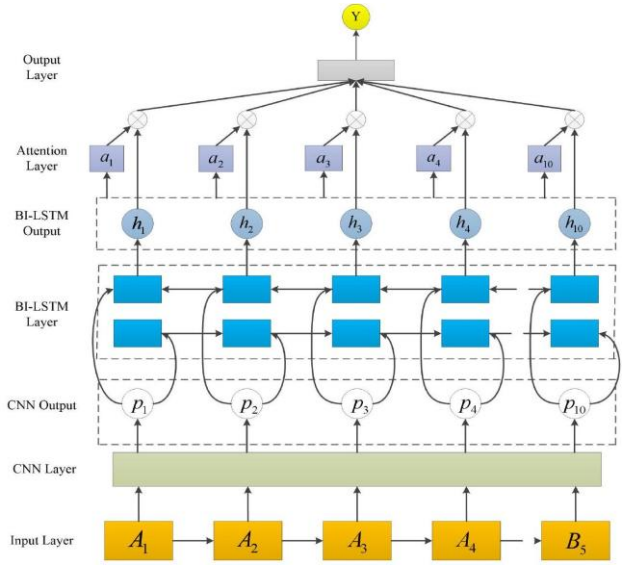


Fig. 7. Structure of the CNN-BILSTM_Att win probability prediction model.

The input and output layers of the CNN-BILSTM_Att model are the same as those of the BILSTM_Att model. In the hidden layers, the CNN-BILSTM_Att model includes an additional CNN layer compared to the BILSTM_Att model. This CNN layer extracts feature vectors from the lineup and then inputs the extracted features into the BILSTM structure. The structure of the CNN model is shown in Fig. 8.

Assuming the team lineup matrix is G, the rows of G represent the vectors of the players in the lineup. Each input lineup consists of 11 players (according to a real football game, each team has 11 field players). The dimensionality of the player vectors is composed of various features such as the technical characteristics, physical data, and tactical role of the players, which is currently set to d. Therefore, G is an 11×d dimensional matrix. $V(W(t))$ represents the d -dimensional feature vector of the t-th player in the lineup matrix G.

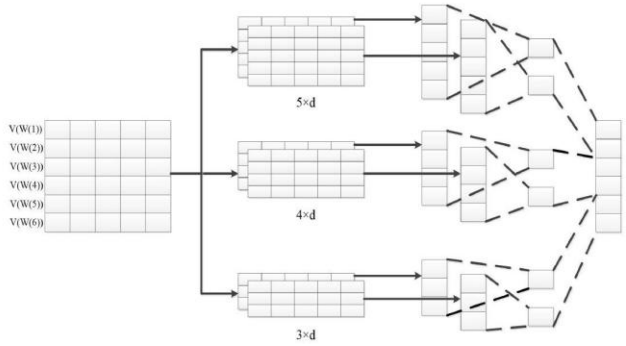


Fig. 8. Convolutional neural network structure.

When the team lineup information is input into the convolutional neural network, the network's convolutional layer uses multiple convolutional kernels of different sizes $l \times d$, to perform convolution on the team lineup matrix G, thus obtaining local feature mappings k_t . The stride of the convolutional kernels is 1. The convolution process is shown as equation:

$$k_t = f(C \cdot G_{t:t+l-1} + b)$$

In equation 5-17, f represents the activation function, and in this model, the Relu function is used as the activation function. C represents the convolutional kernel function, l represents the size of the convolutional window, $G_{t:t+l-1}$ represents the convolutional range of the convolutional kernel, which is from the t-th to the t+l-1-th position of the team lineup matrix G, and b represents the bias term. After convolution, the model can obtain a collection of team lineup feature vectors composed of multiple local features K, with the expression for K shown as equation:

$$K = (k_1, k_2, \dots, k_{m-l+1})$$

In equation, m represents the length of the lineup. In the problem of win probability prediction, each team lineup consists of eleven players, so m=11. After the convolutional layer, a pooling layer can be used to obtain strong features composed of local features. In the CNN-BILSTM_Att model, multiple filters are used to generate the features for each player. The player features after pooling are shown as equation:

$$p = \max(k_1, k_2, \dots, k_{10-l+1})$$

The final representation of the team lineup X^C is derived as equation.

$$X^C = [p_1, p_2, \dots, p_{10}]$$

The model transmits the X^C obtained from the feature extraction of the team lineup data by the convolutional neural network to the bidirectional LSTM. The subsequent data integration and classification steps are the same as those of the BILSTM_Att model.

The improved CNN-BILSTM_Att model first extracts local features, then inputs these features into the bidirectional LSTM network layer and the attention mechanism layer. After integrating features in the attention mechanism layer, it finally uses a softmax classifier for classification. By extracting features from the lineup information through the CNN model, the win probability prediction model can more accurately capture the local information of the lineup, thereby improving the accuracy of predictions.

IV. Case Study

This section will validate the effectiveness of the proposed method using a custom experimental dataset based on a sports stadium scenario.

A. Experimental Environment

The hardware environment for the experiments in this chapter is shown in Table I:

TABLE I. Experimental Software and Hardware Environment Chart

Component	Specification
CPU	Intel i7 8700k
GPU	GTX3080
Memory	32G
OS	Ubuntu18.04
CUDA	11.1
Main Frameworks	Pytorch, keras
Main Programming Language	Python3.6

Due to the complexity of the CNN-BILSTM_Att model, Fig. 9 illustrates the training process of this model.

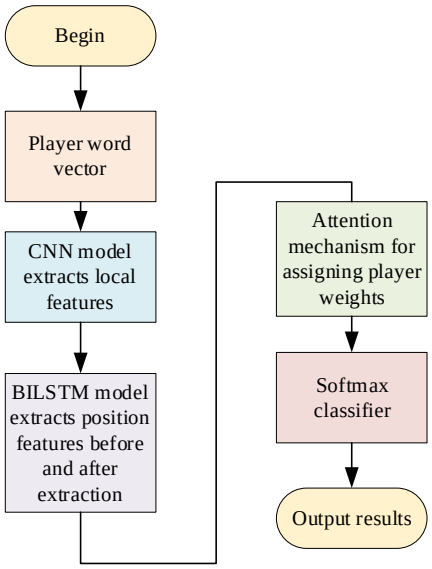


Fig. 9. Training process of CNN-BILSTM_Att.

In the training process of the CNN-BILSTM_Att model, the focus is primarily on two aspects: feature extraction and classification. Initially, the CNN model extracts local features of the word vectors representing players. Subsequently, the BiLSTM model is utilized to extract positional features of the players' lineup. Following that, through the attention mechanism, different weights are assigned to players, reflecting their importance in the match. Finally, a softmax classifier is deployed for classification. The parameters of the CNN-BILSTM_Att win probability prediction model are displayed in Table II.

TABLE II. Hyperparameters of the CNN-BILSTM_Att Prediction Model

Parameter	Parameter Value
Player Word Vector Dimension	110
BILSTM Layer Dimension	16
Learning Rate	0.01
Output Dimension	1
batch_size	200
Convolutional Kernel Sizes	[2,3,4,5]
Parameter	Parameter Value

The model's output employs the player feature vectors generated in Chapter 3, comprising a 100-dimensional feature vector from players' technical statistical data and game performance, along with a 10-dimensional extension vector (potentially including the player's position, health status, and psychological factors). The BiLSTM layer was optimized to select 16 hidden units. The learning rate and batch_size parameters were also finely tuned to achieve the optimal parameters of 0.01 and 200.

Regarding the CNN layers, this study contrasts convolutional kernels of different sizes to assess their impact on the model's performance. Convolutional kernels with sizes 2, 3, 4, and 5 were selected, and the accuracy of the prediction models with these different sizes on the validation set is shown in Fig. 10. Through these experiments, we aim to determine which convolutional kernel size can most effectively enhance the accuracy of football match outcome predictions.

The deep learning techniques applied to process football match data allow the prediction model to extract valuable features from individual technical statistics of players and the overall tactical execution of the team. This aids in more accurately predicting match outcomes. This approach translates the comprehensive performance of players and teams into a format understandable by the prediction model, thus providing technical support for various types of football analysis.

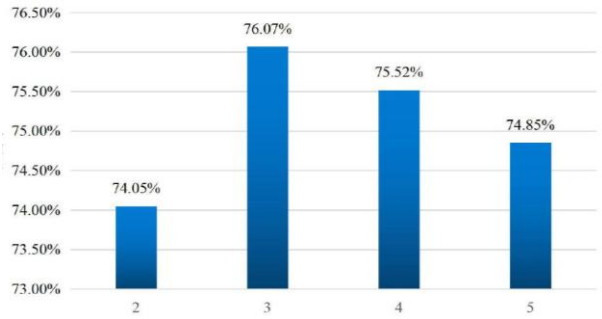


Fig. 10. Accuracy comparison of convolutional kernels of different sizes.

As can be seen from Fig. 10, the accuracy of the four different sizes of convolutional kernels was higher than that of the pure BILSTM_Att model. When the kernel size was 3, the CNN-BILSTM_Att model achieved the highest accuracy in predicting football match outcomes, reaching 76.07%. However, when the kernel size was 2, the model's performance may have been insufficient due to the lack of feature extraction capability. When the kernel size was greater than 3, using larger convolutional kernels meant an increase in the model's weight parameters, which could lead to a decrease in model performance due to issues such as insufficient computational resources.

B. Experimental Results

Fig. 11 shows the accuracy curve of the CNN-BILSTM_Att model for football match win probability prediction over 200 iterations. Overfitting phenomena also occurred during the fitting process of this model. Due to the CNN model's strong feature extraction capability, the CNN-BILSTM_Att model showed a significant improvement in prediction accuracy over the BILSTM_Att model, demonstrating greater expressive power. It can also be seen that the model fitted more quickly, with the CNN-BILSTM_Att model reaching its best performance around the 80th iteration.

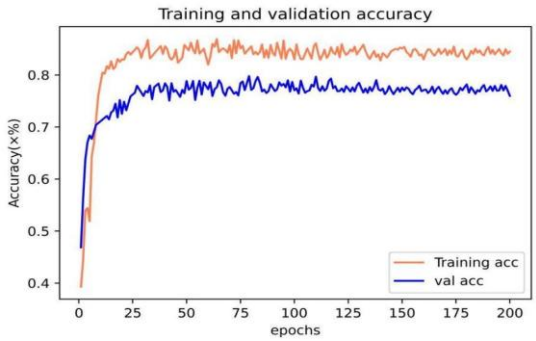


Fig. 11. Fitting curve of the CNN-BILSTM_Att model.

For a more comprehensive comparison, this paper further constructed a CNN-BILSTM_Att model that does not use the extended feature vector as input, namely a CNN-BILSTM_Att model without historical encounter information (No Historical Encounter Relationship CNN-BILSTM_Att, NHR-CNN-BILSTM_Att). This variant differs from the original CNN-BILSTM_Att win probability prediction model only in terms of the input data; it uses only the feature vectors generated from regular team and player statistical data as the model input, in order to assess the impact of historical encounter information on win probability prediction. The model parameters were kept consistent with the complete prediction model. The NHR-CNN-BILSTM_Att model ultimately achieved an accuracy of 74.95%, and the fitting process is shown in Fig. 12.

The results of the NHR-CNN-BILSTM_Att prediction model indicate that the historical encounter information, or the extended vector, has some impact on the final prediction results, suggesting that considering the history of encounters between teams is important when constructing a football match win probability prediction model.

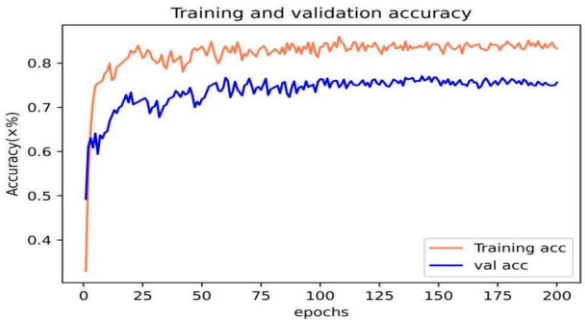


Fig. 12. Fitting curve of the NHR-CNN-BILSTM_Att model.

V. Conclusion

This paper investigates win probability prediction for football matches using deep learning techniques. By analyzing team line-ups and player abilities, the study develops the CNN-BILSTM_Att model, achieving high accuracy in predicting match outcomes. Player statistical data is transformed into numerical vector representations derived from historical performance and ability scores. Initial attempts with three machine learning methods yielded low accuracy, prompting the introduction of an LSTM model to explore key positional information in line-ups. The final models, BILSTM_Att and CNN-BILSTM_Att, incorporate an attention mechanism to weigh player contributions and extract local features effectively. The proposed models enhance win probability predictions by providing valuable insights for coaches, analysts, and fans, thereby enriching the overall viewing experience. Additionally, these methodologies may inspire similar applications in other sports, advancing the field of sports data science.

Acknowledgment

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g.” Avoid the stilted expression, “One of us (R. B. G.) thanks . . .” Instead, try “R. B. G. thanks.”

References

[1] Zheng, H., Liu, S., Zhang, H., et al. (2024). Visual-triggered contextual guidance for lithium battery disassembly: a multi-modal event knowledge graph approach. Journal of Engineering Design, 1-26.

[2] Fu, T., Li, P., & Liu, S. (2024). An imbalanced small sample slab defect recognition method based on image generation. Journal of Manufacturing Processes, 118, 376-388.

[3] Aoki, R. Y. S., Assuncao, R. M., & Vaz de Melo, P. O. S. (2017). Luck is hard to beat: The difficulty of sports prediction. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1367-1376).

[4] Bunker, R., & Susnjak, T. (2022). The application of machine learning techniques for predicting match results in team sport: A review. Journal of Artificial Intelligence Research, 73, 1285-1322.

[5] Li, Y., Wang, L., & Li, F. (2021). A data-driven prediction approach for sports team performance and its application to National Basketball Association. Omega, 98, 102123.

[6] Valero, C. S. (2016). Predicting win-loss outcomes in MLB regular season games: A comparative study using data mining methods. International Journal of Computer Science in Sport, 15(2), 91-112.

[7] Mukherjee, S., Huang, Y., Neidhardt, J., et al. (2019). Prior shared success predicts victory in team competitions. Nature Human Behaviour, 3(1), 74-81.

[8] McGarry, T., & Franks, I. M. (1994). A stochastic approach to predicting competition squash match-play. Journal of Sports Sciences, 12(6), 573-584.

[9] Hodge, V. J., Devlin, S., Sephton, N., et al. (2019). Win prediction in multiplayer esports: Live professional match prediction. IEEE Transactions on Games, 13(4), 368-379.

[10] Chakraborty, S., Dey, L., Maity, S., et al. (2024). Prediction of winning team in soccer game: A supervised machine learning-based approach. In Advances on Mathematical Modeling and Optimization with Its Applications (pp. 170-186). CRC Press.

[11] Buhamra, N., Groll, A., & Brunner, S. (2024). Modeling and prediction of tennis matches at Grand Slam tournaments. Journal of Sports Analytics, 10(1), 17-33.

[12] Wen, Z., Liu, J., & Liu, C. (2024). Football momentum analysis based on logistic regression. Frontiers in Computing and Intelligent Systems, 7(2), 60-64.

[13] Shamshiri, A., Ryu, K. R., & Park, J. Y. (2024). Text mining and natural language processing in construction. Automation in Construction, 158, 105200.

[14] Just, J. (2024). Natural language processing for innovation search: Reviewing an emerging non-human innovation intermediary. Technovation, 129, 102883.

[15] Mekkes, N. J., Groot, M., Hoekstra, E., et al. (2024). Identification of clinical disease trajectories in neurodegenerative disorders with natural language processing. Nature Medicine, 1-11.

[16] Gagliardi, G. (2024). Natural language processing techniques for studying language in pathological ageing: A scoping review. International Journal of Language & Communication Disorders, 59(1), 110-122.

[17] Raza, S., Garg, M., Reji, D. J., et al. (2024). Nbias: A natural language processing framework for BIAS identification in text. Expert Systems with Applications, 237, 121542.

[18] Joshi, A., Dabre, R., Kanojia, D., et al. (2024). Natural language processing for dialects of a language: A survey. arXiv preprint arXiv:2401.05632.

[19] Bunker, R., Yeung, C., & Fujii, K. (2024). Machine learning for soccer match result prediction. arXiv preprint arXiv:2403.07669.

[20] Chakraborty, S., Dey, L., Maity, S., et al. (2024). Prediction of winning team in soccer game: A supervised machine learning-based approach. In Advances on Mathematical Modeling and Optimization with Its Applications (pp. 170-186). CRC Press.

[21] Holmes, B., & McHale, I. G. (2024). Forecasting football match results using a player rating based model. International Journal of Forecasting, 40(1), 302-312.

[22] Xiaoyu, F., & Shasha, W. Evaluating the pinnacle of football match key statistics as in-play information for determining match outcomes in Europe's foremost leagues. Social Science Quarterly.

[23] Saribekyan, G., & Yarovoy, N. (2024). Football prediction model based on teams' Elo ratings and scoring indicators.

[24] Gu, C., De Silva, V., & Caine, M. (2024). A machine learning framework for quantifying in-game space-control efficiency in football. Knowledge-Based Systems, 283, 111123.

Effectiveness of Immersive Contextual English Teaching Based on Fuzzy Evaluation

Mei Niu*

Basic Courses Department, Jiyuan Vocational and Technical College, Jiyuan 459000, Henan, China

***Abstract*—Investigating the real-world impact of immersive contextual instruction on English language education, verifying its contribution to the enhancement of linguistic skills and the improvement of learning attitudes, and evaluating the practicality and worth of fuzzy evaluation in gauging teaching efficacy. A fuzzy complete assessment model was built utilizing the language competency test and the learning attitude questionnaire, and the teaching effect was quantitatively examined based on the experimental data using methods such as affiliation function and weight calculation. The study's findings revealed that students in the experimental group performed much better than students in the control group in terms of language competence and learning attitudes, with an overall fuzzy score of 88.5 compared to 74.8 in the latter. The statistical test indicated a significant difference between the groups (p<0.001). The study also confirmed the scientific and practical validity of fuzzy evaluation in the assessment of multidimensional educational efficacy. Immersion contextual English teaching provides considerable benefits for improving students' language skills and learning attitudes. The fuzzy assessment method introduces a new instrument for quantitative research on teaching efficacy and has a wide range of potential applications.**

***Keywords*—Fuzzy evaluation; immersion; contextual English teaching; teaching effectiveness; teaching assessment**

I. Introduction

With the acceleration of globalization, the improvement of the teaching effect of English, as an important tool for international communication, has increasingly become the focus of research in the field of education [1]. The traditional English teaching mode is mostly based on teachers explaining and students listening to lectures, which lacks contextual immersion, resulting in students' language application ability often being difficult to effectively improve [2]. In recent years, immersion teaching, as an innovative teaching mode, has gradually gained the attention of more and more educators [3]. By simulating the real language use environment, immersion teaching enables students to experience language learning in real situations, thus improving their language proficiency and cultural understanding [4]. However, although the immersion mode of teaching shows obvious advantages in English language teaching, the evaluation of its teaching effectiveness faces a big challenge [5]. Traditional assessment methods often rely on quantitative test scores or qualitative teacher assessment, but these methods are difficult to fully reflect students' comprehensive language proficiency in immersion contexts [6]. Therefore, how to scientifically and accurately evaluate the effects of immersion context teaching has become an important topic in current educational research.

In the complex and dynamic educational context, the fuzzy evaluation method, as a tool adept at handling uncertainty and vagueness, demonstrates its distinctive merits [7]. It can integrate both qualitative and quantitative metrics and conduct a holistic analysis of teaching outcomes from various evaluation perspectives, thereby offering a novel approach for assessing the efficacy of immersive English instruction [8]. Against this backdrop, this study aims to employ the fuzzy evaluation method to carry out a systematic investigation into the teaching effectiveness of immersive contextual English, with the goal of furnishing educators with a more precise and all-encompassing teaching effectiveness assessment instrument [9]. The primary objective of this research is to construct a fuzzy evaluation model suitable for immersive contextual English teaching via empirical research, identify the crucial factors influencing teaching effectiveness, and analyze the data through specific cases [10]. It is anticipated that this research will provide theoretical backing for future English teaching endeavors and an effective basis for decision-making for educational policymakers.

The content structure of this paper is divided into five main sections. Section II of the research review comprehensively combed the domestic and international literature related to this study, including the theoretical foundation of immersive contextual English teaching and the progress of applied research, while analyzing the characteristics of the fuzzy evaluation method and its application value in educational research. Then, Section III of the research method elaborates the design framework of this study in detail, including the selection of experimental samples, data collection methods, the construction process of the fuzzy comprehensive evaluation model, and the specific calculation steps. Subsequently, Section IV of the results and discussion presents the differences in teaching effects between the experimental group and the control group based on empirical data and combines the quantitative results of the fuzzy evaluation with an in-depth discussion of the superiority of immersive contextual teaching and the specific performance of its teaching effects. Finally, Section V of the conclusion summarizes the main findings of the study, analyzes the shortcomings of the study, and looks forward to the future research direction. The parts are interlocked and work together to serve the achievement of the research objectives.

II. Research Review

To fully understand the theoretical background and the current application status of immersive contextual teaching and

*Corresponding Author

fuzzy evaluation, this part will be developed from the following two aspects: the current research status of immersive contextual teaching and its evaluation methods, and the progress of the application of fuzzy evaluation methods in the field of education and its advantages.

A. Research Status and Evaluation Methods of Immersive Contextual Teaching and Learning

The core of the immersive contextual teaching model is to enhance learners' language communication skills by constructing real or simulated language use contexts so that they can practice the language in a near real environment [10]. This model has been widely used in language learning, science experiments cultural courses, etc. Its advantage is that it can enhance students' participation and language acquisition. Some studies have shown that students' language expression and cultural understanding are significantly improved when immersive learning environments are constructed through virtual reality (VR) technology or classroom scenario simulation [11]. However, existing studies have mostly focused on the following three ways of evaluating the effects of immersive teaching: (1) performance assessment based on quantitative data, such as standardized test scores, and comparison of teaching effects through pre-tests and post-tests. (2) Qualitative assessment, such as evaluating teaching effectiveness through students' self-feedback, interviews, or teachers' classroom observations. (3) Mixed assessment, i.e., combining quantitative and qualitative indicators, such as combining test scores and questionnaires for comprehensive analysis. Although these methods can reflect teaching effectiveness to a certain extent, they are often inadequate in dealing with complex and multidimensional evaluation issues [12]. For example, it is difficult to measure the comprehensiveness of students' language useability by relying only on test scores, while subjective interviews and observations are susceptible to the subjective bias of the evaluator [13]. Table I presents for typical immersion teaching evaluation methods and their strengths and weaknesses; therefore, there is a need to introduce evaluation methods that are more scientific, multidimensional, and capable of quantifying complex phenomena.

TABLE I TYPICAL IMMERSION EVALUATION METHODS AND THEIR ADVANTAGES AND DISADVANTAGES

Evaluation Methodology	Vintage	Drawbacks
quantitative assessment	(1) Strong objectivity and easy access to and analysis of data (2) Can be used for comparison and statistical analysis of large samples	(1) Confined to numerical results, it is difficult to fully reflect the effectiveness of teaching and learning (2) Inability to capture learners' subjective experiences and emotional changes
Qualitative assessment	(1) The ability to dig deeper into learners' subjective feelings about teaching and learning (2) Helps capture details and dynamic changes in language use	(1) Highly subjective, with results susceptible to the personal biases of the evaluators (2) A small sample size makes it difficult to generalize to a wider area
Blended assessment	(1) Combines the strengths of quantitative and qualitative methods (2) Teaching effectiveness can be evaluated more comprehensively from multiple dimensions	(1) Complexity of data processing, which may increase the cost of the study (2) The qualitative component remains vulnerable to subjectivity
Assessment based on fuzzy evaluation	(1) Suitable for dealing with complex, multidimensional, and difficult-to-quantify problems (2) Ability to synthesize different indicators and generate overall evaluations (3) Reducing the bias of single data	(1) The model-building process may be subjective in terms of parameter setting (2) The affiliation function and weights should be set scientifically and carefully.

B. Application of Fuzzy Evaluation Methods in Education

The fuzzy evaluation method originated from the fuzzy set theory proposed by Zadeh, which aims to solve the limitations of traditional evaluation methods in dealing with vagueness and uncertainty. In the field of educational evaluation, fuzzy evaluation has attracted much attention because of its ability to integrate multidimensional indicators and handle quantitative and qualitative data [14]. In recent years, fuzzy evaluation methods have been widely used in the fields of teaching quality evaluation, students' comprehensive quality evaluation, and course satisfaction analysis. A study applied the fuzzy comprehensive evaluation method to the assessment of university teaching quality and established a set of multi-dimensional comprehensive evaluation models by setting subjective weights and affiliation functions, which greatly improved the scientificity and persuasiveness of the evaluation results. Similarly, another study proved the high applicability of the fuzzy evaluation model in the assessment of multi-indicator teaching effectiveness in the study of vocational skills teaching. Compared with traditional evaluation methods, fuzzy

evaluation has several advantages [15]. The first is the integration of multidimensional data. It can synthesize and analyze data of multiple dimensions. Then it deals with fuzziness and subjectivity. Through the affiliation function, qualitative evaluations are quantified into manageable mathematical models [16].The fuzzy evaluation method also exhibits significant adaptability in handling intricate and variable teaching scenarios. When integrated with the features of immersive contextual teaching, assessing its outcomes typically entails a multitude of complex and multidimensional factors (such as language proficiency, cultural comprehension, emotional attitudes, etc.), aligning well with the fundamental attributes of fuzzy evaluation [17]. Hence, utilizing fuzzy evaluation for gauging the effectiveness of immersive contextual English instruction can address the limitations of conventional assessment techniques and offer innovative perspectives for teaching effectiveness research. Fig. 1 delineates the construction phases of the fuzzy evaluation model, depicting the entire procedure from identifying evaluation indicators to executing the fuzzy synthesis operation. This aids users in comprehending how to apply the fuzzy evaluation approach to assess the impact of immersive contextual teaching [18]. Initially, it is crucial to establish the evaluation indicator system and choose the evaluation dimension that resonates with the attributes of immersive contextual teaching [19]. Subsequently, the membership function for each evaluation indicator is ascertained to convert qualitative assessments into a quantifiable range of values [20]. Thereafter, the weight of each evaluation indicator is determined based on expert opinions, student feedback, or pertinent literature [21]. Building on this, the fuzzy comprehensive evaluation matrix is formulated by integrating the membership function and weight of each indicator, and the membership values of each evaluation object across each dimension are aggregated into a matrix. Following this, the constructed fuzzy comprehensive evaluation matrix is manipulated to derive the overall score for each evaluation object. Lastly, in accordance with the operational outcomes, the final evaluation results are generated and can be displayed in formats such as scores or grades, assisting decision-makers in making informed judgments.

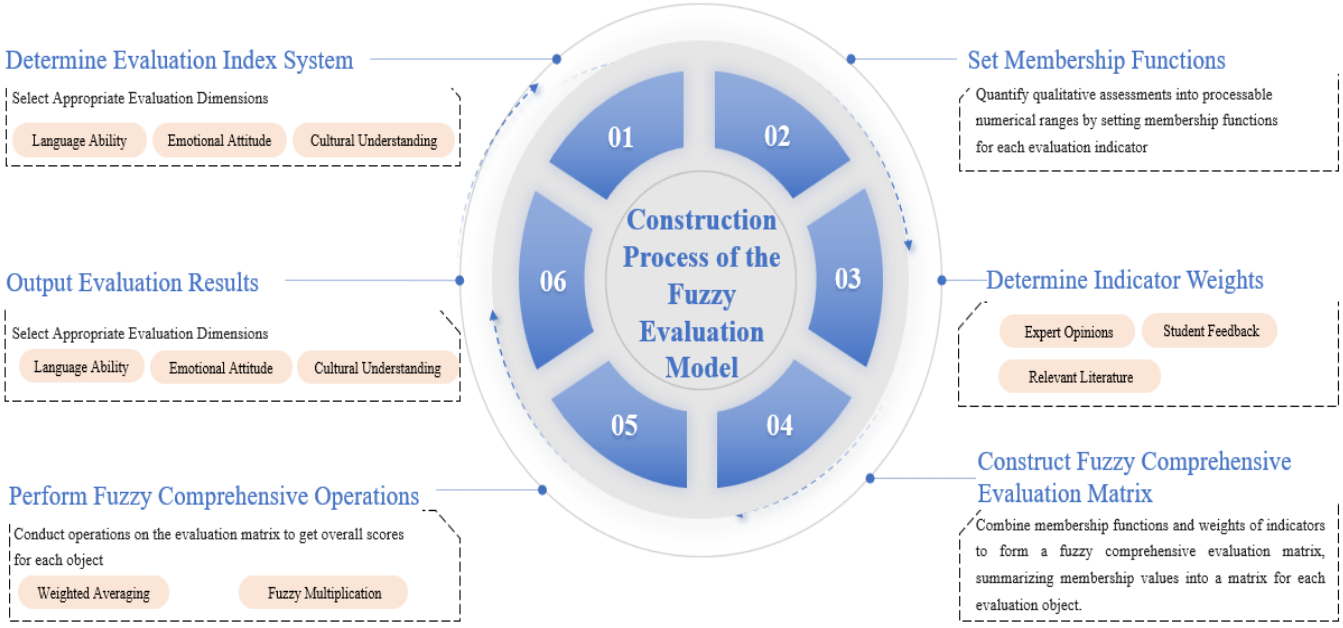


Fig. 1. Construction process of fuzzy evaluation model.

C. Review of the Study and Innovations

To summarize, immersive contextual teaching has received widespread attention in recent years due to its remarkable language-learning effect, but there are still major limitations in the evaluation methods of teaching effectiveness; fuzzy evaluation methods have opened up new paths for educational evaluation research due to their ability to handle multidimensional and complex data and their flexibility [22]. However, few studies have organically combined the two and conducted a comprehensive empirical analysis of actual cases in immersive contextual teaching. A comparison of the innovations of this study with previous studies is shown in Table II. Based on the above analysis, the innovations of this paper are mainly reflected in the following two aspects. (1) Introducing the fuzzy evaluation method into the assessment of the effect of immersive contextual teaching and constructing a multidimensional and highly adaptable fuzzy evaluation model. (2) Through empirical research and data analysis, the validity of the model is verified from actual cases to fill the shortcomings of existing research.

TABLE II COMPARISON OF THE STUDY'S INNOVATIONS WITH PREVIOUS STUDIES

Research	Previous Study	Innovative Points of This Study
Evaluation methods for immersive contextualized instruction	(1) Most studies use traditional quantitative assessments (e.g., achievement assessments, comparison of standardized test scores) (2) or qualitative assessment (e.g., interviews, classroom observations)	Introducing the fuzzy evaluation method into the assessment of immersive contextual teaching, breaking through the limitations of the traditional single method, and constructing a multi-dimensional and highly adaptable fuzzy evaluation model
Comprehensive assessment of teaching effectiveness	Most of the existing research focuses on single dimensions or quantitative indicators, such as academic performance or student feedback, and lacks an integrated approach to assessment.	Based on fuzzy evaluation, a comprehensive assessment of teaching effectiveness is realized through a comprehensive analysis of multiple evaluation dimensions (e.g. language proficiency, cultural understanding, affective attitudes, etc.).
Empirical research and data analysis	Few studies have empirically analyzed real-world cases for immersive contextual instruction.	This study combines real cases with in-depth empirical research to verify the effectiveness of the fuzzy evaluation model in immersive contextual teaching through data analysis.
Evaluation of Teaching Effectiveness Flexibility and Adaptability of	Traditional evaluation methods are often difficult to deal with in complex and dynamically changing teaching and learning contexts.	Fuzzy evaluation methods are flexible and adaptable to complex, dynamic teaching and learning contexts, and are highly adaptable, especially for assessing the effectiveness of immersive contexts.

III. METHODOLOGY

To explore the application of the fuzzy evaluation method in the assessment of English teaching effectiveness in immersive contexts, this study adopts an empirical-based research methodology to construct and validate the fitness model through case studies and data analysis. This section will specify the research design, sample selection, data collection and processing, construction of the fuzzy evaluation model, and its calculation process.

A. Study Design

This study adopts a mixed research method, combining qualitative and quantitative analysis to ensure both an in-depth analysis of the complex teaching phenomenon and the scientific and persuasive nature of the data results. The flow chart of the research design is shown in Fig. 2. The whole research process is divided into four main stages. The first is the contextual teaching implementation phase, in which a school's English course is selected and virtual reality technology is introduced to create multiple simulated learning scenarios, such as ordering food, medical conversations, and international conferences, to provide students with an immersive learning experience. Then comes the data collection phase, in which multidimensional data including standardized language test scores, classroom observation records, student interviews, and questionnaire feedback are collected after the implementation of teaching [23]. The third stage is the fuzzy evaluation model construction, which establishes a suitable fuzzy comprehensive evaluation model to quantify and comprehensively analyze the collected multidimensional data according to the research questions and data characteristics [24]. The final stage is the data analysis and validation stage, which uses tools such as Matlab and Python to process the data, calculate the specific scores of the fuzzy evaluation model, and carry out model adaptation validation [25]. This flowchart provides clear steps and directions for the research and helps to systematically conduct and analyze the research.

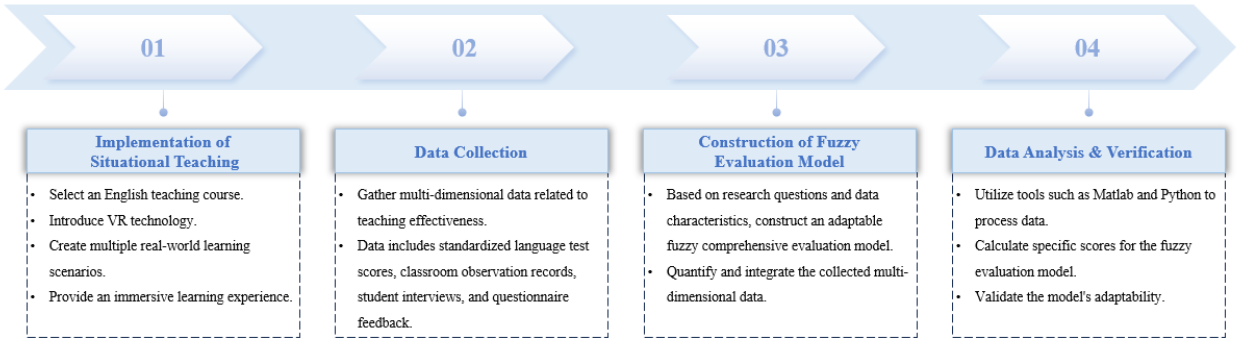


Fig. 2. Flowchart of the study design.

B. Sample Selection and Background

The participants in this study were 120 first-year college students majoring in English, who were evenly divided into an experimental group and a control group, with 60 students in each [26]. The experimental group was subjected to immersive contextual English instruction, while the control group followed the conventional English teaching approach. To reduce the potential impact of sample heterogeneity on the study outcomes, the two groups were matched in terms of age, gender, and English proficiency, which was categorized based on their entrance exam scores [27]. The specific details of the experimental and control groups are presented in Table III. As indicated in the table, both groups comprised 60 participants with average ages of 18.7 and 18.8 years, respectively. A t-test revealed that the age difference between the two groups was insignificant (p=0.653). In terms of gender distribution, the experimental group had 32 males and 28 females, whereas the control group had 31 males and 29 females. A chi-square test indicated no significant difference in gender ratio between the two groups (p=0.853). The mean entrance exam scores for English were 82.3 and 81.9 for the two groups, respectively, and a t-test showed no significant difference in English proficiency (p=0.712). Regarding family background, 35 participants in the experimental group were from urban areas and 25 from rural areas; in the control group, 36 were from urban areas and 24 from rural areas. A chi-square test demonstrated no significant difference in family background between the groups (p=0.896). The duration of English study was 6.5 years and 6.4 years for the two groups, respectively, and a t-test found no significant difference in this regard (p=0.578). In summary, the experimental and control groups exhibited no significant differences in any of the aforementioned basic characteristics, suggesting that the two groups were highly comparable at the outset of the study.

TABLE III SAMPLE GROUPING AND BASIC STATISTICAL CHARACTERISTICS

Variant	Experimental Group (N=60)	Control Group (N=60)	Statistical Test Value	P-Value
Average age (years)	18.7 ± 0.9	18.8 ± 0.8	t = 0.45	0.653
Gender ratio (M/F)	32/28	31/29	$\chi^2 = 0.034$	0.853
The average score on the English test for admission	82.3 ± 5.7	81.9 ± 6.1	t = 0.37	0.712
Family background (urban/rural)	35/25	36/24	$\chi^2 = 0.017$	0.896
Length of English language study (years)	6.5 ± 1.1	6.4 ± 1.0	t = 0.56	0.578

C. Data Collection and Processing

1) Data collection: The data collection for this study covered two main types of data: quantitative and qualitative. Quantitative data were mainly obtained through standardized English test scores, which consisted of four parts: listening, reading, writing, and speaking, and were designed to comprehensively assess students' English proficiency [28]. Qualitative data collection is more diverse and includes instructional observation records, student interviews, and questionnaires. Observation records focus on student engagement and interaction in the classroom to capture the dynamics of the teaching and learning process. Student interviews were conducted to explore students' affective experiences and learning outcomes to obtain their direct feedback on teaching methods and content [29]. In addition, student satisfaction, self-confidence, and changes in interest in teaching and learning were assessed through questionnaires, the content of which helped to quantitatively analyze students' attitudes and affective responses to learning [30]. Synthesizing the data collected through multiple channels, we were able to gain a comprehensive and in-depth understanding of students' learning status and teaching effectiveness, which provided solid data support for the study.

2) Data pre-processing: The collected data were first subjected to data cleaning and pre-processing through SPSS, including outlier removal and data normalization. The data normalization equation is as follows, for the original data x, normalized to:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

Where $X_{\min}$ and $X_{\max}$ are the minimum and maximum values of the sample data, respectively?

D. Fuzzy Evaluation Model Construction

Based on the core logic of the fuzzy evaluation method, the model construction is divided into the following steps:

1) Determination of the evaluation indicator system: Combining the characteristics of the effect of immersion teaching and previous studies, an evaluation system including four first-level indicators and several second-level indicators is established [31]. The first-level indicators include: language ability (A_1), learning attitude (A_2), learning interest (A_3), and emotional experience (A_4). The secondary indicators include listening (A_{11}), speaking (A_{12}), reading and writing sub-competencies in language proficiency, classroom participation, and change in language anxiety. The weights of the first-level indicators are noted as $W = [w_1, w_2, \ldots, w_n]$, where w_i meet $\sum w_i = 1$. Each secondary indicator weight is obtained by hierarchical analysis (AHP) and distributed in the evaluation system in the following way. Table IV shows the weight distribution table of the evaluation indicators, with a total of four first-level indicators (A1, A2, A3, A4), whose sum of

weights is 1, reflecting the importance weights of different dimensions in the total evaluation. Each level 1 indicator is subdivided into several level 2 indicators, and the weight distribution is calculated by the hierarchical analysis method (AHP) to ensure rationality and scientificity [32]. The sum of the weights of the second-level indicators satisfies the corresponding weights of the first-level indicators. The distribution of the indicators reflects the multidimensional nature of the immersive contextual teaching evaluation model. For example, language proficiency is given the highest weight (40%), with higher weights for listening and speaking, indicating its centrality to teaching effectiveness.

TABLE IV DISTRIBUTION OF WEIGHTS OF EVALUATION INDICATORS

Primary Indicators	Weight	Secondary Indicators	Weight
Language Ability (A1)	WA1=0.4	Listening (Listening, A11)	W*A* 11=0.15
		Speaking (Speaking,*A* 12)	W*A* 12=0.15
		Reading (Reading,*A* 13)	WA13=0.05
		Writing (Writing,*A* 14)	WA14=0.05
Learning Attitude (A2)	WA2=0.3	Classroom Participation,*A* 21)	WA21=0.2
		Task Completion (Task Completion,*A* 22)	WA22=0.1
Learning Interest, A3)	WA3=0.2	Interest Improvement (Interest Improvement,*A* 31)	WA31=0.2
Emotional Experience (A4)	WA4=0.1	Language Anxiety Change (Language Anxiety Change,*A* 41)	WA41=0.1

2) Construct fuzzy affiliation function and affiliation matrix: Each secondary indicator corresponds to a different evaluation grade (e.g., excellent, good, fair, poor), and defines the affiliation function for each grade. Define the evaluation level set $\mathrm{V}=\{\mathrm{v}_1,v_2,...,v_m\}$, the corresponding evaluation levels include "excellent", "good", "fair", "poor", etc. For each indicator, define its fuzzy affiliation function. For each indicator A_{ij} , define its fuzzy affiliation function $\mu_{\mathrm{ij}}(x)$. In this study, triangular fuzzy numbers are used for evaluation in the following form:

$$\mu_{\mathrm{ij}}(x)\begin{cases}0, & x\le a \text{ or } x\ge c,\\ \dfrac{\mathrm{x}-a}{b-a}, & \mathrm{a}\le \mathrm{x}\le \mathrm{b},\\ \dfrac{c-x}{c-b}, & b\le x\le c\end{cases} \tag{2}$$

Where*a*, *b* and *c* are the starting point, median and end point of the fuzzy number respectively. For each evaluated object, the affiliation value of each evaluation index is calculated through the data of students' test scores, classroom records, and questionnaire feedback, and the affiliation is formed.

$$\mathrm{R}=\begin{bmatrix}\mu_{11}(v_1) & \mu_{11}(v_2) & \cdots & \mu_{11}(v_m)\\ \mu_{12}(v_1) & \mu_{12}(v_2) & \cdots & \mu_{12}(v_m)\\ \vdots & \vdots & \ddots & \\ \mu_{1n}(v_1) & \mu_{1n}(v_2) & \cdots & \mu_{1n}(v_m)\end{bmatrix} \tag{3}$$

Where each row of R represents the affiliation of a secondary indicator to a different evaluation level.

3) Fuzzy synthesis operation: Based on the affiliation matrix R and the weight vector W , a fuzzy comprehensive evaluation is performed by the following equation:

$$B=W\cdot R=\left[\mathrm{b}_1,b_2,\cdots,b_m\right] \tag{4}$$

Where B is the comprehensive affiliation vector, which represents the comprehensive evaluation affiliation of the evaluation object on different levels; W is the weight vector; and *R* is the affiliation matrix. To quantify the evaluation results, the comprehensive affiliation vector *B* is normalized and the fuzzy score is calculated:

$$\mathrm{S}=\sum_{i=1}^{m}b_i\square v_i \tag{5}$$

Where S is the final score of the fuzzy comprehensive evaluation and V_{i} is the score corresponding to the *i* evaluation level (e.g., 100 for "excellent" and 80 for "good").

E. Model Validation

To verify the applicability of the fuzzy evaluation model, this study conducted a correlation analysis between the model calculation results and the standardized test scores and further verified the reasonableness and scientificity of the model output results through expert review and student feedback. The correlation between the fuzzy composite scores and the standardized test scores was examined through Pearson correlation coefficient analysis to assess the reliability and validity of the model. The equation is as follows:

$$r = \frac{\sum(X_i - \overline{X})\sum(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \cdot \sum(Y_i - \overline{Y})^2}} \quad (6)$$

Where X_i and Y_i are the academic fuzzy score and test score, respectively, and $\overline{X}$ $\overline{Y}$ are the corresponding mean values. Based on the experimental data, the model parameters (e.g., the shape of the affiliation function and the weight distribution) are adjusted to ensure high consistency between the model evaluation results and the actual situation.

IV. Results and Discussion

Through the fuzzy evaluation of the teaching effectiveness of the experimental and control groups, this study reveals the effectiveness of immersion contextualized teaching in enhancing students' language proficiency and learning attitudes [33]. In this part, specific discussions will be made around the results of the data analysis, including the results of the fuzzy comprehensive evaluation, the analysis of the differences between the groups, and the substantive interpretation of the teaching effectiveness.

A. Fuzzy Synthesized Evaluation Results

Utilizing the previously mentioned fuzzy evaluation model, this research carried out an extensive assessment of the teaching outcomes for both the experimental and control groups, and computed their respective final fuzzy scores, as depicted in Fig. 3. This figure contrasts the comprehensive scores of the two groups. The experimental group's comprehensive fuzzy score was notably higher than that of the control group, suggesting that immersive contextual teaching can substantially enhance students' overall language abilities, particularly in communicative competencies like listening and speaking, where the benefits are more pronounced [34]. In the graph, the horizontal axis denotes the Overall Score, while the vertical axis indicates the Group, with the control group symbolized by green dots and the experimental group by purple dots. The graph reveals that the experimental group generally achieved a higher Overall Score than the control group. The score distribution in the experimental group is more clustered, with the majority of scores exceeding 80, whereas the control group's scores are more scattered, primarily ranging between 70 and 80. The experimental group's scores are more tightly packed within the 80 - 90 range, indicating greater consistency and superior performance. In contrast, the control group's scores are more diffused, spanning from 60 to 90, reflecting larger individual variations. The density curves at the base of the figure further illustrate the score distributions of both groups. The experimental group's density curve is more peaked, with the apex situated around 80 points, while the control group's density curve is flatter, with the peak around 70 points. There is a marked disparity in overall scores between the experimental and control groups, with the experimental group markedly outperforming the control group. This could imply that the pedagogical approaches or interventions employed in the experimental group were more efficacious in elevating students' overall scores.

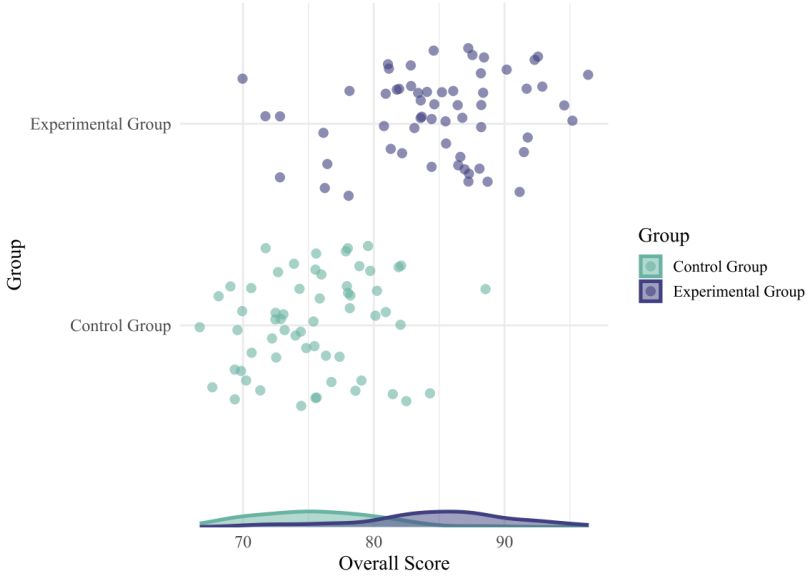


Fig. 3. Comparison of composite scores between experimental and control groups.

B. Analysis of differences between Indicators

Further analysis of the scores for the primary and secondary indicators reveals the following salient features:

1) Effective in improving language skills: The fuzzy evaluation indicates that the experimental group outperformed the control group across all language proficiency subcategories, namely listening, speaking, reading, and writing. For instance, in the listening subcategory, the experimental group's membership distribution predominantly centers on the "excellent" and "good" levels, whereas the control group's distribution is more concentrated on the "good" and "fair" levels. This suggests that immersive contextual teaching facilitates

students' quicker adaptation to real - life language settings and enhances their language comprehension abilities. Table V presents the performance membership matrices for the experimental and control groups in listening and speaking, featuring four levels of membership distribution: excellent, good, fair, and poor. Regarding listening, the experimental group's membership values for excellent and good grades are 0.45 and 0.40, respectively, which are markedly higher than those of the control group (0.10 and 0.60). In terms of speaking ability, the experimental group's membership values for excellent and good grades are 0.50 and 0.35, respectively, also surpassing those of the control group (0.15 and 0.55). This demonstrates that the experimental group exhibits superior overall performance in both listening and speaking, especially in the excellent grade category. Notably, the disparity is more pronounced in the excellent grade performance. Moreover, the membership values for the poorer grades in both groups are 0.05, indicating a low proportion of poor performance.

TABLE V LISTENING AND SPEAKING AFFILIATION MATRIX

Group	Excellent	Favorable (Good)	General (Average)	Mediocre (Poor)
Listening - Experimental Group	0.45	0.40	0.10	0.05
Listening - Control Group	0.10	0.60	0.25	0.05
Speaking - Experimental Group	0.50	0.35	0.10	0.05
Speaking - Control Group	0.15	0.55	0.25	0.05

2) Improvement of positive learning attitudes: The survey results revealed that students in the experimental group exhibited markedly higher classroom engagement, enthusiasm for learning, and associated satisfaction scores compared to the control group (refer to Fig. 4 below), which utilizes a scatter plot coupled with a trend smoothing line to illustrate a contrast between the experimental and control groups regarding learning attitude scores. As depicted in the figure, the experimental group generally achieved higher learning attitude scores than the control group. Despite both groups having a certain degree of score distribution, the experimental group's scores were more concentrated and displayed an upward trend, indicating that students in this group demonstrated a more proactive and positive approach to learning [35]. Conversely, the control group's scores were relatively lower and more scattered, suggesting that their learning attitudes might be less favorable than those of the experimental group. Moreover, the experimental group had a significantly greater number of individuals with high ratings than the control group, further corroborating the efficacy of the experimental teaching method in bolstering students' learning attitudes [36]. In summary, the figure furnishes visual substantiation for the study that the teaching interventions in the experimental group yielded substantial outcomes in enhancing students' attitudes towards learning. This underscores the efficacy of contextualized teaching in stimulating students' intrinsic motivation.

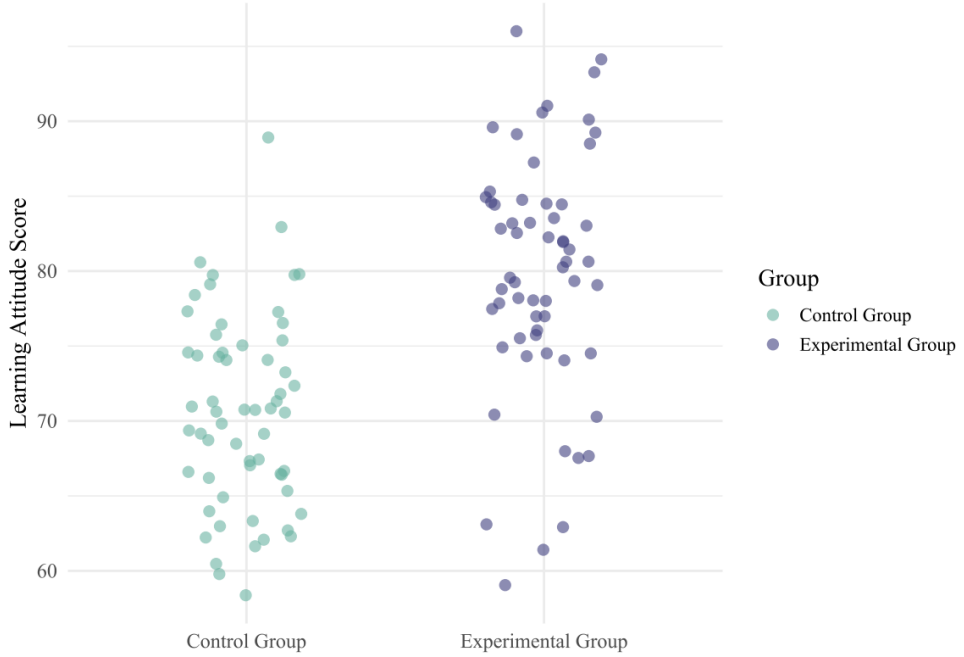


Fig. 4. Statistical comparison of learning attitude indicators between groups.

C. Comparison of Teaching Effectiveness between Experimental and Control Groups

To present a more comprehensive picture of the differences in the effectiveness of immersive contextualized instruction versus traditional instruction, this study conducted a t-test on the composite scores of the two samples, yielding in the Fig. 5.

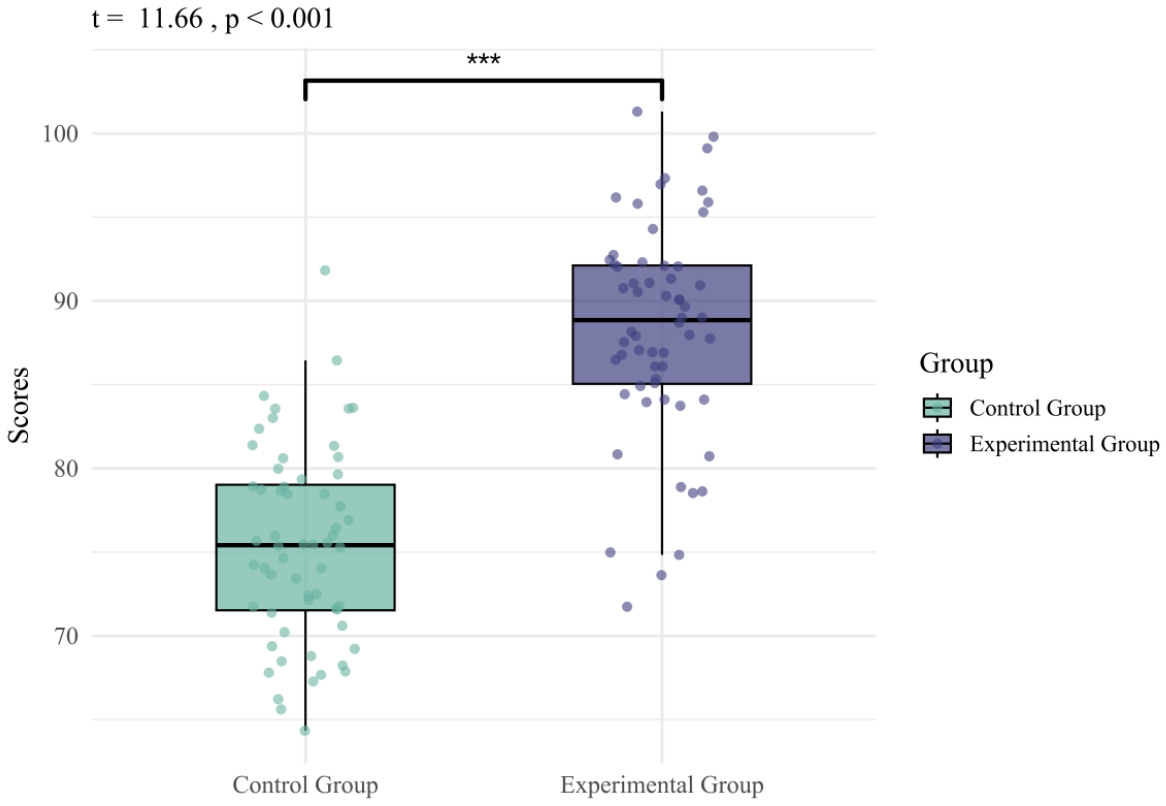


Fig. 5. Graph comparing the effect of experimental and control groups.

The figure provides a detailed comparison of teaching effectiveness between the experimental and control groups, depicted through a box-and-line plot. It is evident that the experimental group's score distribution is more compact and features a higher median, suggesting superior overall performance compared to the control group. Specifically, the experimental group's median score is nearly 90, whereas the control group's median score hovers around 75, highlighting a distinct advantage in teaching effectiveness for the experimental group. Furthermore, the experimental group's score range is relatively narrower, indicating more stable and less variable student performance. The statistical analysis results corroborate this observation. The independent samples *t*-test results revealed a highly significant difference between the two groups (t=11.66, p<0.001), indicating that the experimental group was substantially more effective than the control group. This level of significance (p<0.001) implies that the difference between the experimental and control groups is improbable to be attributed to random factors, but rather to the efficacy of the teaching methods or interventions employed by the experimental group.

D. Interpretation and Discussion of Results

1) Advantages of immersive contextualized instruction: In this study, the effects of immersive contextual teaching were analyzed in depth through the fuzzy evaluation method, and the results showed that this teaching mode can effectively improve students' language communication skills. This finding is consistent with the existing literature on the importance of authenticity of teaching contexts for students' language acquisition [37]. What's more, this study is the first attempt to quantitatively analyze teaching effectiveness through the fuzzy evaluation method, which provides a new perspective and methodology for research in this field. The results of this quantitative analysis not only enhance the scientific nature of teaching research but also provide precise data support for actual teaching design, which helps educators grasp the teaching effect more accurately to optimize teaching strategies.

2) Application value of fuzzy evaluation: Fuzzy evaluation shows its unique advantages in assessing complex teaching phenomena, especially in the multidimensional comprehensive analysis [38]. For example, when assessing highly subjective indicators such as "emotional experience", the affiliation matrix of fuzzy evaluation can effectively describe the distribution of student's satisfaction with the classroom, thus reducing the bias that may be brought by a single scoring model [39]. This method can reflect students' subjective feelings and learning experiences more comprehensively, providing a richer and more detailed perspective for the evaluation of teaching effectiveness.

3) Research limitations and future directions: While this study has yielded significant findings regarding the assessment of teaching effectiveness, it is not without limitations. Firstly, the relatively small sample size may impinge upon the generalizability of the results. Subsequent research endeavors could enhance the generalizability and robustness of the model presented herein by increasing the sample size. Secondly, the parameter configurations of the fuzzy evaluation model, including the form of the membership function, warrant further refinement to augment the model's precision and versatility [40]. Moreover, future investigations might delve into the applicability of fuzzy evaluation techniques across diverse teaching contexts and subject domains, thereby broadening the scope and depth of their utilization. Through such initiatives, it is anticipated that fuzzy evaluation methods will assume a more pivotal role in the realm of teaching effectiveness assessment, offering enhanced support for educational research and practice.

V. Conclusion and Limitations

This study adopts a fuzzy evaluation method to comprehensively analyze the effectiveness of immersive contextual English teaching. By constructing a scientific fuzzy evaluation model, this study assessed the effectiveness of the teaching method in terms of two key dimensions, namely, language proficiency and learning attitude, and proved the significant advantages of immersive contextual teaching in a practical application through empirical research. The results of the study show that immersive contextual teaching is effective in enhancing students' language proficiency. The composite scores of students in the experimental group were significantly higher than those of the control group in each language proficiency index, such as listening, speaking, reading, and writing. This result shows that contextualized teaching can effectively improve students' language practice ability by creating a real language environment. In addition, immersion teaching also shows positive effects in enhancing students' learning attitudes. The results of the fuzzy evaluation show that immersion teaching is effective in enhancing students' motivation, classroom participation, and interest in learning. Scenario simulation and interactive experience can better stimulate students' intrinsic motivation and thus strengthen learning effects. This study also introduces the fuzzy evaluation method into the assessment of English teaching effectiveness and verifies the application value of this method. This method provides a new way of thinking and methodology for the teaching evaluation system, which can quantify the complex teaching process more comprehensively, especially when it involves the comprehensive analysis of multi-dimensional data, which has significant advantages. To summarize, immersive contextual English teaching not only reflects high efficiency in language proficiency cultivation but also achieves positive feedback in students' learning attitudes and subjective experiences, providing important insights for educational reform and practice. The application of the fuzzy evaluation method further highlights the importance of quantitative analysis in teaching research, which is of theoretical promotion and practical guidance significance.

However, there are still some shortcomings in this study. On the one hand, the relatively small size of the experimental sample may affect the external validity of the conclusions; future research can expand the sample coverage and select subjects from students of different age groups and different language bases to verify the generalizability of the findings. On the other hand, the fuzzy evaluation model is somewhat subjective in the selection of the affiliation function; in the future, attempts can be made to optimize the parameter settings of the model by introducing machine learning or artificial intelligence algorithms. In addition, this study mainly focuses on the two dimensions of language proficiency and learning attitude, and subsequent studies can further explore the effects of contextualized teaching on higher-order language proficiency such as critical thinking and intercultural communication skills.

References

[1] Al-Gerafi M, Goswami S, Khan M. Designing of an effective e-learning website using inter-valued fuzzy hybrid MCDM concept: A pedagogical approach[J]. *Alexandria Engineering Journal*, 2024, 97: 61–87.

[2] Jiang L. Research on the integration path and practice of AI intelligent technology and English distance education[J]. *Sciendo*, 2023, 45(6): 23–50. doi: 10.2478/amns.2023.2.01427.

[3] Al-kfairy M, Ahmed S, Khalil A. Factors impacting users' willingness to adopt and utilize the metaverse in education: a systematic review[J]. *Comput Hum Behav Rep*, 2024, 76(56): 100459.

[4] Al-Samarraay M, Salih M, Ahmed M. A new extension of FDOSM based on Pythagorean fuzzy environment for evaluating and benchmarking sign language recognition systems[J]. *Neural Comput Appl*, 2022, 56(5): 1–19.

[5] Caiado R, Scavarda L, Gavião L. A fuzzy rule-based industry 4.0 maturity model for operations and supply chain management[J]. *Int J Prod Econ*, 2021, 231: 107883. doi: 10.1016/j.ijpe.2020.107883.

[6] Mikhailenko M, Maksimenko N, Kurushkin M. Frontiers | eye-tracking in immersive virtual reality for education: a review of the current progress and applications[J]. *Sciendo*, 2023, 34(4): 13–25. doi: 10.3389/feduc.2022.697032.

[7] Cheng H, Zhu L, Meng J. Fuzzy evaluation of the ecological security of land resources in mainland China based on the pressure-state-response framework[J]. *Sci Total Environ*, 2022, 804: 150053.

[8] Chien C, Ho Y, Hou H. Integrating immersive scenes and interactive contextual clue scaffolding into a decision-making analysis ability training game[J]. *J Educ Comput Res*, 2024, 62(1): 376–405.

[9] Purnama Y, Fransiska F, Muhdi A. Long-term effects evaluation of using artificial intelligence-based automated learning systems in improving English content understanding at the secondary education level[J]. *Indones J Educ (INJOE)*, 2023, 3(3): 622–636.

[10] Chua C, Kosnin A, Yeo K. Fuzzy Delphi method for a-level mathematics technological pedagogical and content knowledge module[J]. *Int J Eval Res Educ*, 2024, 13(1): 441–453.

[11] Yuan R. The other side of the coin: A socio-cultural analysis of pre-service language teachers' learning to teach critical thinking[J]. *Think Skills Creat*, 2023, 48: 101265. doi: 10.1016/j.tsc.2023.101265.

[12] Restall G, Yao Y, Niu X. Exploring the experience of year 10 South Korean students' English language learning in immersive virtual reality[J]. *TESOL Context*, 2023, 31(2): 21–67.

[13] Yuan R. Cultivating CT-oriented teachers in pre-service teacher education: what is there and what is missing?[J]. *Taylor Fr*, 2023, 76(9): 67–99.

[14] Sanfilippo F, Blazauskas T, Salvietti G. A perspective review on integrating VR/AR with haptics into STEM education for multi-sensory learning[J]. *Robotics*, 2022, 11(2): 41. doi: 10.3390/robotics11020041.

[15] Meccawy M, Alzahrani A, Mattar Z. Assessing EFL students' performance and self-efficacy using a game-based learning approach[J]. *Educ Sci*, 2023, 13(12): 1228. doi: 10.3390/educsci13121228.

[16] Curran V, Xu X, Aydin M. Use of extended reality in medical education: an integrative review[J]. *Med Sci Educ*, 2022, 33(1): 275–286. doi: 10.1007/s40670-022-01698-4.

[17] Yuchen X. Application of immersive artificial intelligence based on machine vision in education management of children with autism[J]. *Int J Syst Assur Eng Manag*, 2023, 28(4): 1–10.

[18] Fang C. Intelligent online English teaching system based on SVM algorithm and complex network[J]. *J Intell Fuzzy Syst*, 2021, 40(2): 2709–2719.

[19] Shi L, Muhammad Umer A, Shi Y. Utilizing AI models to optimize blended teaching effectiveness in college-level English education[J]. *Cogent Education*, 2023, 10(2): 2282804.

[20] Manabe K, Hwang W, Chuang Y. English learning is enhanced by collaborative contextual drama in an authentic context[J]. *Interact Learn Envir*, 2023, 31(7): 4490–4506.

[21] Zhou S. Gamifying language education: the impact of digital game-based learning on Chinese EFL learners[J]. *Humanit Soc Sci Commun*, 2024, 11(1): 1–14. doi: 10.1057/s41599-024-04073-3.

[22] Lu C, He B, Zhang R. Evaluation of English interpretation teaching quality based on GA optimized RBF neural network[J]. *J Intell Fuzzy Syst*, 2021, 40(2): 3185–3192.

[23] Li Y. The digital transformation of college English classroom: application of artificial intelligence and data science[J]. *EAI Endorsed Trans Scalable Inf Syst*, 2024, 11(5): 23–53.

[24] Song C, Shin S-Y, Shin K-S. Optimizing foreign language learning in virtual reality: a comprehensive theoretical framework based on constructivism and cognitive load theory (VR-CCL)[J]. *Appl Sci*, 2023, 13(23): 12557.

[25] Zhang L. An IoT-based English translation and teaching using particle swarm optimization and neural network algorithm[J]. *Soft Comput*, 2023, 27(19): 14431–14450.

[26] Zhang Y. An analytical study on the design and implementation of a comprehensive web-based learning environment for Chinese as a foreign language[J]. *Soft Computing*, 2023, 27(23): 18147–18164.

[27] Li N. A fuzzy evaluation model of college English teaching quality based on analytic hierarchy process[J]. *Int J Emerg Technol Learn (iJET)*, 2021, 16(2): 17–30.

[28] Khalaf OI, Srinivasan D, Algburi S. Elevating metaverse virtual reality experiences through network-integrated neuro-fuzzy emotion recognition and adaptive content generation algorithms[J]. *Eng Rep*, 2024, 34(10): e12894.

[29] Wang H, Wang J, Wang G. A survey of fuzzy clustering validity evaluation methods[J]. *Inf Sci*, 2022, 618: 270–297.

[30] Zolghadri M, Jafarpour Mamaghani H. Pathology of language teaching ineffectiveness: a case study exposing teacher cognition stepwise[J]. *J Mod Res Engl Lang Stud*, 2021, 9(1): 29–51.

[31] Jin S, Huang J, Zhong Z. Application of immersive technologies in primary and secondary education[J]. *Front Digit Educ*, 2024, 1(2): 142–152.

[32] Huang J, Sang G. Conceptualising critical thinking and its research in teacher education: a systematic review[J]. *Teach Teach*, 2023, 29(6): 638–660.

[33] He Q, Attan S, Zhang J. Evaluating music education interventions for mental health in Chinese university students: a dual fuzzy analytic method[J]. *Sci Rep*, 2024, 14(1): 19727.

[34] Wang W, Huang S. The application of artificial intelligence teaching software in college English teaching[J]. *Sciendo*, 2021, 34(34): 44–62. doi: 10.2478/amns.2023.2.00657.

[35] Han L. Students' daily English situational teaching based on virtual reality technology[J]. *Mobile Inf Syst*, 2022, 2022(1): 1222501.

[36] Zheng B. Translanguaging in a Chinese immersion classroom: an ecological examination of instructional discourses[J]. *Int J Biling Educ Bi*, 2021, 133(34): 126–144.

[37] Xie Y, Liu Y, Zhang F. Virtual reality-integrated immersion-based teaching to English language learning outcome[J]. *Front Psychol*, 2022, 12: 767363.

[38] Weng Y, Schmidt M, Huang W. The effectiveness of immersive learning technologies in K–12 English as second language learning: a systematic review[J]. *Recall*, 2024, 45(6): 1–20.

[39] Gong J, Liu H, You X. An integrated multi-criteria decision-making approach with linguistic hesitant fuzzy sets for E-learning website evaluation and selection[J]. *Appl Soft Comput*, 2021, 102: 107118.

[40] Fernández-Herrero J. Evaluating recent advances in affective intelligent tutoring systems: A scoping review of educational impacts and prospects[J]. *Humanities and Social Sciences Communications*, 2024, 122(16): 67–98. doi: 10.3390/educsci14080839.

Multi-Classification Convolution Neural Network Models for Chest Disease Classification

Noha Ayman[1], Mahmoud E. A. Gadallah[2], Mary Monir Saeid[3]
Department of Computer Science-Faculty of Computers and Artificial Intelligence, Fayoum University, Egypt[1]
Egypt Department of Computer Science, Modern Academy for Computer Science and Management Technology, Cairo, Egypt[2]
Department of Information System-Faculty of Computers and Artificial Intelligence, Fayoum University, Egypt[3]

***Abstract*—Chest diseases significantly affect public health, causing more than one million hospital admissions and approximately 50,000 deaths annually in the United States. Chest X-ray imaging technology, which is a critically important imaging technique, helps in examining, diagnosing, and managing chest conditions by providing essential insights about the presence and severity of disease. This study introduces a novel chest X-ray classification framework leveraging a fine-tuned VGG19 model (16 layers) enhanced with CLAHE for improved contrast, binary mask attention to highlight abnormalities and advanced data augmentation for better generalization. Key innovations include the use of a Probabilistic U-Net for lung segmentation to isolate critical features and weighted masks to focus on pathological regions, addressing class imbalance with computed class weights for fair learning. By achieving 95% accuracy and superior class-specific metrics, the proposed method outperforms existing deep learning approaches, providing a robust and interpretable solution for real-world healthcare applications, where a test accuracy of 94.8% is achieved using different customized models based on VGG19 without using a mask. The experimental results indicate that our proposed method surpasses current deep learning techniques in terms of overall classification accuracy for chest disease detection.**

Keywords*—Convolution neural network; classification; chest X-ray; image preprocessing; U-Net; deep learning

I. Introduction

Chest pain is the most common reason for consultations and emergency room visits. Globally, chest radiography is the most frequently used imaging examination, essential for the screening, diagnosis, and management of numerous life-threatening thoracic conditions. The expertise and observational skills of radiologists are crucial for interpreting chest X-rays (CXRs). However, the complexity of the pathologies and the subtle differences in lung lesions mean that even experts can sometimes miss minute details. Additionally, there is a shortage of trained and experienced radiologists. Consequently, recent research has focused on developing systems to detect thoracic diseases and generate reports. These studies predominantly employ deep learning and neural network models. This paper aims to explore these diseases accurately.

The chest is the upper part of the trunk. It gets support from the rib cage, the girdle of the shoulder, and the spine that also protects it. It is the region of the body formed by the sternum, the thoracic vertebrae, and the ribs [1]. It resides between the neck and diaphragm excluding the upper limb. The heart and lungs reside in the thoracic cavity, as well as many blood vessels that play a vital role in feeding (esophagus), breathing, and pumping blood to all parts of the body [2].

Deep learning techniques that implement deep neural networks became popular due to the increase in high-performance computing facilities. Deep learning achieves higher power and flexibility due to its ability to process a large number of features when it deals with unstructured data [3].

Deep learning models have been used successfully in many areas such as classification, segmentation, and lesion detection of medical data. Analysis of image and signal data obtained with medical imaging techniques such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and X-ray with the help of deep learning models [4]. As a result of these analyses, detection and diagnosis of diseases such as diabetes mellitus, brain tumor, skin cancer, and breast cancer are provided convenience [5].

The contributions of this research can be summarized as follows:

- A method has been devised based on the VGG19 model for the classification of chest X-ray images as COVID, Lung Opacity, Normal, and Viral Pneumonia.
- Preprocessing of images using CLAHE, Data Augmentation, etc., has been considered for improving X-ray image quality to enhance the model performance.
- Attention mechanism is employed here by first generating binary masks that tune the model to learn regions of interest for better detection of pathological features.
- Class weighting while training was done to handle class imbalance, hence the model performed well on all categories.
- VGG19 was pre-trained on ImageNet and fine-tuned on chest X-ray images with very good generalization and a validation accuracy of 95%.

This paper is organized as follows. The Literature survey is given in Section II. Section III describes the Methodology that is used. The result is presented in Section IV. Conclusions and future work are provided in Section V.

II. Literature Survey

Many papers proposed chest disease detection and classification using deep learning techniques. For instance,

Wang et al. [6] proposes COVID-Net, a deep convolutional neural network to enable the detection of COVID-19 cases from CXR images. COVID-Net leverages an open human-machine collaborative design strategy, marrying both principled network prototyping with machine-driven exploration utilizing a novel lightweight architecture enhanced in representational capacity and computational efficiency. It was then trained on COVIDx, the benchmark dataset of 13,975 CXR images, incorporating data from five public repositories. Quantitatively, COVID-Net achieved 93.3% accuracy, 91% sensitivity, and 98.9% positive predictive value for COVID-19 detection, outperforming traditional models such as VGG-19 and ResNet-50 in terms of computational efficiency and sensitivity. A qualitative audit using the GSInquire explainability method confirmed that its decision-making relied on clinically relevant lung regions, thus providing transparency and reliability. Though limited by available data, and non-production ready, the core ideas presented in COVID-Net present a very nice starting point to advance further the use of AI approaches for COVID-19 screening, and triaging with possible extension to risk stratification.

In research [7], the author introduce CoroNet, a deep convolutional neural network (CNN) model, was introduced for detecting COVID-19 infection using chest X-ray images. Using the Xception architecture pre-trained on ImageNet dataset, the study trained CoroNet on a curated dataset of COVID-19, bacterial pneumonia, viral pneumonia, and normal chest X-rays. The model achieved an accuracy of 89.6% for four-class classification (COVID-19, bacterial pneumonia, viral pneumonia, and normal); 95% for three-class classification; and 99% for binary classification (COVID-19 vs. others). CoroNet performed better than previous studies, particularly in detecting COVID-19 patients. However, the study acknowledges its limitations, including reliance on a small dataset, and indicates the need for access to larger datasets.

In study [8], the author brings to light the importance of preprocessing the chest X-ray images for better classification of diseases. Using the ChestX-ray8 dataset, this research identifies some quality issues in images that affect classification performance. Prior methods used CNNs with preprocessing techniques such as contrast enhancement and feature extraction, often relying on metadata or manual preparation. In this paper, an automated preprocessing method is proposed that combines Sobel and Scharr edge detection with a shallow CNN to classify images into clear and low-quality images. It achieved 95% accuracy and provided a scalable solution for dataset cleaning and reduced dependence on metadata, although it had problems with extreme image defects.

In study [9], the use of transfer learning for pneumonia classification from chest X-ray pictures is examined, with a focus on differentiating between SARS-CoV-2, generic viral, and bacterial infections as the origins of pneumonia. This study assesses 12 well-known ImageNet pre-trained neural network models, building on earlier studies that frequently concentrated on binary categorization (COVID-19 versus healthy) or required particular data preprocessing. Using a dataset of 6,330 publicly sourced images, cropped to uniform dimensions, and separated into four classes—healthy pneumonia, bacterial pneumonia, viral pneumonia, and COVID-19—these models were optimized. MobileNet v3 demonstrated good classification performance and computational efficiency, with the best F1 score of 84.46%. Additionally, resilience tests were performed by reducing the amount of available training data to 50%, 20%, and 10%, showing varying levels of performance loss. The paper highlights the effectiveness of transfer learning in medical diagnosis while highlighting a number of challenges, such as model convergence issues and data scarcity. Expanding the datasets to different types of lung disorders and establishing preprocessing methods to reduce classification errors are future priorities.

In study [10], this work fills important gaps in previous work on deep learning-based COVID-19 identification, including restricted attention to binary or three-class classification problems, imbalanced datasets, and limited metric reporting. With confidence fusion, it suggests COVDC-Net, a hybrid deep learning model that combines MobileNetV2 and VGG16. On balanced datasets, it performs better, with 96.48% accuracy for three classes and 90.22% accuracy for four classes. The work overcomes the drawbacks of single architectural techniques and ensures scalability for practical clinical applications by providing comprehensive metrics for each class and enhancing model robustness through fusion.

In study [11], Sahin et al, a novel CNN model for COVID-19 detection utilizing chest X-ray pictures is proposed and tested on a dataset of 13,824 images alongside MobileNetV2 and ResNet50. After standardizing the images to 224 x 224 pixels through preprocessing, the models were trained with the Adam optimizer using the same training/testing splits (80%/20%). The proposed CNN outperformed MobileNetV2 (95.73%) and ResNet50 (91.54%) with an F1 score of 97% and test accuracy of 96.71%. It was also computationally economical, using fewer convolutional layers and parameters. The model was limited to single image slices and lacked severity classification capability, which hindered its ability to differentiate between COVID-19 and other viral pneumonia despite its high accuracy. This study highlights the potential of deep learning for COVID-19 diagnosis while addressing gaps like dataset limitations and computational efficiency.

In study [12], it presents a framework for detecting COVID-19 from chest X-ray images using transfer learning, addressing key limitations of prior studies. What was done: The authors used pre-trained VGG19 and EfficientNetB0 models to classify chest X-ray images as COVID-19 or normal in binary classification and extended it to a 4-class task (COVID-19, normal, viral pneumonia, and lung opacity). How it was done: Preprocessing included histogram equalization, CLAHE, and complement techniques to enhance image quality, while segmentation utilized lung masks for region-specific analysis. Training was performed with fine-tuned CNN models using benchmark datasets (e.g., COVID-19 Radiography Database) divided into training, validation, and testing sets. Results achieved: The VGG19 model achieved the best binary classification accuracy of 95% with CLAHE-enhanced images and 93.5% for 4-class classification using EfficientNetB0. Sensitivity, specificity, and F1 scores were also high, demonstrating reliable performance. Limitations or gaps: The study highlights that classification accuracy on segmented images was lower than on full images, suggesting the need for

further optimization. Additionally, dataset diversity and scalability remain areas for future research.

III. METHODOLOGY

Apply deep learning models to train data and achieve a high accuracy, through datacollecting and make preprocessing on it, then train data. In this section, the proposed method for chest disease classification using chest X-ray images is presented shown in Fig. 7. The data is split to 70 train, 15 validation and 15 test.

A. Dataset

As emphasized in the companion paper [11], it is our aim to optimize the performances of the models proposed there. The dataset for this study was the COVID-19 Radiography Database, a large repository of thousands of publicly available benchmark X-ray images of both lungs along with ground-truth masks. In order to make the set of images uniform for efficient processing/analysis, images are available in PNG format with 299 × 299 pixel resolution.

The database consists of 6012 lung opacity images, 10,192 normal cases, 3616 positive COVID-19 cases, and 1345 viral pneumonia cases as depicted. The database was developed by a team from Qatar University and Dhaka University in Bangladesh with collaborators from Malaysia and Pakistan of medical experts. The many classifications of the COVID-19 Radiography [13] Database are represented through samples in Fig. 1.

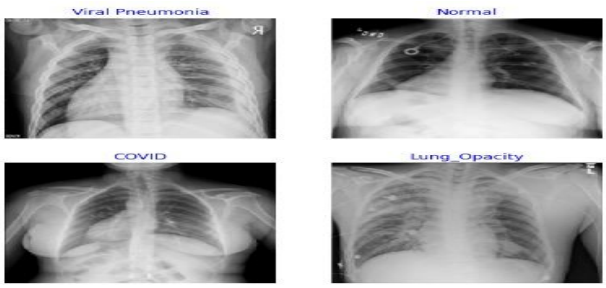


Fig. 1. Sample of the X-ray images used in dataset.

B. Data Preprocessing

Preprocessing is kind of important to increase the quality of an image and hence boost the performance of the model. The following steps are performed on the dataset:

- Apply CLAHE (Contrast Limited Adaptive Histogram Equalization)
- CLAHE improves the overall contrast of the chest X-ray images but at the same time, it enhances darker areas particularly. It enhances the contrast of the weak patterns like lesions and opacities and makes the abnormalities conspicuous as shown in Fig. 2.

Image Transformations (Data Augmentation):

Also, in order to increase the variance of the dataset and prevent overfitting, the following operations are performed: rotation, flipping, zooming, and changes in brightness. These changes allowed the model to generalize better over unseen data.

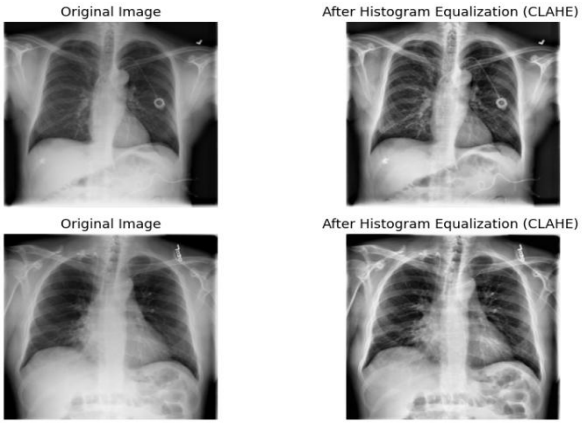


Fig. 2. Image after and before Histogram Equalization.

C. Attention Mechanism

In order to draw the model's attention to critical regions in X-ray images during learning, a binary mask-based attention mechanism was implemented as follows:

Mask attention steps:

1) Creating a binary mask: The masks were then converted to binary 0 or 1 based on a threshold value. Regions containing abnormalities were labeled as class 1, while regions containing no abnormalities were labeled as 0.

2) Weighted mask application: The binary mask was then adjusted using the following formula: This calculation is represented by Eq. (1).

$$\text{Weighted mask} = \text{mask} + 0.4 * (1\text{-mask}) \quad (1)$$

- The alpha parameter (0.4) was applied to weight the non-masked (background) regions; thus, the model gives greater importance to abnormal regions but doesn't completely ignore the rest.
- Masked image: The weighted mask was applied to the input image to emphasize pathological regions. This calculation is represented by Eq. (2)

$$\text{Masked image} = \text{image} * \text{weighted mask} \quad (2)$$

- This approach made the model focus on relevant clinical features, resulting in better performance. As shown in Fig. 3.

D. Weights for Classes in Unbalanced Data

Fig. 4 shows how the dataset was imbalanced, with the number of images associated with COVID-19 being less than those associated with normal or lung opacity classes. To overcome this, class weights were calculated and passed during model fitting: This calculation is represented by Eq. (3)

$$\text{class\_weights} = \{\text{class\_id}: \text{total\_samples} / (\text{num\_classes} * \text{samples\_in\_class})\} \quad (3)$$

By penalizing misclassifications of underrepresented classes, these class weights ensured that the model learned as equally as possible for each class.

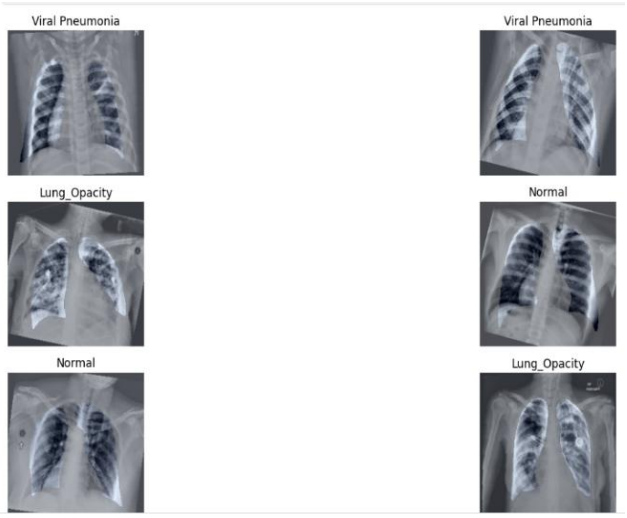


Fig. 3. Image after multiplication the mask.

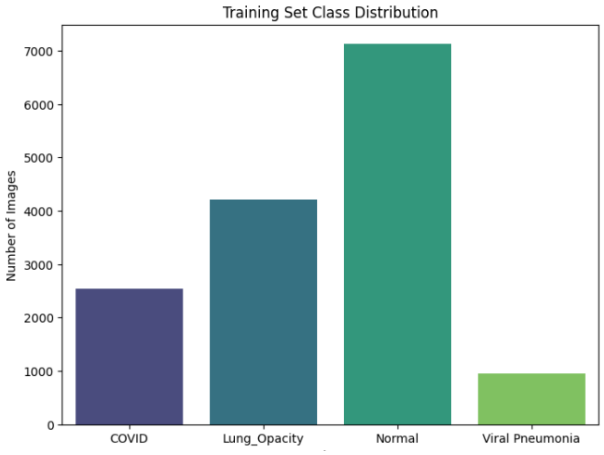


Fig. 4. Different classes in the training dataset.

In this study, CNN models were trained using several versions of segmented and original chest X-ray images as inputs. Both the raw and segmented lung X-ray pictures, as well as improved versions of these datasets, were used in the experiments. CNN models that had already been trained, especially VGG19, were used in the classification procedure. The most successful model was chosen as the final framework after the models were assessed using a variety of performance indicators. The pre-trained models used are briefly described in the sections that follow.

1) VGG19 Model: The VGGNet (Visual Geometry Group Network) is a deep neural architecture characterized by multiple processing layers. What makes VGG-19 particularly effective is its straightforward design, which incorporates stacked 3 × 3 convolutional layers that increase in depth. The network manages dimensional reduction through strategically placed max pooling layers [14]. The architecture includes two fully connected (FC) layers, each containing 4096 neurons. During training, the network extracts features using convolutional layers, while accompanying max pooling layers help reduce the dimensionality of these features. The initial convolutional layer processes input images using 64 kernels, each with a 3 × 3 filter size, to extract features. The network then utilizes fully connected layers to organize these features into a vector format.

Fig. 5 shows the architecture of VGG19. The diagram consists of a series of vertical bars representing different layers in the network: peach-colored bars indicate 3×3 convolutional layers, light blue bars represent max pooling layers, green bars show fully connected layers, and a final red bar represents the SoftMax output layer [15].

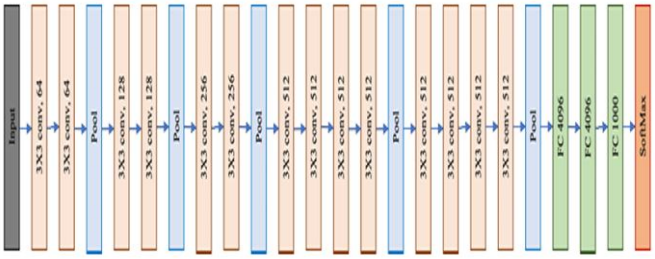


Fig. 5. Architecture of VGG19 [16].

E. Fine-Tuning VGG19 (16 Layers)

- The pre-trained VGG19 model was used (from ImageNet) with fine-tuning on top 16 layers
- The previous layers were frozen to retain general image feature extraction capabilities as shown in Fig. 6.

The last layers were thawed and fine-tuned on the chest X-ray dataset to specialize in features radiography.

Fine-tuning allowed the model to take advantage of the pre-learned weights in adaptation to domain-specific patterns like lung lesions or opacities.

2) Result of fine-tuning: Fine-tuning greatly improved the model performance, resulting in better accuracy and F1 score for all classes. Validation accuracy leveled off at 95%, indicating great generalization.

F. Image Segmentation Using Probabilistic U-Net

Goal: To segment the lung regions and then feed them into the classification model for better feature extraction.

Procedures Adopted:

1) Segmentation model: A Probabilistic U-Net was used to generate masks of the lung region from chest X-rays as shown in Fig. *7.*

2) Mask overlay: The generated segmentation masks were applied to pre-process images by isolating lung regions.

3) VGG19 integration: The segmented images were then passed through the classification pipeline, reducing the impact of irrelevant regions and generally improving accuracy.

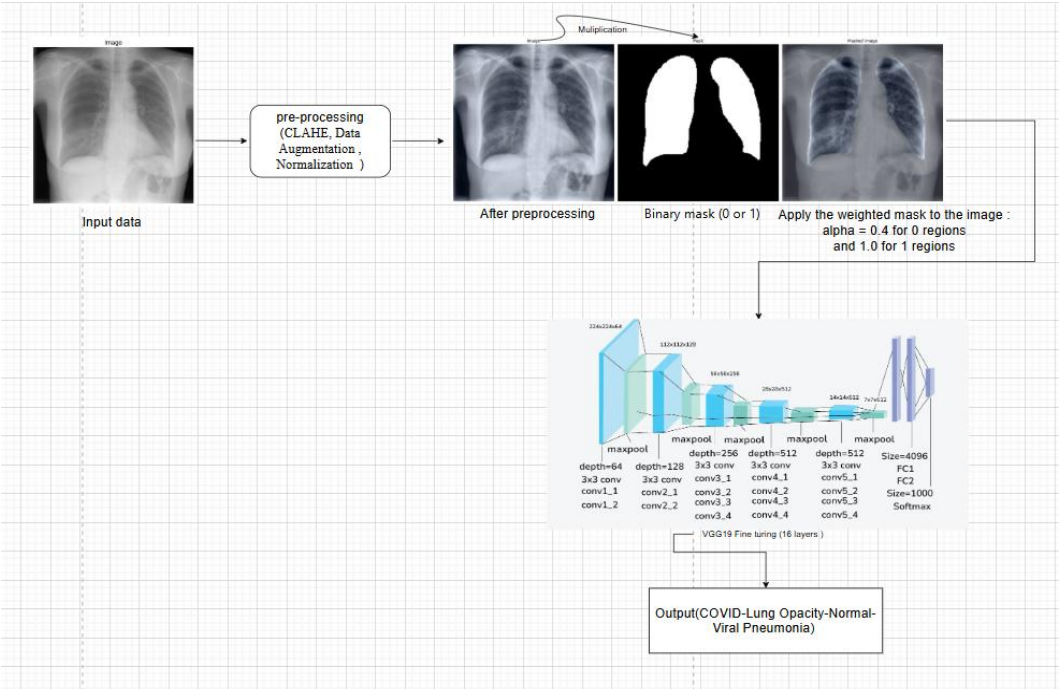


Fig. 6. The framework of the used methodology for Chest X-ray image classification.

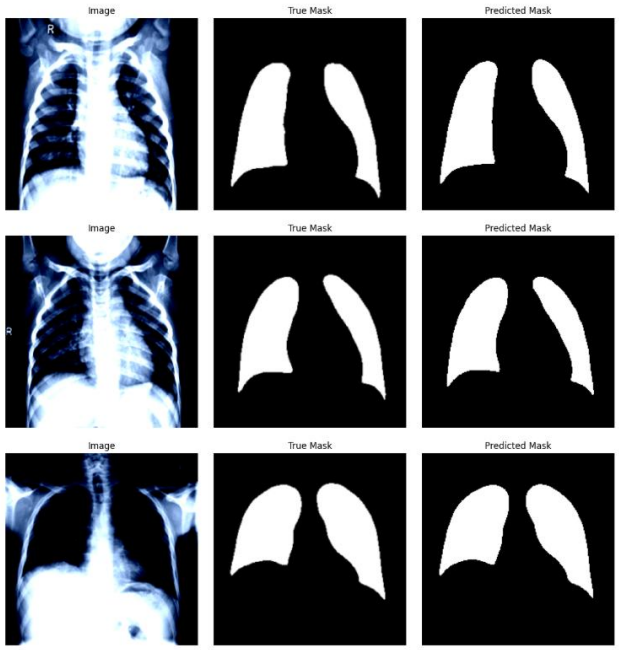


Fig. 7. U-Net was implemented to generate masks.

IV. Performance Metrics and Results

This section discusses the results of models using different evaluation measures.

A. Evaluation Criteria

Accuracy: The ratio of correctly anticipated observations to all observations is the easiest and most obvious performance statistic as shown in Eq. (4). Given in the equation below:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (4)$$

In these formulas:

TP (True Positives) represents the number of correctly predicted positive instances.

TN (True Negatives) represents the number of correctly predicted negative instances.

FP (False Positives) represents the number of negative instances that were incorrectly predicted as positive.

FN (False Negatives) represents the number of positive instances that were incorrectly predicted as negative.

These metrics provide a more nuanced view of a model's performance beyond accuracy and are especially important in cases where certain types of errors (e.g., false positives or false negatives) have different consequences or costs.

Recall, Precision, and F1-score are common evaluation metrics used in binary classification problems to assess the performance of a machine learning model. They are derived from the confusion matrix, which summarizes the model's predictions in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Here are the formulas for Recall, Precision, and F1-score:

Recall (Sensitivity or True Positive Rate):

Recall measures the ability of a model to identify all relevant instances (true positives) out of all actual positive instances. Eq. (5) shown the Recall or Sensitivity of result.

Formula:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (5)$$

Precision (Positive Predictive Value):

Precision measures the accuracy of the model's positive predictions and answers the question: "Of all the instances predicted as positive, how many were positive?" Eq. (6) shown the Prevision of data.

Formula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (6)$$

F1-score:

The F1-score is a harmonic mean of Recall and Precision. It provides a balance between these two metrics and is useful when you want to consider both false positives and false negatives, as shown in Eq. (7).

Formula:

$$\text{F1 score} = \frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall+Precision}} \quad (7)$$

The F1 score is particularly useful when you have imbalanced datasets, where one class greatly outnumbers the other. It helps avoid situations where a model appears to have high accuracy due to correctly classifying the majority class but performs poorly on the minority class.

The experimental achieved exceptional results, confirming the effectiveness of the methodology that shown in Table I:

TABLE I. EVALUATION MEASURES OF THE MODELS

Model	Precision (%)	Recall (%)	F1-score (%)	Support
COVID	0.98	0.96	0.97	543
Lung Opacity	0.93	0.93	0.93	902
Normal	0.95	0.96	0.95	1529
Viral Pneumonia	0.98	0.97	0.98	202
Accuracy			**0.95**	3176
macro avg	0.96	0.95	**0.96**	3176
Weighted avg	0.95	0.95	**0.95**	3176

A classification model's performance across four classes—COVID-19, lung opacity, normal, and viral pneumonia—is assessed using the image's confusion matrix. The actual class (True) is shown in each row, and the anticipated class is shown in each column. shown in Fig. 8.

Generalization: Consistent training and validation loss/accuracy curves demonstrate minimal overfitting as shown in Fig. 9.

The research demonstrates that combining image segmentation techniques with attention mechanisms and model fine-tuning leads to improved classification of radiographic images. This approach proves valuable in clinical healthcare settings, as it achieves high prediction accuracy while also identifying medically significant features within the images.

When using the model without the mask:

Our study focused on the performance of pre-trained convolutional neural network (CNN) architectures after they were fine-tuned for particular classification objectives. The job involved four classes. Every model configuration included a basic base model that was enhanced by extra layers that were specially designed to meet the particular requirements of the task at hand VGG19 and ResNet101 models were used in our investigation for the 4-class classification scenario. The classification tests were ResNet101 and VGG19. All models received uniform extra layers that were customized to fit each model's architecture.

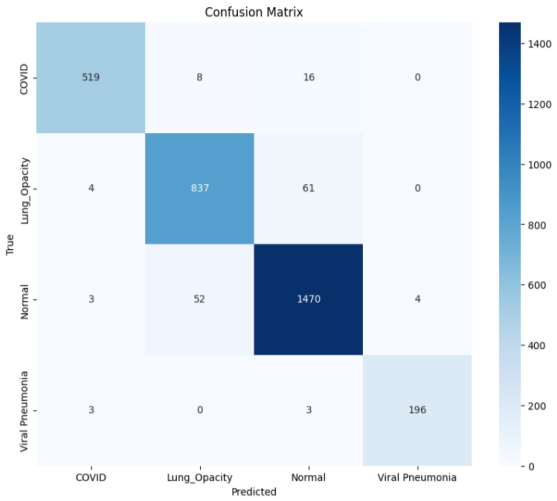


Fig. 8. Classification model's performance.

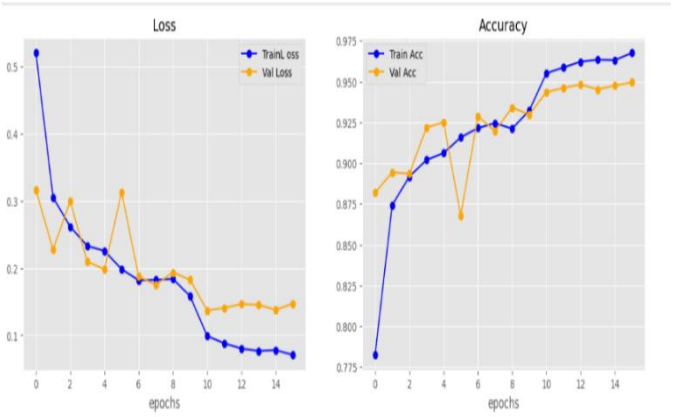


Fig. 9. Loss and accuracy for training and validation over several epochs.

TABLE II. EVALUATION MEASURES OF THE MODELS

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
VGG19	94.85	96.25	94.5	95.25
ResNet101	94.28	96	94.25	95

Table II, compares the performance metrics of two popular convolutional neural network models: VGG19 and ResNet101. VGG19 slightly outperforms ResNet101 across all metrics, with an accuracy of 94.85% versus 94.28%. However, the differences are quite minimal, with both models achieving very high-performance scores above 94% across all measures. Looking at these metrics, VGG19 shows a marginal advantage of about 0.25-0.57 percentage points across different measures, though in practical applications this difference might be negligible depending on the specific use case. Fig. 10 and Fig. 11 show the Confusion Matrices of this model.

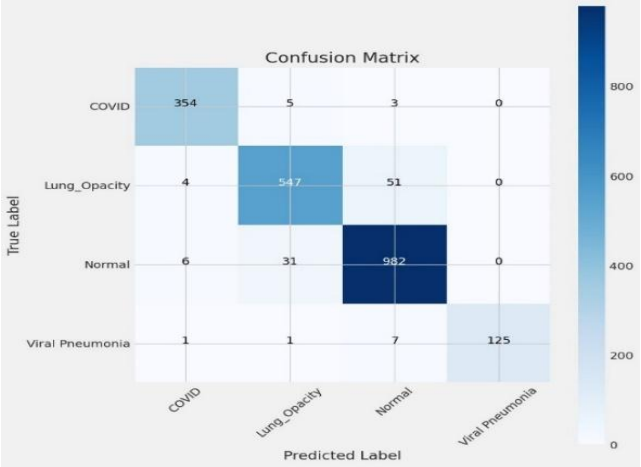


Fig. 10. VGG19 Confusion matrices.

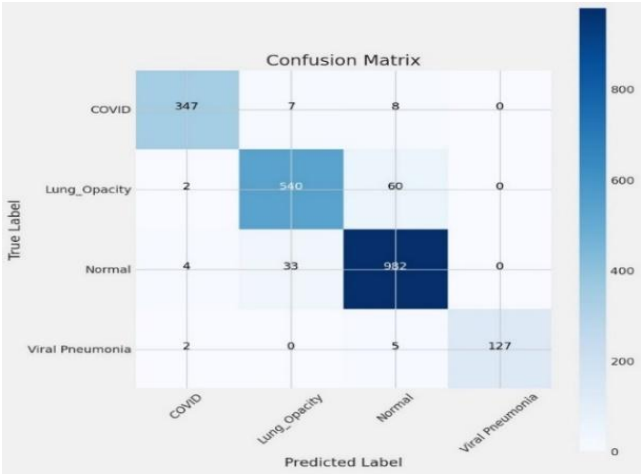


Fig. 11. ResNet101 Confusion matrices.

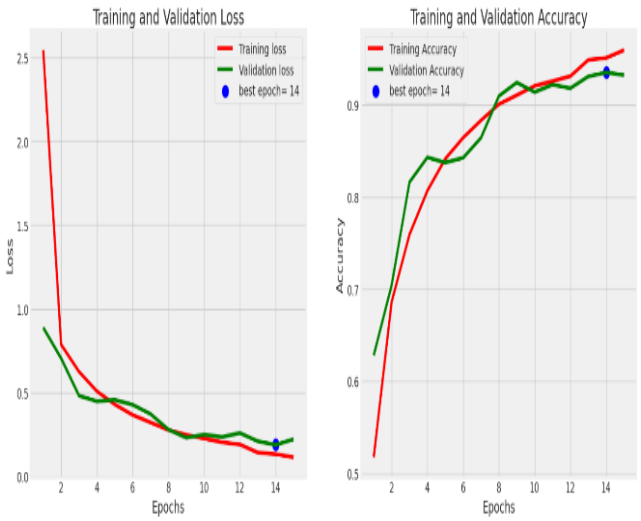


Fig. 12. Loss and accuracy for training and validation over several epochs for the resnet101model.

Fig. 12, shows two graphs side by side, illustrating the performance of a machine learning model during the training and validation phases. Let me break down each graph for you:

1) Left graph - training and validation loss: The horizontal axis displays epochs, which represent complete passes through the dataset. The vertical axis indicates the loss values, where lower values indicate better performance. Training loss (red) and validation loss (blue) both demonstrate a rapid initial decline before plateauing. Throughout the training process, validation loss maintains a slightly elevated position compared to training loss.

2) Right graph - training and validation accuracy: Similar to the loss graph, epochs are shown on the horizontal axis, while accuracy values range from 0 to 1 on the vertical axis. The training accuracy (red) and validation accuracy (green) lines show swift improvement early on before leveling off. Training accuracy consistently outperforms validation accuracy by a small margin.

3) Notable findings: The model demonstrates rapid improvement during the initial epochs for both metrics. A small, consistent gap exists between training and validation metrics, which is typical. Performance improvements become minimal after epochs 14-22, suggesting convergence. The relatively parallel trajectories of training and validation metrics indicate proper model fitting, without significant overfitting concerns.

This visualization helps in understanding how well the model is learning and generalizing to unseen data over time. It can be used to determine the optimal number of training epochs and to check for potential overfitting or underfitting issues.

Our comparative analysis against established models, which primarily used training and testing data splits for validation, demonstrated superior performance of our model, as shown in Table III.

TABLE III. COMPARISON OF THE MODEL'S OBTAINED RESULTS WERE COMPARED WITH THOSE OF OTHER PROPOSED MODELS

Study	Model architecture	Accuracy (in %)
El Houby, E. M. [12]	VGG19	93.5
Our Model	VGG19	**95.0**

Here we have proven that using the mask with the image and making Attention Mechanism the mask work is much better and more accurate than using the normal x-ray image the opposite of what this paper says [12].

V. CONCLUSION AND FUTURE WORK

This Paper proved successful in developing an accurate COVID-19 chest X-ray classification system by combining multiple sophisticated techniques. The approach utilized VGG19 fine-tuning, attention mechanisms, and image segmentation. Image quality was enhanced through CLAHE processing, while data augmentation techniques improved the model's ability to handle diverse cases. The attention component helped the model identify crucial areas in X-ray images, particularly benefiting the detection of COVID-19 and Viral Pneumonia cases. By adjusting class weights and optimizing 16 layers of VGG19 specifically for X-ray analysis, the model achieved 95% validation accuracy and a 96% macro-average

F1-score. The integration of Probabilistic U-Net for segmentation helped isolate lung regions and reduce interference from surrounding areas, leading to more accurate predictions.

The model demonstrated strong performance metrics across all categories with balanced precision and recall scores. The similar trajectories of training and validation metrics indicated good generalization without overfitting, suggesting the model's readiness for clinical application.

Looking ahead, researchers plan several enhancements: expanding the training data for better generalization, improving segmentation through advanced networks like Attention U-Net and DeepLabV3+, and incorporating additional imaging modalities such as CT and MRI scans. The team also aims to implement interpretability tools like Grad-CAM and SHAP to make the model's decisions more transparent to healthcare providers. Future developments will explore mobile and cloud deployment options, with additional validation using external datasets to ensure reliable performance in clinical settings. These improvements aim to create a dependable automated system for detecting and diagnosing respiratory conditions early.

REFERENCES

[1] Caseneuve, Guy, et al. "Chest X-Ray image preprocessing for disease classification." Procedia Computer Science 192 (2021): 658-665.

[2] Mathew, Amitha, P. Amudha, and S. Sivakumari. "Deep learning techniques: an overview." Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2020 (2021): 599-608.

[3] Nasr, Mona, Alaa El Din M. El Ghazali, and Amr I. Shehta. "Deep Learning Models for Early Detection of Blood Cancer Disease." In International Conference on Advanced Intelligent Systems and Informatics, pp. 53-65. Springer, Cham, 2024.

[4] Alqudah, Ali Mohammad, Shoroq Qazan, and Ihssan S. Masad. "Artificial intelligence framework for efficient detection and classification of pneumonia using chest radiography images." Journal of Medical and Biological Engineering 41.5 (2021): 599-609.

[5] Minaee, Shervin, et al. "Image segmentation using deep learning: A survey." IEEE transactions on pattern analysis and machine intelligence 44.7 (2021): 3523-3542.

[6] Wang, Linda, Zhong Qiu Lin, and Alexander Wong. "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images." Scientific reports 10.1 (2020): 19549.

[7] Khan, Asif Iqbal, Junaid Latief Shah, and Mohammad Mudasir Bhat. "CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images." Computer methods and programs in biomedicine 196 (2020): 105581.

[8] Caseneuve, Guy, et al. "Chest X-Ray image preprocessing for disease classification." Procedia Computer Science 192 (2021): 658-665.

[9] Avola, Danilo, et al. "Study on transfer learning capabilities for pneumonia classification in chest-x-rays images." Computer Methods and Programs in Biomedicine 221 (2022): 106833.

[10] Sharma, Anubhav, Karamjeet Singh, and Deepika Koundal. "A novel fusion based convolutional neural network approach for classification of COVID-19 from chest X-ray images." Biomedical Signal Processing and Control 77 (2022): 103778.

[11] Sahin, M. Emin. "Deep learning-based approach for detecting COVID-19 in chest X-rays." Biomedical Signal Processing and Control 78 (2022): 103977.

[12] El Houby, Enas MF. "COVID-19 detection from chest X-ray images using transfer learning." Scientific Reports 14.1 (2024): 11639.

[13] Rahman, T., COVID-19 radiography database. https://www.kaggle.com/tawsifurrahman/covid19-radiography-database (2021)

[14] Shehta, A.I., Nasr, M. & El Ghazali, A.E.D.M. Blood cancer prediction model based on deep learning technique. Sci Rep 15, 1889 (2025). https://doi.org/10.1038/s41598-024-84475-0

[15] Hamdy, Walid, Amr Ismail, Wael A. Awad, Ali H. Ibrahim, and Aboul Ella Hassanien. "An Optimized Ensemble Deep Learning Model for Predicting Plant miRNA–IncRNA Based on Artificial Gorilla Troops Algorithm." Sensors 23, no. 4 (2023): 2219.

[16] Ibrahim, Dina M., Nada M. Elshennawy, and Amany M. Sarhan. "Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases." Computers in biology and medicine 132 (2021): 104348.

Deep Learning-Based Attention Mechanism Algorithm for Blockchain Credit Default Prediction

Wangke Lin[1], Yue Liu[2]*
College of Artificial Intelligence, Zhejiang College of Security Technology, Wenzhou 35000, Zhejiang, China[1]
College of Digital Business, Wenzhou Polytechnic, Wenzhou 325000, Zhejiang, China[2]

***Abstract*—With the rise of internet finance and the increasing demand for personal credit risk management, accurate credit default prediction has become essential for financial institutions. Traditional models face limitations in handling complex and large-scale data, especially in the blockchain domain, which has emerged as a crucial technology for securing and processing financial transactions. This paper aims to improve the accuracy and generalization of blockchain-based credit default prediction models by optimizing deep learning algorithms with the Special Forces Algorithm (SFA) and attention mechanism (AM) networks. The study introduces a hybrid approach combining SFA with AM to optimize hyperparameters of the credit default prediction model. The model preprocesses blockchain credit data, extracts critical features such as user and loan information, and applies the SFA-AM algorithm to improve classification accuracy. Comparative analysis is conducted using other machine learning algorithms like XGBoost, LightGBM, and LSTM. Results: The SFA-AM model outperforms traditional models in key metrics, achieving higher precision (0.8289), recall (0.8075), F1 score (0.8180), and AUC value (0.9407). The model demonstrated better performance in identifying both default and non-default cases compared to other algorithms, with significant improvements in reducing misclassifications. The proposed SFA-AM model significantly enhances blockchain credit default prediction accuracy and generalization. While effective, the study acknowledges limitations in dataset diversity and model interpretability, suggesting future research could expand on these areas for more robust applications across different financial sectors.**

***Keywords*—Deep learning; attention mechanism; blockchain credit default prediction; special forces algorithm**

I. Introduction

The rapid development of international Internet finance, more and more third-party lending institutions, Internet finance companies continue to emerge and appear [1]. At the same time, people's living standards are getting better, consumption level rises, consumption demand increases, the number of people borrowing and consuming is also increasing, and credit institutions are paying more and more attention to personal credit risk management [2]. Credit risk management focuses on the measurement of default risk, and when a borrower applies for a loan, the lending institution needs to evaluate the risk of the borrower with the help of certain methods to decide whether to borrow or not [3]. Credit default prediction as one of the core of credit risk management, the accurate identification and assessment of default customers, not only can avoid the loss of credit default customers to the counterparty, but also can be based on the credit risk assessment model to provide customers with more accurate, more personalised, better quality products and services [4].

With the arrival of the era of big data, the use of a large amount of information data to establish an effective credit default prediction model helps financial institutions to analyse the user's consumption, capital, and creditworthiness in a certain period of time, and then predict whether there is a user's default to reduce the risk of loss [5]. Therefore, researching scientific and accurate blockchain credit default prediction methods is very significant for financial institutions to make decisions. Analysing a large number of previous studies, scholars around the world and beyond have launched a large number of research works on credit default prediction [6].

Artificial intelligence technology is updated and iterative, machine learning and deep learning algorithms are gradually applied to credit default prediction [7]. Han et al [8] applied the decision tree model to the field of credit risk assessment. Oliver [9] used the personal loan data of LC company and built KNN, SVM, Logistic and RF models, and found that the RF model performed the best. Liu et al [10] compared the performance of different classification models such as RF, ANN, and LR based on different sampling strategies according to different evaluation indexes, and the results showed that the oversampling strategy has obvious advantages in dealing with unbalanced data. Ragab and Saleh [11] constructed a credit assessment model of Lasso-LR, and the results showed that it can effectively screen out features. Kriebel and Stitz [12] proposed to build an XGBoost-RF credit assessment model, and the results show that XGBoost can improve the accuracy of RF model after filtering the features. Lin and Liu [13] used a hybrid whale bat optimization algorithm to optimize the hyper-parameters of the machine learning model, and constructed a credit default prediction model for users, and the results show that the proposed prediction model not only has higher classification accuracy, but also has higher accuracy. model not only has higher classification accuracy, but also has strong interpretability.

Based on previous research, this paper uses intelligent optimisation algorithm and deep learning algorithm to optimise and improve the personal credit default prediction model, which improves the model classification accuracy and makes the model more explanatory. In this paper, the research will be carried out through the following architectures:

- The personal credit default prediction problem is analysed, relevant features are extracted, and data preprocessing is carried out;

*Corresponding Author

- The SFA algorithm [14] is used to optimise the hyper-parameters of the Attention Mechanism Model [15], and the personal credit default prediction model based on the SFA-AM is constructed; and
- The proposed method is validated and analysed using credit default data.

Based on the research objectives, the remainder of this paper is structured as follows: Section II provides an overview of blockchain technology and the key issues in credit default prediction. Section III describes the problem-solving approach, including feature analysis, data preprocessing, model construction, hyperparameter optimization, and model evaluation. Section IV introduces the improved SFA algorithm and the attention mechanism network, presenting the SFA-AM hybrid model. Section V details the experimental design, dataset description, parameter settings, and evaluation metrics, followed by a comparative analysis of model performance. Section VI summarizes the key findings, highlights the limitations, and proposes future research directions.

II. Blockchain Credit Default Analysis

A. Blockchain Technology

1) Overview: Blockchain technology is an innovative distributed ledger technology originally proposed by Satoshi Nakamoto, the anonymous creator of Bitcoin, to create a decentralised digital currency system [16-18]. The blockchain is maintained and updated by multiple nodes in the network, with each "block" containing a batch of transaction records that are cryptographically linked to the previous block to form a tamper-proof chain. This structure ensures data transparency and security, and any attempt to modify existing information will be detected and rejected by nodes in other parts of the network, as shown in Fig. 1.

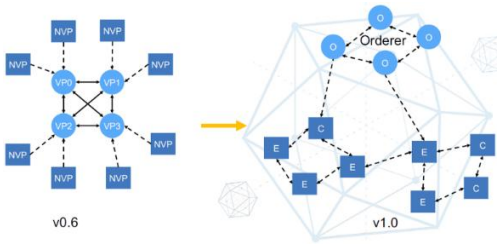


Fig. 1. Blockchain technology.

2) Blockchain characteristics: According to the principles of blockchain technology, blockchain has the following characteristics [19] (Fig. 2): 1) decentralisation; 2) immutability; 3) transparency; and 4) cryptographic security.

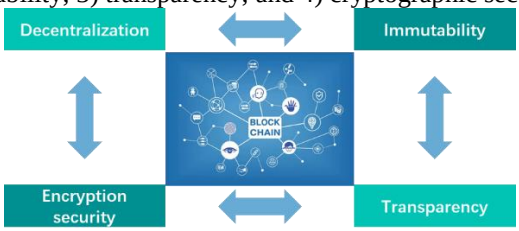


Fig. 2. Blockchain characteristics.

3) Blockchain applications: The application of blockchain technology has been extended to a number of fields, including financial services, supply chain management, healthcare, real estate, and voting systems [20], as shown in Fig. 3.

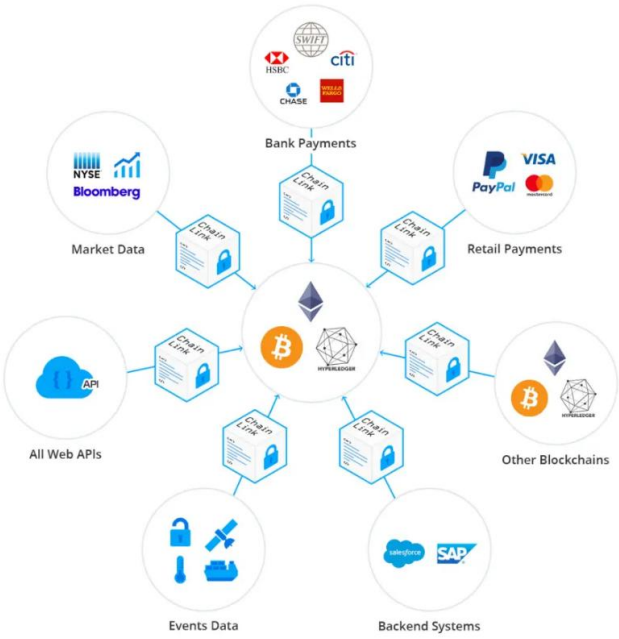


Fig. 3. Blockchain applications.

B. Credit Default Prediction Issues

1) Credit rating: Credit rating is an important indicator for financial institutions in assessing the credit risk of users. There are seven grades of loan users, which are A, B, C, D, E, F, and G, and their credit ratings decrease in order [21]. The percentage of users in each grade is given in Fig. 4. From Fig. 4, it can be seen that users of grades A, B, and C in the dataset occupy 73.12%, and the remaining grades occupy 26.88%, in which users of grades A, B, and C are higher credit rating users, which indicates that most of the users have fewer defaults.

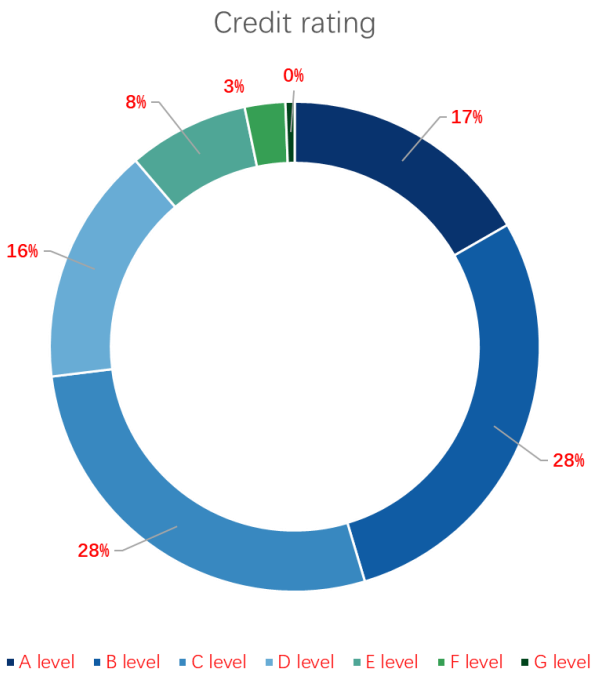


Fig. 4. Distribution of credit ratings.

2) Characterisation: The features that need to be analysed in the user credit default prediction problem mainly include basic information about the user, basic information about the borrowing project, and historical information about the borrowing project [22].

a) Basic information of loan users: In addition to the user's credit rating, the basic profile of the loan user includes the user's years of employment (Fig. 5), the user's home ownership (Fig. 6), the distribution of the user's annual income (Fig. 7), and the distribution of the loan user's loan amount (Fig. 8).

As can be seen from Fig. 5, 34.64% of the users have worked for more than 10 years, 26.05% have worked for 0 to 3 years, and the rest have worked for 3 to 10 years. The number of years of working experience reflects whether the user has the ability to make repayments, and usually the higher the number of years of working experience, the lower the possibility of their default.

Fig. 6 gives the distribution of home ownership among users. It can be seen from Fig. 6 that 50 % of the users are in home ownership and 40.13 per cent of the users are still renting their homes.

Fig. 7 shows that more than 90 % of the users have an annual income of less than $500,000, and very few users have more than $500,000 per year.

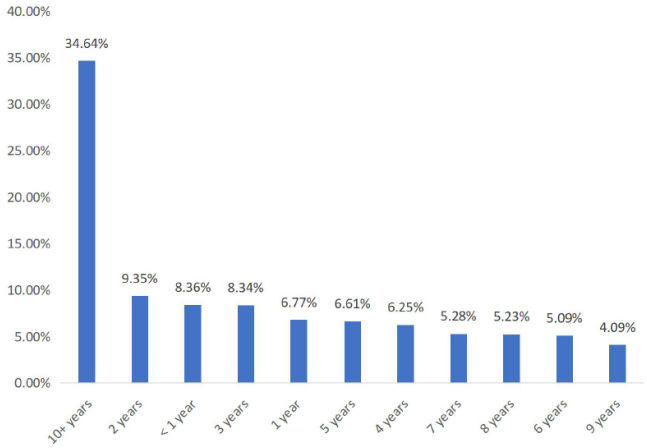


Fig. 5. Distribution of users' years of working experience.

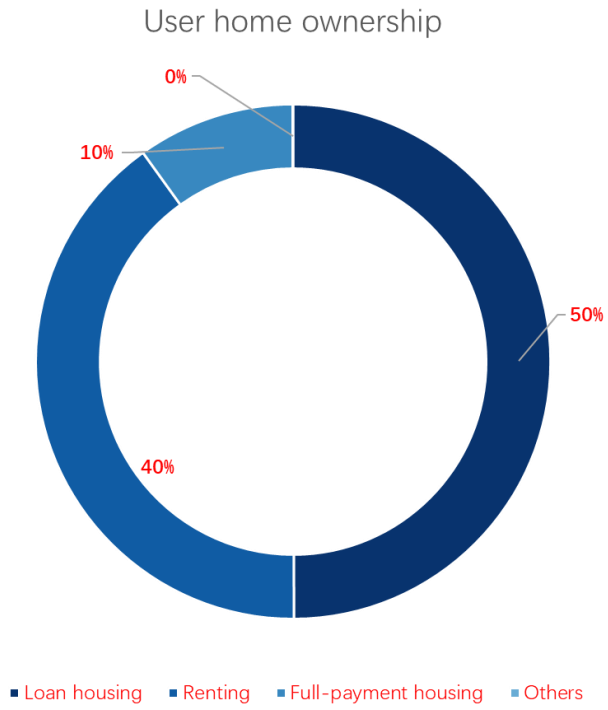


Fig. 6. Distribution of users' home ownership.

Fig. 8 gives the distribution of users' loan amounts. Most of the loans are between 5,000 and 20,000 yuan, with 10,000 yuan having the highest number of loans, and relatively few users having loans of more than 20,000 yuan.

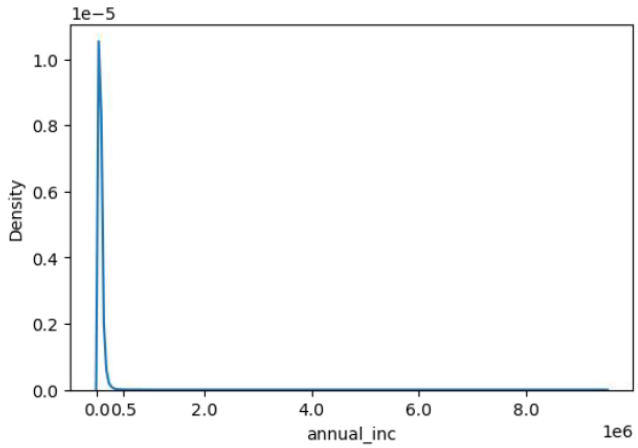


Fig. 7. Distribution of annual income of users.

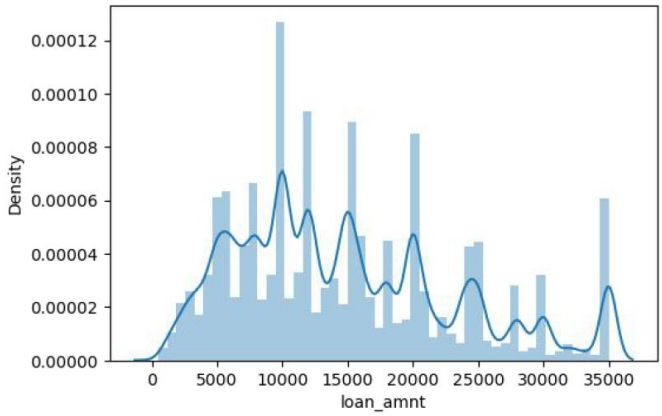


Fig. 8. Distribution of users' loan amount.

b) Basic information of the borrowing project: The basic information about the borrowing item mainly includes information such as the amount of the loan requested by the customer, the term of the loan, the interest rate of the loan, the income status verified by the bank, and the current total balance of all accounts.

c) Borrowing project history information: Borrowing item history information mainly includes information such as the number of enquiry cases in the past six months, the number of months since the last default, the number of months since the public record, the number of open lines in the credit line, and the total collection amount ever owed.

d) Description of the credit default prediction problem: The credit default prediction problem is essentially a classification and identification problem, where the inputs are user credit default characteristic variables and the outputs are user defaults, i.e., non-defaults versus defaults, as shown in Fig. 9.

Fig. 9. Description of the user default prediction problem.

C. Problem Solving Ideas and Design

In order to solve the blockchain credit default prediction problem, this paper adopts hybrid machine learning algorithm to construct blockchain credit default prediction model and design blockchain credit default prediction method based on hybrid machine learning algorithm. The design idea of this method mainly solves the blockchain credit default prediction problem from the aspects of feature analysis, data preprocessing, feature selection, credit default prediction model building, model optimisation, model evaluation, etc., as shown in Fig. 10.

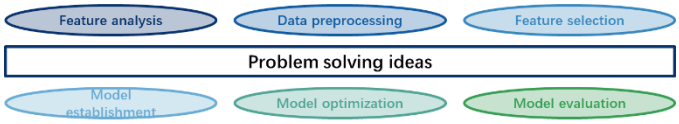


Fig. 10. Problem solving ideas.

III. Improving the SFA Algorithm to Optimise Networks of Attention Mechanisms

A. Network of Deep Attention Mechanisms

A class of deep learning models known as attention mechanism networks (AMNs) [23] enable neural networks to concentrate on pertinent components of the input data during the processing process, thereby emulating the human visual and cognitive systems. The performance and generalization of the model are enhanced by the Attention mechanism, which enables neural networks to automatically learn and selectively concentrate on the critical information in the input. The structure of the Attention mechanism network is illustrated in Fig. 11. In order to focus on the most relevant portions of each sequence element when processing it, the attention mechanism is often applied to the processing of sequential data, such as text, speech, or image sequences. This allows the model to assign various weights to different positions in the input sequence.

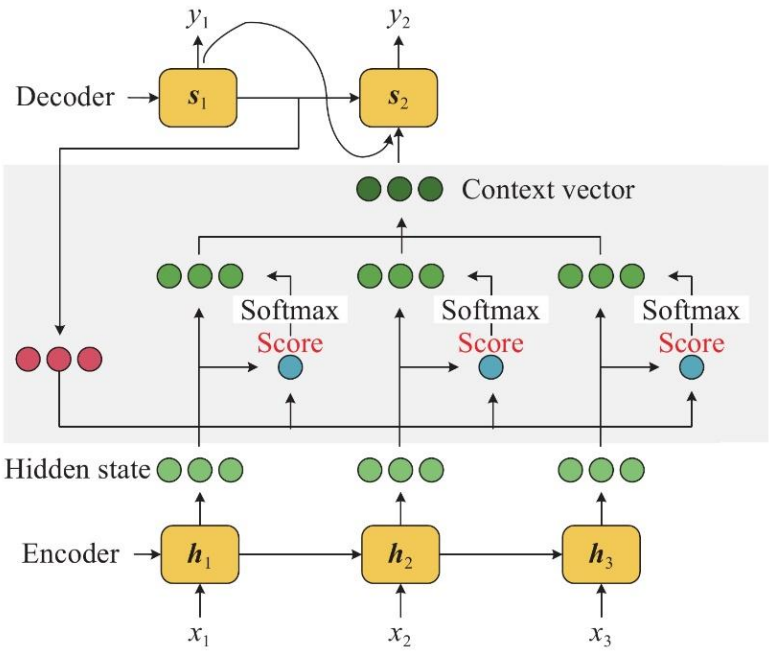


Fig. 11. Structure of the network model of the deep attention mechanism.

The core architecture of the attention mechanism consists of three main components, Query, Key and Value. Query represents the element currently being processed or the target to be attended to, Key represents the identity or characteristic of each element in the input sequence, and Value contains the specifics or information about each element in the input sequence. In attention computation, Key is used to determine how well each element matches the Query, while Value provides the actual information related to the Query.

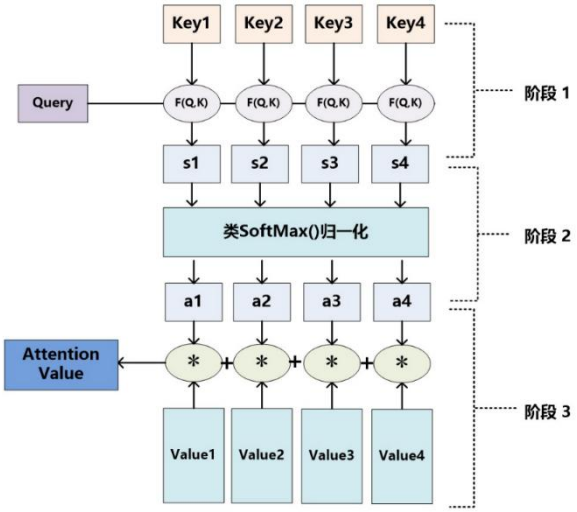


Fig. 12. Principles of attention mechanisms.

B. SFA Algorithm

Based on the strategies and behaviors of Special Forces engaged in counter-terrorism combat operations, the Special Forces Algorithm (SFA) [14] is a swarm intelligence algorithm. In order to satisfy the optimization requirements, SFA can simulate genuine dynamic behaviors by integrating a variety of strategies and incorporating unique mechanisms into the algorithm. According to the common characteristics of MAs, the process of SFA is divided into three phases: exploration phase, transition phase, and development phase (shown in Fig. 13).

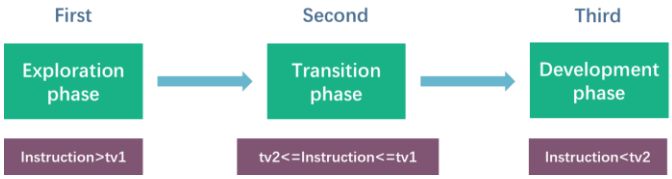


Fig. 13. Analysis of optimisation strategies for the special forces algorithm.

A "directive" is a feature of SFA that serves as an identifier to direct all team members in the completion of the mission. The directive and the threshold value, which is represented as follows, will alter in accordance with the number of iterations, allowing for the identification of the specific task type:

$$Instruction(t) = (1 - 0.15rand)\left(1 - \frac{t}{T}\right) \quad (1)$$

Where t is the current iteration number, T is the maximum iteration number, and $rand$ is a random number between 0 and 1.

2 thresholds tv_1 and tv_2 are set in SFA to clarify the phase transition as follows:

$$\begin{cases} \text{Exploration phase} & Instruction > tv_1 \\ \text{Transition phase} & tv_2 \le Instruction \le tv_1 \\ \text{Development phase} & Instruction < tv_2 \end{cases} \quad (2)$$

During the execution of mission engineering, team members can access the location information of their colleagues; however, there is a possibility that communication terminals may be lost by any team member throughout the algorithm's process, resulting in the potential loss of some team members' information:

$$p(t+1) = p_0 \cos\left(\frac{\pi t}{2T}\right) \tag{3}$$

Where, p is the lost connection probability at the current iteration t, and p_0 is the initial lost connection probability, the specific trend is shown in Fig. 14.

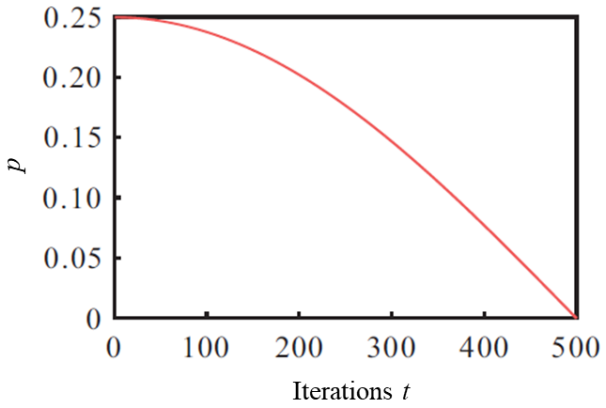


Fig. 14. Trends in probability of missing a connection.

1) Exploratory phase: Following the algorithm's initialization, the investigation phase commences. Two strategies, assault search and mass search, comprise the exploration aspect of SFA.

a) Large-scale search: Mass search missions are the primary responsibility of special forces during the exploration phase. The team members' activity area will be quite vast during mass search, and they are free to look for any possible target anywhere within the practical range at any time. Given that there are two types of jobs that team members may complete during the exploration phase, this study adds a random number to provide the team members with the ability to divide the work and conduct a random search. In other words, the location is updated based on the following equation:

$$X(t+1) = r_1\left(X_{best} - X(t)\right) \pm (1-r_1)\,range$$
$$r_1 \geq 0.5 \tag{4}$$

Where, $X(t+1)$ is the position vector of the player in the next iteration, $X(t)$ is the position vector of the current player, X_{best} is the optimal position of the previous population, r_1 is a uniformly distributed random number, and $range$ is the solution space range.

b) Raids and searches: The Special Forces occasionally conduct raids on potential locations during large-scale search missions, as they already possess some information about hostages or miscreants. The known direction of the closest and most skilled team member affects the location of every maneuver. When the random number r_1 decides to perform a surprise raid, the team members perform a position update according to the following equation:

$$X_i(t+1) = X_i(t) + w(t)\,A_i(t), r_1 < 0.5 \tag{5}$$

where $X_i(t)$ is the search and capture vector of player i for the tth iteration. For any player, the search vector is:

$$A_i(t) = \frac{f_i(t)}{f_i(t) + f_{aim}(t)}\left(X_{aim}(t) + X_i(t)\right) \tag{6}$$

Among them, $X_{aim}(t)$ is the position of player aim No. i, i.e., the optimal position known to player i, and $f_i(t)$ and $f_{aim}(t)$ are the values of their positional adaptations respectively.

The search coefficient w decreases until 0 depending on the number of iterations:

$$w(t+1) = w_0 - 0.55\arctan\left(\left(\frac{t}{T}\right)^{2\pi}\right) \tag{7}$$

Among them, this paper sets $w_0 = 0.75$, and the specific trend is shown in Fig. 15.

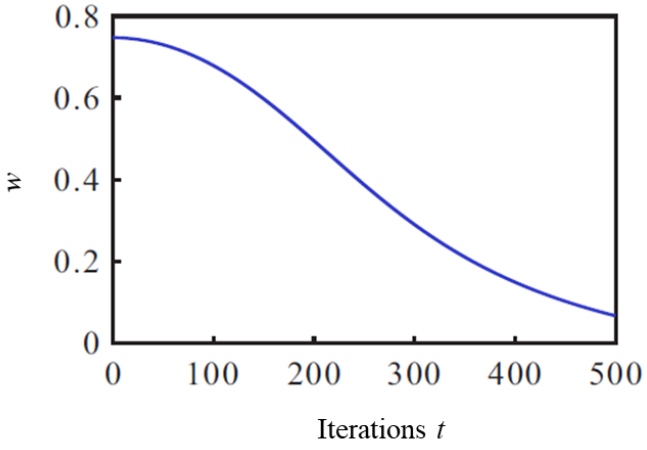


Fig. 15. Schematic diagram of search coefficient.

2) Transition phase: A buffer between the Exploration and Exploitation phases is the Transition Phase. In this period, the team will progressively transition to the exploitation phase while continuing to accomplish the previous tasks. The details are shown below:

$$X(t+1)=\begin{cases} X(t)+w(t)A(t) & r_2 \geq 0.5 \\ Instruction(t)\cdot(X_{best}-X(t))+0.1\cdot X(t) & r_2 < 0.5 \end{cases} \quad (8)$$

where r_2 is a random number satisfying a uniform distribution.

3) Development phase: A significant amount of information on the location of the criminals or hostages has been gathered by the special forces throughout the development phase, and they have now officially started the "capture" phase of the activity. The term "capture and rescue" refers to their mission of apprehending the offenders or freeing the captives.

The special operations team members in the development phase decisively approach and take a concentrated approach to surround and attack the hostage or robber based on the most likely point known to the entire team (i.e., the location of the hostage or robber). At this point, the team members use the positional updates shown below:

$$X(t+1)=X_{best}+r\cdot\left|X_{best}-X_{ave}(t)\right| \quad (9)$$

Where r is a uniformly distributed random number and X_{ave} is the current average position, calculated as follows:

$$X_{ave}(t)=\frac{1}{N}\sum_{i=1}^{N}X_i(t) \quad (10)$$

where $X_i(t)$ is the position of each player for the tth iteration and N is the total number of the whole team.

According to the SFA optimisation strategy, the SFA pseudo-code is shown in Table I.

TABLE I. SFA ALGORITHM PSEUDO-CODE

Algorithm 1: SFA algorithm pseudo-code	
1	Initialise the parameters tv1, tv2, p0, w0, and the population size N with the maximum number of iterations T;
2	Initialise population X;
3	While t<=T do
4	Calculate the fitness value;
5	Update the optimal fitness value and the optimal value;
6	Calculate the instructions;
7	If command >= tv1 do
8	If r1>=0.5 do Execute the mass search strategy;
9	Else if r1<0.5 do Execute the raid and search strategy;
10	Else if tv2< instruction<=tv1 do
11	Implementation of the transition phase;
12	Else if instruction <= tv2 do
13	Implementation development phase;
14	End if
15	Update p and w;
16	t=t+1;
17	End while
18	Return the optimal solution

The initialization, adaptation value calculation, missing information screening, and position update processes comprise the majority of the SFA computational volume. Let N represent the number of players in the SFA algorithm, T denote the maximum number of iterations, and D signify the number of issue dimensions. The computational complexity of the initialisation is $O(N)$, the computational complexity of the adaptation value update calculation is $O(N\times T)$, the computational complexity of the lost information screening is $O(N\times T)$, the computational complexity of the location updating is $O(N\times T\times D)$, and the total computational complexity is $O(N\times(1+2T+T\times D))$.

C. SFA Improved Attention Mechanism Network

In order to improve the credit default prediction accuracy of the attention mechanism network, this paper adopts the SFA algorithm to optimise the parameters of the attention mechanism network, with the RMSE value as the adaptation value and the SFA algorithm optimisation strategy as the optimisation iteration process, and the specific improvement principle is shown in Fig. 16.

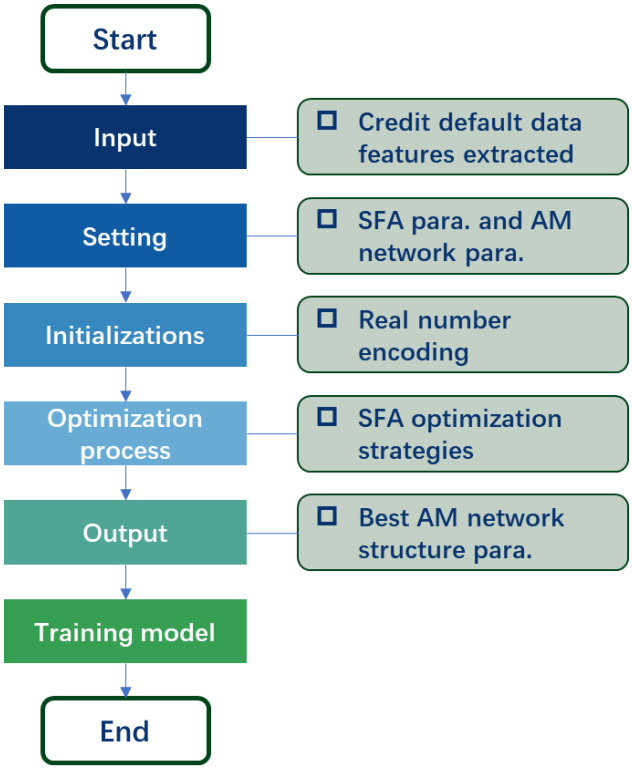


Fig. 16. Schematic diagram of improvement principle.

IV. Algorithmic Applications

In order to solve the blockchain credit default prediction problem, this paper proposes a blockchain credit default prediction method based on SFA-AM network structure. The method analyzes the blockchain credit default prediction problem, extracts the blockchain credit default feature vectors, preprocesses the data in terms of missing value processing, category variable processing, continuous variable processing, etc., constructs the blockchain credit default prediction model by using the Attention Mechanism Network, combines with the SFA Search Optimization Algorithm to optimize the blockchain credit default prediction model based on the AM network, and adopts the blockchain Bank loan dataset as the research object, the performance of the constructed blockchain credit default prediction model is verified, and the specific algorithm application idea is shown in Fig. 17.

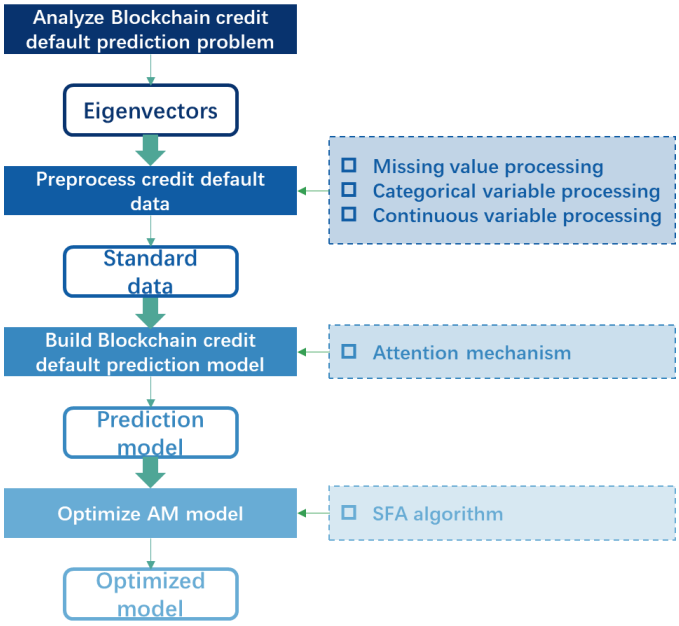


Fig. 17. SFA-AM network model application idea.

V. Results and Discussion

A. Environmental Settings

In this paper, we take Bank load data based on blockchain framework as the research object, firstly, we analyse the basic situation of the dataset, including the size of the data volume, the number of features, and the basic situation of the users; and then, we complete the work of data cleansing for the data, including irrelevant feature deletion, missing value processing, and category coding.

In order to verify the effectiveness and superiority of the blockchain credit default prediction problem based on SFA-AM network, this paper uses XGBoost, CatBoost, LightGBM, LSTM, AMnet and SFA-AM for comparative analysis, as presented in Table II.

In this paper, we use the credit dataset bank loan from kaggle website. The dataset contains information about more than 500,000 different users of Indesa bank in September 2016, out of which there are 406,601 honest customers and 125,827 defaulted customers. The percentage of honest users versus defaulted users is given in Fig. 18.

TABLE II. Parameter Settings of different Credit Default Prediction Comparison Algorithms

serial number	arithmetic	Algorithm setup
1	XGBoost	Booster= gbtree, max_depth=6, learning_rate=0.3, n_estimators=100
2	CatBoost	Iterations=1000, learning_rate=0.11, depth=6, l2_leaf_reg=3
3	LightGBM	max_depth=-1, learning_rate=0.1, n_estimators=100, min_child_weight0.003, min_child_samples=20
4	LSTM	The optimiser is Adam, the hidden layer nodes are 50 and the activation function is ReLu
5	AMnet	The optimiser is Adam, the hidden layer nodes are 100 and the activation function is tanh
6	SFA-AM	SFA population size 100, maximum number of iterations 1000, AM parameters set as in AMnet

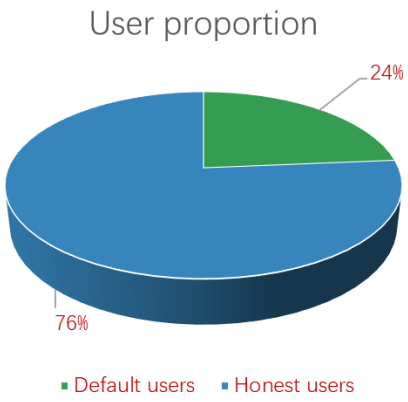


Fig. 18. Schematic representation of users.

The algorithm validation in this paper is carried out in Win11 system, the programming software is Matlab2023a, and the visualisation software includes Pycharm, PPT and Excel.

B. Comparative Analysis of Results

1) Analysis of data preprocessing results: Firstly, the missing values of the data are processed to plot the true scale of the features as shown in Fig. 19. From Fig. 19, it can be seen that, there are 16 features with missing values. In this paper, we take 50% as the limit, and the features with missing rate more than 50% are deleted. mths_since_last_recor, mths_since_last_major_derog, mths_since_last_delinq have more serious missing values, and their missing rate is already more than 50%, so these three features are deleted.

The most significant data gaps are found in "mths_since_last_record" and "mths_since_last_major_derog," which may influence model accuracy if not properly handled (e.g., through imputation or feature elimination). The features with fewer missing values, such as "tot_cur_bal" and "tot_coll_amt," are more reliable for modeling.

A 50% missing rate for "mths_since_last_delinq" means this feature could still be usable but requires careful imputation strategies. Features with more than 50% missing data might need to be dropped, depending on the importance of the feature and the model being used.

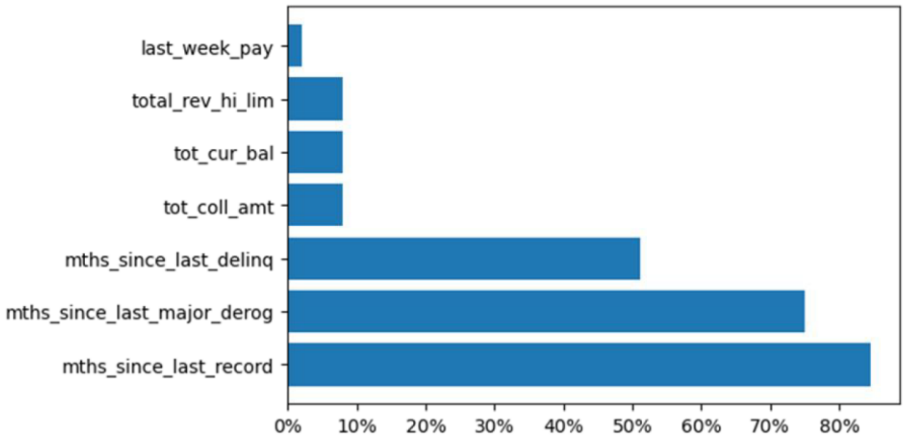


Fig. 19. Mapping the rate of absence of special diagnosis.

2) SFA optimisation of AMnet network processes: The SFA-AM optimisation iteration process is given in Fig. 20. From Fig. 20, it can be seen that the AUC value of the validation set stabilises after the number of iterations reaches 5, and reaches a maximum value of 0.951432 after the number of iterations is at 8. The AUC score starts at around 0.9495 and rapidly increases during the first few iterations, reaching approximately 0.9510 after 2 iterations. This indicates that the model performance improves significantly at the early stages of training or optimization. After about 5 iterations, the AUC reaches a peak of 0.9515, and the curve flattens, indicating that further iterations do not significantly improve the model's performance.

The model reaches a near-optimal performance (AUC of 0.9515) within a small number of iterations (around 5). Beyond that, the improvement is minimal, suggesting that the model has converged. The plateau in AUC after 5 iterations indicates that the model maintains consistent performance and does not suffer from overfitting or performance degradation, which is a good sign of stability.

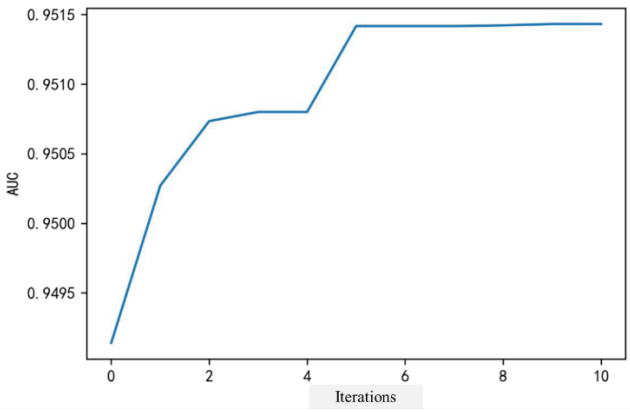


Fig. 20. Iterative process of SFA optimisation.

3) Algorithm comparison results: The model results were evaluated using the test dataset and the results of XGBoost, CatBoost, LightGBM, LSTM, AMnet and SFA-AM comparison were obtained as shown in Table III. From Table III, it can be seen that the SFA-AM model has the highest Precision, the highest Recall, the highest F1 value, and the highest AUC value, which are 0.8289, 0.8075, 0.8180, and 0.9407, respectively, which indicates that the model has a better ability to identify the defaults and non-defaults, and the model's generalisation ability is also better. The SFA-AM model demonstrates the best overall performance across all metrics, especially in Precision, Recall, and AUC, indicating that the optimization of the attention mechanism with the Special Forces Algorithm (SFA) significantly improves the predictive accuracy. CatBoost is a close second and might be a viable alternative when considering slightly lower computational complexity.

TABLE III. COMPARISON OF RESULTS

Serial number	Arithmetic	Precision	Recall	F1-score	AUC
1	XGBoost	0.7973	0.7901	0.7937	0.9324
2	CatBoost	0.8188	0.8074	0.8130	0.9400
3	LightGBM	0.7865	0.7225	0.7531	0.9156
4	LSTM	0.8166	0.7936	0.8050	0.9353
5	AMnet	0.8065	0.7309	0.7668	0.9188
6	SFA-AM	0.8289	0.8075	0.8180	0.9407

Table IV gives the blockchain credit default prediction results for AMnet and SFA-AM. As can be seen from Table IV, the ability of AM to identify defaulted customers has been improved, 27236 defaulted users can be identified before SFA is used, 27551 defaulted users can be identified after SFA is used, while identifying defaulted users as non-defaulted users has been reduced from 10460 to 10145. The SFA-AM model correctly classifies 115,423 non-default cases, which is 785 more than the AMnet model. Additionally, the number of misclassified defaults (mistakenly predicted as non-defaults) drops from 7,395 in AMnet to 6,610 in SFA-AM, showing an improvement in identifying true default cases.

SFA-AM also correctly identifies 27,551 default cases, slightly better than AMnet's 27,236. Furthermore, it reduces the number of non-defaults incorrectly predicted as defaults from 10,460 in AMnet to 10,145, reducing false positives.

The SFA-AM model consistently performs better than the AMnet model in both the identification of non-default and default cases. The reduction in misclassifications (both false negatives and false positives) indicates that the SFA-AM model has superior classification accuracy and a better ability to handle both default and non-default scenarios in credit risk prediction

TABLE IV. COMPARISON OF RESULTS BETWEEN AMNET AND SFA-AM MODELS

Real value	Projected value	Non-default	Default (on a loan or contract)
AMnet	non-default	114638	7395
	default (on a loan or contract)	10460	27236
SFA-AM	non-default	115423	6610
	default (on a loan or contract)	10145	27551

VI. CONCLUSION AND FUTURE WORK

In order to improve the accuracy of blockchain credit default prediction, this paper adopts SFA algorithm and attention mechanism model to construct blockchain credit default prediction optimisation model. It proposes a credit default prediction model based on the Attention Mechanism Network (AM) optimized by the Special Forces Algorithm (SFA). The key contributions include: (1) analyzing the credit default prediction problem, extracting relevant features, and performing data preprocessing; (2) constructing the credit default prediction model by optimizing the hyperparameters of the AM network using SFA; (3) validating the proposed model using blockchain credit default data. The experimental results demonstrate that the SFA-optimized model improves classification accuracy and generalization, achieving an AUC value of 0.9407.

We also recognise the following three areas of weakness: First, the blockchain credit default data used comes from a single bank, lacking diverse datasets from multiple industries or regions, which may limit the model’s generalizability. Then, while the paper compares several common algorithms (e.g., XGBoost, CatBoost, LSTM), it does not explore other emerging deep learning models or hybrid models in depth. Finally, although the model shows high prediction accuracy, there is insufficient analysis of the model's interpretability, particularly for the deep learning model.

Future research could incorporate more diverse datasets from different industries and regions to test the model's robustness and applicability in various scenarios. Investigating other deep learning or reinforcement learning algorithms and comparing their performance with the current model could further enhance prediction accuracy. As the complexity of models increases, focusing on the interpretability of the model—

especially the role of the attention mechanism in credit default prediction—would help financial institutions better understand and trust the model's decisions.

REFERENCES

[1] Chen, Z., Wu, Z., Ye, W., Wu, S. An Artificial Neural Network-Based Intelligent Prediction Model for Financial Credit Default Behaviors[J]. Circuits, Systems and Computers, 2023, 32(10).

[2] Andrés, A. R., JoséManuel, C. M. Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction[J]. Financial Innovation (English), 2022, 8(1):1930-1964.

[3] Amzile, K., Habachi, M.Assessment of Support Vector Machine performance for default prediction and credit rating[J].Banks and Bank Systems, 2022.

[4] Ye, X., Yu, F., Zhao, R.Credit derivatives and corporate default prediction[J].Journal of Banking & Finance, 2022, 138.

[5] Ma, H. D., Li, G., Liu, R.Y., Meng, S., Liu, X. H.The personal credit default discrimination model based on DF21[J].Journal of Intelligent & Fuzzy Systems: applications in Engineering and Technology, 2023, 44(3):3907-3925.

[6] Chang, C.Research on User Default Prediction Algorithm Based on Adjusted Homogenous and Heterogeneous Ensemble Learning[J]. 2024, 14.

[7] Gupta, P.K., Jain, K. K. Credit default prediction for micro-enterprise financing in India using ensemble models[J].Global Business and Economics Review, 2022, 26.

[8] Han, D., Guo, W., Chen, Y., Wang, B., Li, W. Personal credit default prediction fusion framework based on self-attention and cross-network algorithms[J]. Engineering Applications of Artificial Intelligence, 2024, 133.

[9] Oliver, B. Multiperiod default probability forecasting[J].Journal of Forecasting, 2022, 41(4):677-696.

[10] Liu, J., Zhang, S., Fan, H.A two-stage hybrid credit risk prediction model based on XGBoost and graph-based deep neural network[J].Expert Syst. Appl. 2022, 195:116624.

[11] Ragab, Y. M., Saleh, M. A.Non-financial variables related to governance and financial distress prediction in SMEs-evidence from Egypt[J]. .Journal of Applied Accounting Research, 2022, 23(3):604-627.

[12] Kriebel, J., Stitz, L.Credit default prediction from user-generated text in peer-to-peer lending using deep learning[J]. Operational Research, 2022(1):302.

[13] Lin, S. Y., Liu, D. R., Huang, H. P.Credit default swap prediction based on generative adversarial networks[J].Data Technol. Appl. 2022, 56:720-740.

[14] Pan, K., Zhang, W., Wang, Y. G. Special forces algorithm:a new metaheuristic algorithm[J]. Control and Decision Making,2022,37(10):2497-2504.

[15] Gretel, L. D. L. P. S, Rosso, P.Correction to: offensive keyword extraction based on the attention mechanism of BERT and the eigenvector centrality using a graph representation[J].Personal and Ubiquitous Computing, 2024, 28(2):443-444.

[16] Srinadh, V., Swaminathan, B., Vidyadhari, C.Blockchain-Integrated Advanced Persistent Threat Detection Using Optimized Deep Learning-Enabled Feature Fusion[J].Journal of Uncertain Systems, 2023, 16(03).

[17] Saurabh, K., Upadhyay, P., Rani, N.A study on blockchain-based marketplace governance platform adoption: a multi-industry perspective[J].Digital policy, regulation and governance, 2023, 25(6):653-692.

[18] Sangal, S., Nigam, A., Sharma S .Integrating blockchain capabilities in an omnichannel healthcare system: a dual theoretical perspective[J]. of Consumer Behaviour, 2023.

[19] Xiao, L., Linda, D.Bitcoin daily price prediction through understanding blockchain transaction pattern with machine learning methods[J]. of combinatorial optimisation, 2023.

[20] Kumar, V., Ali, R., Sharma, P. K.IoV-6G+: A secure blockchain-based data collection and sharing framework for Internet of vehicles in 6G-assisted environment[J].Vehicular Communications, 2024, 47.

[21] WANG Peipei, ZHOU Xiaoping, CHEN Jiajia, WANG Hanqi. A corporate credit bond default risk assessment model based on KMV-CatBoost enhancement[J]. Journal of Shanghai Normal University (Natural Science Edition in English),2024,53(02):247-253.

[22] Nguyen, S., Chen, P., Du, Y.Blockchain adoption in container shipping: an empirical study on barriers, approaches, and recommendations[J].Marine Policy, 2023.

[23] Huang, Y., Liu, Q., Peng, H., Wang, J., Yang, Q., David, O. Sentiment classification using bidirectional LSTM-SNP model and attention mechanism[J].Expert Syst. . appl. 2023, 221:119730.

Modeling Cloud Computing Adoption and its Impact on the Performance of IT Personnel in the Public Sector

Noorbaiti Mahusin[1], Hasimi Sallehudin[2], Nurhizam Safie Mohd Satar[3],
Azana Hafizah Mohd Aman[4], Farashazillah Yahya[5]
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Malaysia[1, 2, 3, 4]
Faculty of Computing and Informatics, Universiti Malaysia Sabah, Malaysia[5]

***Abstract*—This study investigates the factors influencing cloud computing adoption in the public sector, emphasizing the performance of IT personnel. Through qualitative interviews with five IT management professionals in the public sector, we identify key challenges in integrating cloud computing systems. The primary issues include technical complexity, skill and knowledge deficits in data governance and budget constraints. These insights inform the development of the Cloud Computing Capacity and Integration Model for the Public Sector, which proposes a comprehensive strategy to address these barriers. Our findings identified five key challenges to cloud computing adoption in the public sector. First, compatibility issues and system integration challenges resulting from conflicts between cloud platforms and older infrastructure contributed to operational inefficiency. Second, data migration issues due to incompatible formats and structures resulted in data loss and delays. Third, network constraints, such as limited bandwidth and high latency, hampered cloud service performance. Finally, a lack of staff training and budget constraints hampered successful cloud integration, emphasizing the importance of focused capacity-building initiatives and additional financial support. Thus, the "Cloud Computing Integration and Cloud Computing Acceptance and Performance Model" (CCAPM) presented in this research paper aims to deliver a comprehensive model that tackles a wide array of technical, operational, and human resource challenges to create an effective cloud computing ecosystem, enhance the adoption of cloud computing within the public sector, elevate the capabilities of public sector IT personnel, and develop a secure, resilient, and sustainable cloud computing environment in the public sector.**

***Keywords*—Cloud computing; cloud integration model introduction; performance of IT personnel; public sector; system integration**

I. INTRODUCTION

Cloud computing has emerged as one of the most influential and potential technologies in the modern IT world. The study in [1] highlight cloud computing's emergence as a powerful force in contemporary IT, offering on-demand services for computation, storage, and applications. This technology offers a variety of advantages including the ability to reduce costs, increase operational flexibility, and increase the scale and efficiency of IT services. The study in [2] point out that cloud computing delivers significant cost advantages. Organizations can leverage virtualization and cloud infrastructure to pay only for what they use, eliminating the upfront costs associated with traditional IT investments. This translates to increased operational flexibility and improved scalability and efficiency of IT services. In the context of the public sector, the application of cloud computing is seen as a strategic step capable of bringing digital transformation to government services. The research in [3] argues that cloud computing adoption is a strategic move for the public sector, enabling them to deliver government services with greater efficiency and agility. However, there are various challenges and obstacles that public institutions have to face in the process of adoption and integration of cloud computing. According to study [4], public institutions face several hurdles when adopting and integrating cloud computing solutions. These challenges include security risks, network connectivity issues, cost and budget constraints, policy formulation, skills availability, and infrastructure readiness. This study will focus on two main problems: the complexity of integrating cloud computing systems with existing infrastructure and the lack of knowledge and skills in data governance.

The complexity in the integration of cloud computing systems with existing infrastructure in public institutions is one of the main challenges in the application of this technology. The study in [3] highlights the complex integration of cloud computing systems with existing public institution infrastructure as a significant challenge. The existing IT infrastructure in most public institutions may not be designed to support cloud technology, making the integration process complex and requiring significant changes. Integrating cloud technology into existing IT infrastructure in public institutions can indeed be complex due to various challenges highlighted in the research papers. The study in [5] identify security concerns, data leakage, and legal implications as key hurdles to cloud computing adoption in government organizations. This includes changes in IT architecture, software customization, and data restructuring. The study in [6] advocates for software restructuring, which entails modifications to IT architecture, software customization, and data restructuring to enhance software system performance and efficiency. The cost of integrating cloud computing can also be very high, especially if it requires the purchase of new equipment or major upgrades to existing infrastructure. Additionally, [7] said that supporting a large number of users in the cloud can necessitate increased network bandwidth, which

translates to higher expenses for both the service provider and potentially the customer.

In addition, the lack of knowledge and skills in data governance is also a big obstacle. Cloud computing requires a high level of understanding of data management, security, and compliance with relevant regulations. As the study [8] emphasizes, cloud computing necessitates a deep understanding of data management, security, and regulatory compliance. This ensures the integrity and protection of sensitive information entrusted to the cloud, mitigating the risk of breaches and unauthorized access. According to study [9], inadequate data and storage security can trigger a domino effect of negative consequences. These include data security breaches, non-compliance with regulations, and difficulties managing technological advancements. In the public sector context, a lack of training and professional development for IT personnel can result in slow and ineffective adoption of cloud computing technology. Building on the work of [10], a lack of awareness and training among IT staff emerges as a critical barrier to public sector information technology adoption. This highlights the importance of well-designed training programs to address these challenges and facilitate successful IT implementation.

This study attempts to achieve two main objectives regarding cloud computing adoption in public sector. It aims to establish the details behind the complexity of cloud computing systems integration with existing infrastructure, with special emphasis on technical, financial and organizational challenges and their correlation with the decision to accept or reject cloud technology. Secondly, this research intends to evaluate the impact of knowledge and skill gaps in data governance on cloud adoption of IT personnel. It will evaluate their knowledge and understanding of data governance principles and best practices, as well as the effectiveness of existing training and development efforts in strengthening their skills and competencies.

This study has great importance in the context of technology development and digital transformation in the public sector. The study in [11] research underscores the critical role of technological advancements and digital transformation in the public sector. These studies shed light on the multifaceted nature of digital transformation, emphasizing the need for a holistic approach that extends beyond mere technological innovation. According to study [12], several factors significantly contribute to IT project complexity. These factors include the involvement of diverse stakeholders, the implementation of new technologies, the presence of conflicting goals, and the ambiguity of project objectives. In addition, by analyzing the lack of knowledge and skills in data governance, this study can help in designing more effective training and professional development programs to increase the level of readiness of IT personnel to accept cloud computing technology. The study in [13] posit that analyzing IT personnel's data governance knowledge and skill gaps can be instrumental in designing targeted training programs. These programs would enhance their preparedness for adopting cloud computing expertise.

Interests and Contributions

This article has great importance in various aspects, especially to the government, community, industry, and the development of science. Specifically, this study seeks to contribute significantly in the following ways:

First of all, this study provides an important contribution to the government in the context of the digital transformation of the public sector. However, the study in [14] integrating these systems presents complexities that require a thorough understanding to develop effective policies and strategies. This will not only increase operational efficiency but also reduce costs and improve the effectiveness of services to the public. The study in [15] explored how cloud computing integration enhances operational efficiency, reduces costs, and improves public service delivery which requires consulting key references for a comprehensive understanding.

Second, in terms of the community, this study will have a positive impact by strengthening the ability of public institutions to provide better services to the community. Wider use of cloud computing technology can open up new opportunities for innovation and development of applications that benefit local communities. According to study [16], cloud computing is an IT service model that delivers computing resources, including hardware and software, on-demand. This means they are readily available to users whenever needed, regardless of their device or location.

Third, from an industry perspective, this article provides valuable insight to decision-makers and IT managers in providing more appropriate infrastructure and facilitating the integration of cloud computing technology in the private sector. The study in [17] discuss the significant cost benefits of cloud computing, highlighting its ability to minimize IT infrastructure costs, ease deployment of new products, and reduce maintenance costs.

Finally, in terms of knowledge, this article will contribute to the existing literature by providing a deeper understanding of the challenges of cloud computing integration and deficiencies in data governance in public institutions. Cordella and Paletti's exploration of public value delivery complexities in public agencies sheds light on the challenges of effective governance and value creation. Understanding these intricacies is key to comprehending the obstacles public institutions face when integrating cloud computing and addressing data governance deficiencies [18]. This will provide a foundation for further research and strengthen the knowledge base in the field of information technology and management in Malaysia.

Overall, this article not only provides an in-depth analysis of the issues faced by public institutions in accepting cloud computing technology but also provides valuable guidance for policy preparation, community development, industry improvement, and the development of science in our country [19]. This study has a limited scope which allows a clear focus on specific and relevant issues to IT personnel without extending the study to aspects that may differ in the context of other grades in the public sector. [20], in their guidance from the JBI Scoping Review Methodology Group, advocate for engaging knowledge users in scoping reviews. This emphasis stems from the value stakeholders bring to the process, ensuring the research findings are relevant and applicable. Besides, [21] emphasizes the challenges associated with technology integration, particularly the complexities arising from merging diverse data types and

sources, as evidenced by a survey on data integration problems. Thus, this study aims to identify and analyze these factors in the hope of providing useful guidance to improve the adoption process of cloud computing in the public sector. Through this study, public institutions can be more prepared and competitive in the era of digital transformation.

II. Literature Review

A. Current State of Cloud Computing Adoption

Cloud computing revolutionizes IT by offering cost-effective, scalable solutions that streamline government operations and improve service delivery. Studies by [22] support this, highlighting cost reduction and enhanced services as key benefits [23]. However, challenges like integration complexity and a skilled workforce gap need to be addressed for successful implementation.

B. Recent Studies and Theories

Cloud adoption in the public sector is a double-edged sword, according to recent research by [24] and [25]. While it offers potential benefits, there are significant challenges to overcome. Data security, regulatory compliance, and integrating cloud computing with existing infrastructure are all major hurdles [21][3]. Notably, the complexity of this integration remains a critical barrier to wider adoption.

C. Contrasting Views

Cloud computing's impact on the public sector sparks debate. Research by [26] highlights its potential for increased flexibility and cost savings [2]. However, [27] point to initial investment costs and integration hurdles [7]. This suggests the actual value of cloud computing depends on each organization's preparedness and existing infrastructure.

D. Identified Gaps

A notable gap in the literature is the specific challenges faced by IT personnel in the Malaysian public sector. While there is a wealth of general information on cloud computing adoption, few studies focus on the detailed impacts on IT personnel's performance and the specific barriers they encounter. This study aims to fill this gap by providing an in-depth analysis of these factors, focusing on the Malaysian public sector context [21].

E. Literature Gap Addressed by the Study

This study tackles a critical gap by zeroing in on the twin challenges of complex system integration and the public sector's data governance skills shortage. According to study [28], the implementation of data integration in the public sector faces four main challenges: the lack of management support, policy standards and politics, the inability of human resources and the lack of governance. By delving into these areas, the research sheds light on how these factors significantly impact cloud adoption. The factors influencing cloud adoption are relative advantage, service quality, risk management, top management support, facilitating conditions, influence on cloud providers, server location and computer self-efficacy [29]. Furthermore, it offers practical solutions to overcome these barriers, paving the way for smoother cloud implementation.

F. Introduction of Theoretical Framework

To strengthen the study's foundation, the well-established Technology Acceptance Model (TAM) is incorporated into the research framework. Developed by [30], TAM sheds light on how users within organizations perceive and adopt new technologies. This model emphasizes two key factors: perceived usefulness and perceived ease of use. By integrating TAM, the study aims to provide a more comprehensive understanding of cloud adoption in the public sector.

G. Relevance of TAM

The TAM proves particularly valuable to this study. Perceived availability acts as a significant predictor of user perception and system/service quality, significantly influencing the adoption of sustainable cloud computing solutions [30]. TAM's core focus aligns perfectly: it delves into how IT personnel in the public sector perceive cloud computing, specifically the factors influencing their acceptance and use of this technology. By leveraging TAM, the research gains a structured framework to analyze the key drivers of cloud adoption in this context. TAM provides empirical evidence supporting the positive inclination of end users toward cloud computing adoption [31].

H. Application of TAM in the Study

This study leverages TAM to explore key hypotheses. It predicts that the perceived ease of use will be negatively impacted by the complexity of integrating cloud computing with existing infrastructure. According to study [32], cloud computing adoption and utilization determinants encompass perceived ease of use, compatibility, security, technological readiness, organizational size, competitive pressures, trialability, cost-efficiency, and individual innovativeness. Conversely, the perceived usefulness is expected to rise with the improvement in IT personnel's performance driven by cloud adoption. Perceived availability emerges as a critical determinant of users' perceived utility and system/service quality, thereby influencing the adoption of sustainable cloud computing solutions [30]. By examining these relationships through TAM's structured framework, the study offers a systematic analysis of factors influencing cloud adoption in the public sector.

I. Strengthening Academic Rigor

This study's integration of the TAM bolsters its academic credibility in several ways. TAM serves as a robust theoretical foundation, allowing for the development of well-defined hypotheses. TAM constitutes a theoretical framework for evaluating technology adoption within organizational contexts, a critical component of successful digital transformation initiatives [33]. Furthermore, it acts as a roadmap for data collection and analysis, guaranteeing that the research findings are firmly anchored in established theoretical concepts.

J. Methodology Supported by Theoretical Framework

To understand IT personnel's perspectives on cloud computing adoption in the Malaysian public sector, this study utilizes qualitative interviews guided by the TAM. These in-depth discussions will explore perceived challenges, benefits, and the impact on IT staff performance—all key areas influencing TAM's core constructs. By analyzing the collected

data, the research aims to test TAM-derived hypotheses and shed light on the model's practical application within this specific context. This will offer valuable insights into the factors driving or hindering cloud adoption in the Malaysian public sector.

III. Research Methodology

A. Qualitative Research Methods

This study delves into the factors influencing cloud adoption and its impact on public sector IT personnel performance through qualitative research methods. These methods are ideal for exploring complex issues in depth, as they allow researchers to capture the nuanced perspectives and lived experiences of individuals. This approach provides a richer understanding of the human element within cloud adoption in the public sector.

B. Nature of Qualitative Research

Qualitative research serves as a launchpad for deeper understanding. According to study [34], qualitative research is characterized by its naturalistic, contextualized, and interpretive nature, prioritizing the exploration of processes and developmental patterns over the delineation of definitive products or outcomes. Unlike quantitative methods that focus on numbers and broad generalizations, qualitative research delves into the 'why' behind human experiences. Qualitative research employs methodologies such as interviews, focus groups, and observation to delve into the underlying motivations and experiences that shape human behavior [35]. Through interviews, observations, and document analysis, it gathers rich, non-numerical data like words and meanings. This approach helps us uncover motivations, opinions, and underlying reasons, laying the groundwork for future studies and potentially informing quantitative research with well-defined hypotheses.

C. Suitability for the Study

1) Unpacking the intricacies of cloud adoption within the public sector demands a research approach that goes beyond the surface. Qualitative methods rise to this challenge. Qualitative research employs methodologies such as interviews, focus groups, and observation to investigate participants' perspectives and lived experiences, typically involving smaller sample sizes and iterative data collection until theoretical saturation is achieved *[35]*. By enabling in-depth exploration of IT personnel's perceptions and the factors influencing adoption and performance, qualitative research offers a nuanced understanding often missed by quantitative approaches. This rich data provides valuable insights that can guide successful cloud implementation in the public sector *[36]*.

2) Qualitative research methods shine a light on the human dimension of cloud adoption in the public sector. According to Omar Ali et al., the adoption of cloud-based services within local government entities is significantly influenced by compatibility, complexity, cost, security considerations, anticipated benefits, and organizational scale [37]. By gathering detailed narratives and insights from IT personnel, researchers can uncover the rich tapestry of their experiences. This in-depth approach goes beyond just identifying barriers and facilitators; it reveals the nuanced contextual factors that influence how cloud computing is adopted and implemented. Compatibility, relative advantage, security, trust, and reduced complexity are key determinants of positive attitudes toward cloud adoption [38]. This knowledge is crucial for crafting successful strategies for the public sector.

D. Data Collection Methods

1) To gather rich data on cloud adoption, this study employed semi-structured interviews. This method strikes a balance between focus and flexibility. A pre-defined interview guide ensures we cover key areas while allowing the conversation to flow organically and explore unexpected insights that emerge. This approach yields detailed and comprehensive data, providing a deeper understanding of IT personnel's experiences in the public sector.

2) This study went beyond interviews, to strengthen the credibility and comprehensiveness of the findings. Document analysis and observations were incorporated, employing a technique called triangulation. This multifaceted approach allowed for the verification and enrichment of the data. By examining relevant documents and potentially observing IT personnel interacting with cloud computing, the research paints a more complete picture of cloud adoption in the public sector.

E. Rationale for Choosing Qualitative Interviews

1) Depth of understanding: In-depth interviews are the key to unlocking the complexities of cloud adoption in the public sector. Research conducted in diverse geographic contexts, such as South Africa, underscores the efficacy of in-depth interviews in elucidating the facilitators, obstacles, and intricacies inherent in cloud computing adoption processes [4]. Unlike surveys, interviews allow for a nuanced exploration of IT personnel's experiences, perceptions, and attitudes. Such interviews afford IT professionals a venue to articulate their perspectives on factors impacting workplace well-being, including interpersonal relationships and individual work environment characteristics [39]. This deep dive reveals the 'why' behind their challenges and the specific benefits they anticipate. This rich understanding is crucial for crafting effective cloud implementation strategies tailored to the public sector.

2) Flexibility: The semi-structured nature of the interviews empowers researchers to delve deeper. Qualitative research offers a rich and nuanced exploration of participants' subjective realities, allowing for the emergence of unforeseen patterns and insights [40]. When IT personnel offer intriguing responses or areas of ambiguity arise, the flexible format allows for follow-up questions. This adaptability ensures researchers capture a comprehensive understanding of the challenges and nuances surrounding cloud adoption in the public sector.

3) Rich data: Interviews are the cornerstone of this study, unearthing the intricate tapestry of cloud adoption in the public sector. Research conducted by [3], [4] and [41] underscore the pivotal role of interviews in elucidating the determinants, advantages, and obstacles associated with cloud computing adoption within governmental organizations. Unlike

quantitative methods, interviews capture rich, qualitative data that reveal the interplay of technical hurdles, organizational dynamics, and human factors influencing the process. This nuanced understanding is essential for navigating the complexities of cloud adoption within the public sector.

F. Ethics Statement

This ethical statement explains the steps taken to ensure that this study is conducted in compliance with high ethical standards. This includes how participant consent is obtained and the steps taken to ensure data privacy.

G. Obtaining Participant Consent

Before the study was conducted, all participants were given a detailed explanation of the purpose of the study, the procedures to be carried out, and their rights as participants. Each participant was asked to sign a written consent form before participating in this study.

H. Ensuring Data Privacy

Regarding [42], data security and privacy issues in cloud storage systems are significant concerns, including unauthorized access, data leakage, and privacy disclosure. To ensure the privacy of participant data, the following steps are taken:

1) Anonymity: All information collected is anonymized [43], [44].

2) Data storage: Study data is stored in a password-protected electronic format [45].

3) Confidentiality: Participants' personal information is kept confidential [46].

4) Data destruction: After the study is completed, all personal data will be securely destroyed [43].

IV. Result

This section adheres to a rigorous approach by presenting the interview findings objectively. Each theme identified within the data is reported without interpretation, allowing the participants' voices to shine through. Supporting quotes are carefully chosen to illustrate these themes, providing a clear picture of the key issues surrounding cloud adoption in the public sector.

A. Compatibility Issues and System Integration

Findings: Interview data revealed a significant hurdle: technical complexity and compatibility issues during cloud integration with existing infrastructure. Participants described difficulties aligning new cloud computing with legacy software and hardware, leading to operational disruptions and inefficiencies. One participant (Participant 1) echoed this sentiment, stating, 'The old system used was difficult to integrate because of numerous compatibility problems.'

Supporting Quotes:

- Participant 1: "This old system is kind of hard to integrate; it seems like there are a lot of bugs. It's really hard."
- Participant 3: "Integrating the cloud computing with our existing infrastructure has been a major headache due to compatibility issues.

B. Data Migration Problems

Findings: Data migration emerged as a critical barrier due to incompatible data formats and structures. Participants consistently reported challenges transferring data from legacy systems to the cloud, often resulting in data loss, corruption, or significant delays. As Participant 2 succinctly stated, 'Data migration is hampered by the incompatibility between old data formats and structures and new cloud computing systems.'

Supporting Quotes:

- Participant 2: "This data is not directly compatible with the cloud; it is difficult to transfer."
- Participant 4: "We faced numerous problems migrating our legacy data to the cloud due to different data formats."

C. Network Problems

Findings: Network limitations emerged as a major roadblock to cloud adoption. Participants consistently described the challenges of slow infrastructure and high latency. They reported that, insufficient bandwidth and unreliable internet connections hampered the smooth operation of cloud services, leading to frequent downtime and decreased productivity. As Participant 3 noted, 'Slow network infrastructure and latency problems impede the smooth integration of cloud computing.'

Supporting Quotes:

- Participant 3: "Insufficient network speed and delays in data transmission make the user experience less satisfactory."
- Participant 5: "Our internet connection is often too slow to handle the demands of cloud computing effectively."

D. Lack of Training and Knowledge

Findings: A critical knowledge gap emerged as a significant barrier to cloud adoption. Participants consistently highlighted the lack of training and knowledge among IT staff. They pointed out that many personnel are not equipped to effectively manage and implement cloud technologies, leading to improper usage and integration failures. This sentiment was echoed by Participants 4 and 1, who stressed the need for 'in-depth training in cloud technology management and integration' for IT personnel.

Supporting Quotes:

- Participant 4: "Without adequate training, staff do not have the necessary skills to carry out the integration process properly."
- Participant 1: "We need comprehensive training programs to equip our team with the necessary cloud computing skills."

E. Budget Constraints

Findings: Budgetary limitations emerged as a persistent hurdle. Participants consistently highlighted the challenges posed by limited financial resources. Upgrading hardware and software to meet cloud integration requirements proved difficult due to a lack of funds for these vital investments. This financial

constraint significantly hampers cloud adoption efforts, as Participant 5 poignantly noted: 'The lack of funds caused difficulties in updating the hardware and software required for cloud computing integration.'

Supporting Quotes:

- Participant 5: "The budget is not enough; the hardware upgrade is not affordable, and the software integration is crazy expensive."
- Participant 2: "We struggle to secure the funding needed to support our cloud computing initiatives."

These results provide insight that can be used to plan more effective strategies for integrating cloud computing and improving the performance of IT personnel in the public sector.

TABLE I. SUMMARY OF KEY FINDINGS

Theme	Issues Identified	Supporting Quotes
Compatibility Issues	Technical complexity, system compatibility problems	"The old system used was difficult to integrate because there were many compatibility problems."
Data Migration Problems	Incompatible data formats, transfer difficulties	"Data migration is hindered by the incompatibility of old data formats and structures with new cloud computing systems."
Network Problems	Slow network infrastructure, high latency	"Slow network infrastructure and latency problems impede the smooth integration of cloud computing."
Lack of Training and Knowledge	Insufficient training, skill deficits	"IT personnel need more in-depth training in cloud technology management and integration."
Budget Constraints	Insufficient funds for upgrades	"The budget is not enough; the hardware upgrade is not affordable, and the software integration is crazy expensive."

Table I summarizes key findings into five themes which are compatibility issues, data migration problems, network problems, lack of training and knowledge, and budget constraints.

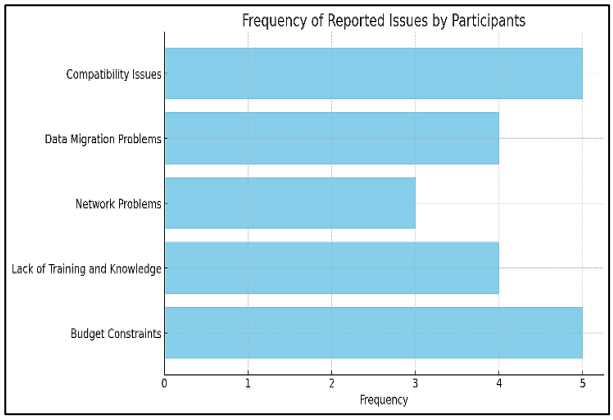


Fig. 1. Frequency of reported issues by participants.

Fig. 1 dives deeper into the challenges faced during cloud adoption within the public sector. This bar chart visually translates the participants' interview responses, highlighting the frequency of various issues reported. A quick glance allows for easy comparison, revealing which challenges were most prevalent among IT personnel.

Fig. 1 paints a clear picture of the roadblocks hindering cloud adoption in the public sector. Compatibility issues and budget constraints emerge as the most significant hurdles, with all 5 participants citing them. Data migration woes and the knowledge gap among IT personnel are also prevalent, reported by 4 participants each. Network limitations, though affecting slightly fewer (3 participants), remain a noteworthy challenge. This visual representation serves as a roadmap for successful cloud implementation. By prioritizing solutions that address these key areas compatibility, budget constraints, data migration, and IT skill development public sector organizations can pave the way for smoother adoption and ultimately unlock the full potential of cloud technologies for enhanced performance.

V. COMPARISON WITH PREVIOUS STUDIES

In order to understand how the findings of this study fit into a broader research context, it is important to compare the results of this study with those of previous studies. This section will highlight the similarities and differences between the findings of this study and the existing literature, providing a clearer picture of the challenges and opportunities in the adoption of cloud computing in the public sector.

A. Compatibility and System Integration Issues

- Findings: Participants reported significant technical complexity and compatibility issues when integrating cloud computing with existing infrastructure.
- Previous Research: The study in [4] stated that system compatibility is the main barrier to cloud adoption in the public sector. The study in [47] also emphasized the technical difficulties in integrating the new system with the old infrastructure.
- Comparison: Findings are consistent with these studies, reinforcing the view that compatibility issues are a common barrier to cloud adoption. This consistency across different contexts underscores the need for strategic planning and investment in compatible technologies to facilitate smoother integration.

B. Data Migration Problems

- Findings: Data migration was identified as a major challenge due to incompatibility of data formats and structures.
- Previous Research: The study in [32] emphasized the need for effective data migration strategies in cloud environments, noting similar challenges with compatibility and data migration difficulties. The study in [48] also documented data migration issues in their study.

- Comparison: Findings are consistent with these studies, showing that the problem of data migration is widespread and critical. This emphasizes the importance of robust data management strategies and tools in cloud computing adoption [49] [50].

C. Network Problems

- Findings: Network issues, including slow infrastructure and high latency, are significant barriers.
- Previous Studies: The study in [32] recommend increased data transmission speed and reduced latency in cloud environments, emphasizing similar network issues. The study in [19] also emphasized the need to upgrade network infrastructure to support cloud computing.
- Comparison: Findings are consistent with these studies, showing that network problems are a common barrier to cloud adoption. This consistency shows that upgrading the network infrastructure is essential for successful cloud integration.

D. Lack of Training and Knowledge

- Findings: Participants highlighted lack of training and knowledge as the main barrier.
- Previous Studies: The study in [10] emphasized the importance of training in improving the readiness of IT personnel for cloud adoption. The study in [10] also emphasized the need for ongoing professional development programs.
- Comparison: Findings are consistent with previous studies, emphasizing the need for comprehensive training programs to equip IT personnel with the skills necessary for effective cloud computing adoption [51] [52].

E. Budget Constraints

- Findings: Budget constraints were a recurring theme, with participants reporting difficulties in updating hardware and software.
- Previous Studies: The study in [25] emphasized the role of financial resources in successful IT integration. The study in [21] also documents the importance of budget allocation for IT upgrades and training.
- Comparison: Findings are consistent with these studies, suggesting that budget constraints are a significant barrier to cloud adoption. This consistency highlights the critical need for adequate funding and resource allocation to support cloud initiatives [53].

The findings of this study align closely with existing literature, validating the use of the Technology Acceptance Model (TAM) to analyze cloud computing adoption and its impact on IT personnel performance in the public sector. Consistent with previous research, participants reported significant challenges, including compatibility and system integration issues, which [4] and [47] identified as major barriers to cloud adoption. Additionally, data migration problems, such as incompatible data formats, align with findings from [32] and [48], emphasizing the need for robust migration strategies. Network issues, including slow infrastructure and high latency, echo concerns highlighted by [19] and [32], reinforcing the necessity of upgrading network infrastructure to support cloud integration. The lack of training and knowledge was another critical barrier, supporting [10]'s assertion that professional development programs are essential for cloud readiness. Furthermore, budget constraints, a recurring theme in this study, align with [21] and [25], emphasizing the need for sufficient financial resources for technology upgrades and training. These findings collectively validate TAM's constructs of perceived usefulness and perceived ease of use, demonstrating how technical challenges, organizational readiness, and resource availability influence cloud adoption. Thus, incorporating these sector-specific barriers into CCAPM's model, the study offers a comprehensive model that not only reinforces established theories but also provides practical insights for overcoming adoption challenges and enhancing IT personnel performance in the public sector.

Overall, the findings of this study are consistent with many previous studies, emphasizing common issues such as system compatibility, data migration problems, network issues, lack of training, and budget constraints. Although some specific contexts are unique to the Malaysian public sector, the key challenges identified are consistent with the global literature. This suggests that better strategies and careful planning are needed to overcome these obstacles and maximize the benefits of cloud computing.

VI. Implications

In this study, findings obtained from interviews with public sector IT personnel revealed several key issues that hinder the adoption of cloud computing. These findings include system compatibility and integration issues, data migration issues, network issues, lack of training and knowledge, and budget constraints. The practical implications of these findings are important to note for policymakers, IT managers, and public sector organizations to ensure a more effective and efficient implementation of cloud computing.

The findings of this study offer valuable insight into the real challenges faced in integrating cloud technology in public sector environments. By understanding and overcoming these barriers, stakeholders can develop more robust and relevant strategies to harness the full potential of cloud computing. Therefore, it is important to explore the practical implications arising from the findings of this study and provide recommendations that can help in increasing the uptake and use of cloud technology in the public sector.

This section will elaborate on the implications of this study's findings for three main groups: policymakers, IT managers, and public sector organizations. Each of these groups plays an important role in ensuring the successful implementation of cloud computing, and understanding the implications of this will allow them to take appropriate actions to address the identified challenges.

A. For Policy Makers

1) Strategic planning and investment: Policymakers should prioritize strategic planning and investment in compatible technologies to facilitate the integration of cloud computing with existing infrastructure. This includes allocating funds to upgrade legacy systems and ensure compatibility with new cloud-based solutions.

2) Infrastructure development: There is a need for significant investment in network infrastructure to address the issues of slow speed and high latency. Policymakers should consider funding initiatives to improve broadband infrastructure and ensure reliable internet connectivity, which is essential for effective cloud computing.

3) Training and development programs: Policy makers should support the development of comprehensive training programs for IT personnel. This includes funding professional development initiatives focused on cloud technology management and integration, ensuring staff have the necessary skills and knowledge.

B. For IT Managers

1) Implementation of best practices: IT managers should adopt best practices for data migration, including robust data management strategies and tools that address compatibility issues. This involves thorough planning and testing to ensure smooth data transfer.

2) Resource allocation: IT managers need to support adequate budget allocations to support necessary hardware and software upgrades. This includes presenting a strong business case for funding for cloud initiatives and training programs.

3) Continuous learning: IT managers should foster a culture of continuous learning and professional development in their teams. This can be achieved by organizing regular training sessions, workshops, and certifications in cloud technology.

C. For Public Sector Organizations

1) Policy development: Public sector organizations should develop clear policies and guidelines for cloud adoption, addressing compatibility, data migration, network infrastructure, training, and funding. This policy should align with broader government strategy and ensure consistent implementation across departments.

2) Collaboration and partnerships: Organizations should collaborate with technology vendors, industry experts, and other public sector entities to share knowledge and best practices. Partnerships can also help in negotiating better terms and prices for cloud services and related infrastructure upgrades.

3) Monitoring and evaluation: Implement robust monitoring and evaluation mechanisms to assess the effectiveness of cloud adoption initiatives. This includes tracking performance metrics, identifying areas for improvement, and making data-driven decisions to optimize cloud computing implementations.

VII. CONCLUSIONS

This study has conducted in-depth research on the factors that influence the adoption of cloud computing in the public sector, particularly on the performance of IT personnel. Through interviews with five participants who are IT experts in the public sector, the findings of this study offer valuable insight into the various challenges faced in the cloud computing integration process.

A summary of the key findings shows that there are several key barriers that affect the adoption of cloud computing. Among the main issues faced are system compatibility problems, deficiencies in the data migration process, inefficient network issues, lack of training and skills in data governance, as well as tight budget constraints. Study participants have provided insight into how these factors affect their ability to integrate and optimize the use of cloud technology in daily operations.

In the discussion of the findings, it is clear that these challenges are directly related to the objective of the study, which is to identify and analyze the factors that influence the adoption of cloud computing. From the data collected, main themes such as 'Compatibility and System Integration Issues', 'Data Migration Issues', 'Network Issues', 'Lack of Training and Knowledge', as well as 'Budget Constraints' were formed. These themes not only clarify the various technical and operational aspects that need to be addressed but also outline the need for a more comprehensive strategy that includes financial aspects and human resource development.

This study provides an important contribution to the field of cloud computing in the context of the public sector. First, it broadens the understanding of the specific challenges public institutions face in adopting cloud technologies, which are often different from the challenges faced by the private sector. Second, this study offers empirical evidence that supports the need for a more integrated approach in the implementation of this technology, emphasizing the importance of adapting infrastructure, training, and access to sufficient financial resources.

In conclusion, the findings from this study clearly show that although cloud computing offers many potential benefits, there are significant barriers to overcome to maximize its benefits in the public sector. Solutions to the issues identified in this study will require joint efforts between policy makers, technology managers, and IT practitioners to develop a comprehensive strategy that focuses not only on technology but also on improving organizational and individual capabilities. This will ensure that public institutions can take advantage of cloud technology to improve the efficiency and effectiveness of their services to the public.

In the context of solving the problem stated in the problem statement of this study—complexity in the integration of cloud computing with existing infrastructure and the lack of knowledge and skills in data governance in public institutions—the appropriate model needs to take into account technical, operational, and human resource development aspects. The proposed model is the "Cloud Computing Integration and Cloud Computing Acceptance and Performance Model" (CCAPM).

A. CCAPM Model Formation

1) Preliminary assessment phase:

a) Diagnostics of existing infrastructure: The first step is to perform a thorough diagnostic of the existing information technology infrastructure to assess compatibility with cloud computing systems. This includes an assessment of software, hardware, and network configuration.

b) Constraint and potential analysis: Identify constraints such as compatibility issues, data security, and network requirements, as well as potential improvements through cloud computing.

2) Development and integration phase:

a) Infrastructure development: Upgrade or replace infrastructure components that are not compatible with cloud technology. This may include replacing old software, network upgrades, and installing better data security systems.

b) System integration: Integrate cloud computing with upgraded infrastructure using protocols and tools that ensure smooth data migration and continuous operations.

3) Capacity expansion phase:

a) Training and human resource development: Implement a comprehensive training program to improve the technical skills of IT personnel in managing and maintaining cloud computing systems. This includes training in data governance, cyber security, and disaster recovery.

b) Expertise development: Establish a specialized unit in cloud computing responsible for providing technical support, guiding daily operations, and guaranteeing compliance with public sector policies.

4) Continuous evaluation phase:

a) Performance audit and evaluation: Conduct periodic audits to evaluate cloud computing system performance, effectiveness of training provided, and adherence to security protocols.

b) Iteration and refinement: Based on the results of the audit, improvements should be made to the infrastructure and operational procedures to continue improving the performance and security of the system.

This CCAPM model is expected to solve the problem of the complexity of cloud computing integration by taking a holistic approach that does not only focus on technical aspects but also the development of human resource skills and capacity in the public sector. This model also aims to create a resilient and sustainable ecosystem that can adapt new technologies more quickly and efficiently, while maintaining a high level of security and compatibility.

This model includes four main phases in the process of integration and expansion of cloud computing capacity in the public sector, ranging from initial assessment, development and integration, capacity expansion, to continuous assessment for sustainable improvement. Each step in the model has a specific task aimed at overcoming the challenges stated in the problem statement of this study.

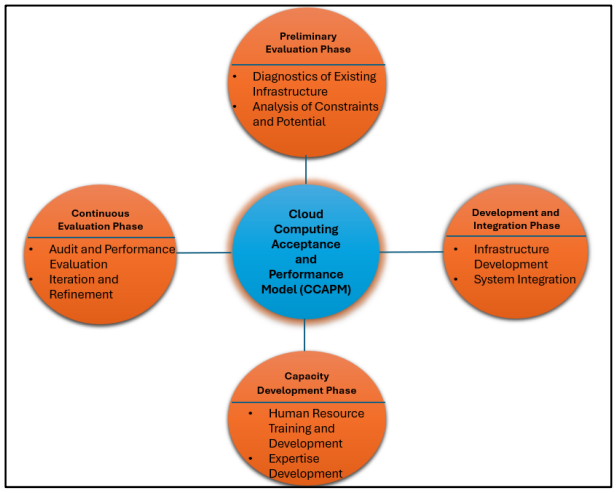


Fig. 2. Cloud Computing Acceptance and Performance Model (CCAPM).

This study has identified several key challenges in the adoption of cloud computing in the Malaysian public sector. Key findings include system compatibility and integration issues, data migration issues, network issues, lack of training and knowledge, and budget constraints. Complex system compatibility and technical problems hinder the seamless integration of cloud technology with existing infrastructure. Data migration problems, including the incompatibility of old data formats and structures with the new system, also add to these challenges. Network issues such as slow infrastructure and high latency reduce the effectiveness of using cloud technology. In addition, the lack of training and knowledge among IT personnel makes the process of adopting this technology difficult. Finally, budget constraints prevent the necessary investments to upgrade the hardware, software, and training needed to support cloud adoption. These findings emphasize the need for strategic planning, adequate investment, and comprehensive training programs to ensure the successful adoption of cloud computing in the public sector.

VIII. Future Research

Based on the findings of this study, there are several areas that require further research to deepen understanding and overcome challenges in cloud computing adoption.

1) First, future studies can focus on developing more effective data management strategies and tools to deal with data migration problems. This study can investigate the technical and methodological approaches that can be used to ensure a smoother and safer transfer of data.

2) Second, further research needs to be done to assess the effectiveness of training and professional development programs in improving the readiness of IT personnel for cloud computing. This includes research on what types of training are most effective and how best to deliver training content.

3) Third, future studies could examine methods to overcome budget constraints, including alternative funding models and strategies for obtaining financial support for cloud computing initiatives.

4) Finally, there is a need for more in-depth studies on the long-term impact of cloud computing on the performance of operations and services in the public sector. This study can help in understanding how this technology can be optimized to improve the efficiency and effectiveness of public services.

IX. LIMITATIONS

This study only involved five participants, which may not reflect the views of all IT personnel in the public sector. Also, this study is limited to the Malaysian context only.

1) Research implications: Findings from this study provide a basis for further research in cloud computing in the public sector. It suggests the need for a broader study involving more participants from a variety of geographic and technical backgrounds. Further research could explore the relationship between various technical, operational, and organizational factors with the effectiveness of cloud computing, including a comparative analysis between the public and private sectors.

2) Practical implications: These findings offer practical guidance for government agencies in planning and implementing cloud computing systems, emphasizing the need for improved infrastructure and better training programs.

ACKNOWLEDGMENT

This work was supported in part by Universiti Kebangsaan Malaysia, and in part by the Ministry of Higher Education Malaysia under Grant FRGS/1/2024/TK01/UKM/02/2.

NOORBAITI MAHUSIN received a B.Sc. from Universiti Malaysia Pahang, in 2006, and M.Sc. in Information Science from the Universiti Kebangsaan Malaysia, Bangi, Malaysia, in 2018, where she is currently pursuing her Ph.D. degree in Management Information Systems. She currently working on publish papers in reputed journals. Her current research interests including IoT and Artificial Intelligence.

HASIMI SALLEHUDIN is a Senior Lecturer with the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia. His research interests include computer security and networks, and management information systems.

NURHIZAM SAFIE MOHD SATAR received the Diploma degree from UPM, with specialization in E-learning technology information systems adoption and diffusion cloud computing, the B.H.Sc. degree from IIUM, the M.I.T. degree.

AZANA HAFIZAH MOHD AMAN received her PhD, MSc and BEng in Computer and Information Engineering from International Islamic University Malaysia. She is currently working as senior lecturer at Research Center for Cyber Security, Faculty of Information Science and Technology (FTSM), The National University of Malaysia (UKM), Malaysia. Her research areas are computer system & networking, information & network security, IoT, cloud computing, and big data.

FARASHAZILLAH YAHYA is a senior lecturer at the Faculty of Computing and Informatics at Universiti Malaysia Sabah. Her research interests include data management, digital technology, cloud computing and cybersecurity. She holds a PhD. in Computer Science from the University of Southampton, United Kingdom and a BSc. (Hons) Information System Engineering from UiTM, Shah Alam.

REFERENCES

[1] Abdelhakim, M., Abdeldayem, M. M., & Aldulaimi, S. H. (2022, June). Information technology adoption barriers in public sector. In *2022 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)* (pp. 355-360). IEEE. https://doi.org/10.1109/ICETSIS55481.2022.9888805

[2] Agarwal, P. K. (2018). Public administration challenges in the world of AI and bots. *Public Administration Review, 78*(6), 917–921. https://doi.org/10.1111/puar.12979

[3] Ali, O., Shrestha, A., Osmanaj, V., & Muhammed, S. (2020). Cloud computing technology adoption: An evaluation of key factors in local governments. *Information Technology & People, 34*(2), 666–703. https://doi.org/10.1108/ITP-03-2019-0119

[4] Amin, R., & Vadlamudi, S. (2021). Opportunities and challenges of data migration in cloud. *Engineering International, 9*(1), 41–50. https://doi.org/10.18034/ei.v9i1.529

[5] Atuluku, A. R., Osang, F. B., & Adebiyi, A. A. (2022). Cloud computing adoption and use: A systematic review. *Advances in Multidisciplinary and Scientific Research Journal Publication, 13*(4), 29–64. https://doi.org/10.22624/AIMS/CISDI/V13N4P3

[6] Beatrice, B., Burke, L., Kariuki, J., & Sereika, S. (2023). Lessons learned: Successes and challenges with technology-based data collection and intervention in a behavioral weight loss study. *Journal of the Academy of Nutrition and Dietetics, 123*(9), A16. https://doi.org/10.1016/j.jand.2023.06.042

[7] Bolin, J., & Yang, M. (2018, March). Cloud computing: cost, security, and performance. In *Proceedings of the ACMSE 2018 conference* (pp. 1-1). https://doi.org/10.1145/3190645.3190706

[8] Cordeiro, D., Francesquini, E., Amarís, M., Castro, M., Baldassin, A., & Lima, J. V. (2023). Green cloud computing: Challenges and opportunities. *Anais Estendidos do XIX Simpósio Brasileiro de Sistemas de Informação*, 129-131. https://doi.org/10.5753/sbsi_estendido.2023.229291

[9] Cordella, A., & Paletti, A. (2019). Government as a platform, orchestration, and public value creation: The Italian case. *Government Information Quarterly, 36*, 1-15. https://doi.org/10.1016/j.giq.2019.101409

[10] Cresswell, K., Hernández, A. D., Williams, R., & Sheikh, A. (2022). Key challenges and opportunities for cloud technology in health care: Semistructured interview study. *Jmir Human Factors, 9*, e31246. https://doi.org/10.2196/31246

[11] Dash, B., Sharma, P., & Swayamsiddha, S. (2023). Organizational digital transformations and the importance of assessing theoretical frameworks such as TAM, TTF, and UTAUT: A review. *International Journal of Advanced Computer Science and Applications, 14*(2), 1-6. https://doi.org/10.14569/IJACSA.2023.0140201

[12] Davis, F. D. (1986). A Technology Acceptance Model for Empirically Testing New End-User Information Systems. *Theory and Results/Massachusetts Institute of Technology*.

[13] Denny, E., & Weckesser, A. (2022). How to do qualitative research? *BJOG: An International Journal of Obstetrics & Gynaecology, 129*(7), 1166–1167. https://doi.org/10.1111/1471-0528.17150

[14] Gkika, E. C., Anagnostopoulos, T., Ntanos, S., & Kyriakopoulos, G. L. (2020). User preferences on cloud computing and open innovation: A case study for university employees in Greece. *Journal of Open Innovation: Technology, Market, and Complexity, 6*(2), 1-21. https://doi.org/10.3390/joitmc6020041

[15] Goodman, H. B., & Rowland, P. (2021). Deficiencies of Compliancy for Data and Storage: Isolating the CIA Triad Components to Identify Gaps to Security. In *National Cyber Summit (NCS) Research Track 2020* (pp. 170-192). Springer International Publishing. https://doi.org/10.1007/978-3-030-58703-1_11

[16] Gupta, N., & Sohal, A. (2022). Cloud Computing: Evolution, Research Issues, and Challenges. *Emerging Computing Paradigms: Principles,*

Advances and Applications, 1-17. https://doi.org/10.1002/9781119813439.ch1

[17] Hasan, M. Z., Hussain, M. Z., Mubarak, Z., Siddiqui, A. A., Qureshi, A. M., & Ismail, I. (2023, January). Data security and integrity in cloud computing. In *2023 International Conference for Advancement in Technology (ICONAT)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICONAT57137.2023.10080440

[18] Heinze, C., Hartmeyer, R. D., Ringgaard, L. W., Bjerregaard, A.-L., Krølner, R. F., Allender, S., Bauman, A., & Klinker, C. D. (2023). Study protocol for the Data Health Study-A data-driven and systems approach to health promotion among vocational students in Denmark, 1-28. https://doi.org/10.21203/rs.3.rs-3061625/v1

[19] Hwang, I., Kim, S., & Rebman, C. (2022). Impact of regulatory focus on security technostress and organizational outcomes: the moderating effect of security technostress inhibitors. *Information Technology & People, 35*(7), 2043-2074. https://doi.org/10.1108/itp-05-2019-0239

[20] Junnier, F. (2024). Action and understanding in the semi-structured research interview: Using CA to analyse European research scientists' attitudes to linguistic (dis) advantage. Journal of English for Academic Purposes, 68, 101355. https://doi.org/https://doi.org/10.1016/j.jeap.2024.101355

[21] Kashaija, L. S. (2022). E-records management readiness for implementation of e-government in local authorities of Singida Municipal Council. *Journal of the South African Society of Archivists, 55*, 41-55. https://doi.org/10.4314/jsasa.v55i.4

[22] Khayer, A., Talukder, Md. S., Bao, Y., & Hossain, Md. N. (2020). Cloud computing adoption and its impact on SMEs' performance for cloud supported operations: A dual-stage analytical approach. *Technology in Society, 60*, 101225. https://doi.org/10.1016/j.techsoc.2019.101225

[23] Kim, S., Andersen, K. N., & Lee, J. (2022). Platform government in the era of smart technology. *Public Administration Review, 82*(2), 362-368. https://doi.org/10.1111/puar.13422

[24] Lastanti, N., & Djasuli, M. (2024). Optimizing the use of cloud technology in public sector management control (case study of e-government in Bandung City). *Eduvest - Journal of Universal Studies, 4*(5), 2014-2109. https://doi.org/10.59188/eduvest.v4i5.1245

[25] Lulaj, E., Zarin, I., & Rahman, S. (2022). A Novel Approach to Improving E-Government Performance from Budget Challenges in Complex Financial Systems. *Complexity, 2022*(1), 1-16. https://doi.org/10.1155/2022/2507490

[26] Maelah, R., Al Lami, M. F. F., & Ghassan, G. (2019). Management accounting information usefulness and cloud computing qualities among small medium enterprises. *International Journal of Management Studies, 26*(1), 1-31. https://doi.org/10.32890/ijms.26.1.2019.10511

[27] Dibetle, M., & Kalema, B. M. (2023). Data security governance for Software-as-a-Service Cloud computing environment: A South African perspective, 1-14.

[28] Mehtälä, J., Ali, M., Miettinen, T., Partanen, L., Laapas, K., Niemelä, P. T., Khorlo, I., Ström, S., Kurki, S., Vapalahti, J., Abdelgawwad, K., & Leinonen, J. V. (2023). Utilization of anonymization techniques to create an external control arm for clinical trial data. *BMC Medical Research Methodology, 23*(1), 1-11. https://doi.org/10.1186/s12874-023-02082-5

[29] Mohamed, A., Ali, C., Fakhri, Y., & Noreddine, G. (2022, October). A survey on the challenges of data integration. In *2022 9th International Conference on Wireless Networks and Mobile Communications (WINCOM)* (pp. 1-6). IEEE. https://doi.org/10.1109/WINCOM55661.2022.9966419

[30] Mkhatshwa, B., & Mawela, T. (2023a). Cloud computing adoption in the South African public sector. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI), 11*(2), 537-552. https://doi.org/10.52549/ijeei.v11i2.4464

[31] Nanos, I. (2020, September). Cloud Computing Adoption in Public Sector: A Literature Review about Issues, Models and Influencing Factors. In *Balkan Conference on Operational Research* (pp. 243-250). Cham: Springer International Publishing.

[32] Naresh, P., P, R., Vempati, K., & Saidulu, D. (2020). Improving the data transmission speed in cloud migration by using MapReduce for Big Data. *International Journal of Engineering Technology and Management Sciences, 4*, 73-75. https://doi.org/10.46647/ijetms.2020.v04i05.013

[33] Nassaji, H. (2020). Good qualitative research. *Language Teaching Research, 24*(4), 427–431. https://doi.org/10.1177/1362168820941288

[34] Nghihalwa, E., & Shava, F. B. (2018, May). An assessment of cloud computing readiness in the Namibian government's Information Technology departments. In *2018 19th IEEE Mediterranean Electrotechnical Conference (MELECON)* (pp. 92-97). IEEE. https://doi.org/10.1109/MELCON.2018.8379074

[35] Hassan, N. H. M., Ahmad, K., & Salehuddin, H. (2020). Diagnosing the issues and challenges in data integration implementation in public sector. *Int. J. Adv. Sci. Eng. Inf. Technol, 10*, 529-535.

[36] Novianto, N. (2023). Systematic literature review: Models of digital transformation in the public sector. *Policy & Governance Review, 7*(2), 170. https://doi.org/10.30589/pgr.v7i2.753

[37] Pańkowska, M., Pyszny, K., & Strzelecki, A. (2020). Users' adoption of sustainable cloud computing solutions. Sustainability, 12(23), 1-21. https://doi.org/10.3390/su12239930

[38] Parthasarathy, S., Sivagurunathan, S., & Subramanian, G. H. (2022). What should a startup know about software customization? *International Journal of Information Technology Project Management, 13*(1), 1–13. https://doi.org/10.4018/IJITPM.313945

[39] Pollock, D., Alexander, L., Munn, Z., Peters, M. D., Khalil, H., Godfrey, C. M., ... & Tricco, A. C. (2022). Moving from consultation to co-creation with knowledge users in scoping reviews: guidance from the JBI Scoping Review Methodology Group. *JBI evidence synthesis, 20*(4), 969-979. https://doi.org/10.11124/jbies-21-00416

[40] Sadlier, A., & Baksh, N. (2022, March). Real-Time-as-a-Service Increases Efficiency, Agility and Consistency. In *SPE/IADC Drilling Conference and Exhibition* (p. D031S020R002). SPE. https://doi.org/10.2118/208790-MS

[41] Sallehudin, H. A. S. I. M. I., Razak, R. C., Ismail, M. O. H. A. M. M. A. D., Fadzil, A. F. M., & Baker, R. O. G. I. S. (2019). Cloud computing implementation in the public sector: factors and impact. *Asia-Pacific Journal of Information Technology and Multimedia, 7*(2-2), 27-42. https://doi.org/10.17576/apjitm-2018-0702(02)-03

[42] Yang, P., Xiong, N., & Ren, J. (2020). Data security and privacy protection for cloud storage: A survey. *IEEE Access, 8*, 131723–131740. https://doi.org/10.1109/ACCESS.2020.3009876

[43] Shetty, J. P., & Panda, R. (2023). Cloud adoption in Indian SMEs–an empirical analysis. *Benchmarking: An International Journal, 30*(4), 1345-1366. https://doi.org/10.1108/bij-08-2021-0468

[44] Stieninger, M., Nedbal, D., Wetzlinger, W., Wagner, G., & Erskine, M. A. (2022). Factors influencing the organizational adoption of cloud computing: A survey among cloud workers. *International Journal of Information Systems and Project Management, 6*(1), 5–23. https://doi.org/10.12821/ijispm060101

[45] Thobejane, M., & Marnewick, C. (2020). The effective implementation of cloud computing through project management: Conceptual framework. *Journal of Contemporary Management, 17*(2), 416–444. https://doi.org/10.35683/jcm20079.82

[46] Wang, X., Xia, D., Wang, Y., Xu, S., & Gui, L. (2020). A cross-sectional study of heat-related knowledge, attitude, and practice among naval personnel in China, 1-15. https://doi.org/10.21203/rs.2.20828/v1

[47] Zhao, J., & Wang, W. (2019). Creative Combination of Legacy System and Map Reduce in Cloud Migration. *International Journal of Performability Engineering, 15*(2), 579-590. https://doi.org/10.23940/ijpe.19.02.p22.579590

[48] Zutavern, S., & Seifried, J. (2021). Exploring well-being at work—An interview study on how IT professionals perceive their workplace. *Frontiers in Psychology, 12*, 1-18. https://doi.org/10.3389/fpsyg.2021.688219

[49] Al-Jumaili, A. H. A., Muniyandi, R. C., Hasan, M. K., Singh, M. J., Paw, J. K. S., & Amir, M. (2023). Advancements in intelligent cloud computing for power optimization and battery management in hybrid renewable energy systems: A comprehensive review. *Energy Reports, 10*, 2206-2227. https://doi.org/10.1016/j.egyr.2023.09.029

[50] Badie, N., Hussin, A. R. C., Yadegaridehkordi, E., Singh, D., & Lashkari, A. H. (2023). A SEM-STELLA approach for predicting decision-makers' adoption of cloud computing data center. *Education and Information*

Technologies, 28(7), 8219-8271. https://doi.org/10.1007/s10639-022-11484-9

[51] Usman, L. O., Muniyandi, R. C., & Usman, M. A. (2023). Efficient Neuroimaging Data Security and Encryption Using Pixel-Based Homomorphic Residue Number System. *SN Computer Science, 4*(6), 834. https://doi.org/10.1007/s42979-023-02297-9

[52] Meri, A., Hasan, M. K., Dauwed, M., Jarrar, M. T., Aldujaili, A., Al-Bsheish, M., ... & Kareem, H. M. (2023). Organizational and behavioral attributes' roles in adopting cloud services: An empirical study in the healthcare industry. *Plos one, 18*(8), 1-23. https://doi.org/10.1371/journal.pone.0290654

[53] Naseri, N. K., Sundararajan, E., & Ayob, M. (2023). Smart Root Search (SRS) in Solving Service Time–Cost Optimization in Cloud Computing Service Composition (STCOCCSC) Problems. *Symmetry, 15*(2), 272. https://doi.org/10.3390/sym15020272

TPGR-YOLO: Improving the Traffic Police Gesture Recognition Method of YOLOv11

Xuxing Qi[1], Cheng Xu[2]*, Yuxuan Liu[3], Nan Ma[4], Hongzhe Liu[5]
Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing, China[1, 2, 3, 5]
College of Information and Technology, Beijing University of Technology, Beijing, China[4]
Beijing Qiangqiang Yuanqi Technology Co., Ltd, Beijing, China[1, 2, 3]

***Abstract*—In open traffic scenarios, gesture recognition for traffic police faces significant challenges due to the small scale of the traffic police and the complex background. To address this, this paper proposes a gesture recognition network based on an improved YOLOv11. This method enhances feature extraction and multi-scale information retention by integrating RFCAConv and C2DA modules into the backbone network. In the Neck part of the network, an edge-enhanced multi-branch fusion strategy is introduced, incorporating target edge information and multi-scale information during the feature fusion phase. Additionally, the combination of WIoU and SlideLoss loss functions optimizes the positioning of bounding boxes and the allocation of sample weights. Experimental validation was conducted on multiple datasets, and the proposed method achieved varying degrees of improvement in all metrics. Experimental results demonstrate that this method can accurately perform the task of recognizing traffic police gestures and exhibits good generalization capabilities for small targets and complex backgrounds.**

***Keywords*—Traffic police gesture recognition; loss function; YOLO algorithm; multi-scale feature fusion**

I. INTRODUCTION

In the Road Traffic Law, it is stipulated that when traffic signals and signals issued by traffic control officers conflict, the latter shall prevail [1]. The quick and accurate identification of traffic control gestures is crucial for preventing traffic accidents, alleviating congestion, and improving road usage efficiency.

Traditional gesture recognition methods primarily rely on image processing techniques, separating hand regions through skin color detection [2, 3] and tracking hand movements. However, these methods are effective only in simple and controlled environments. To address this issue, researchers have explored the use of various types of sensors [4, 5, 6, 7, 8] to directly capture motion data. Although this approach achieves higher accuracy, its expensive hardware installation costs and complex real-time communication requirements limit practical implementation.

In recent years, multi-modal recognition technologies have gained prominence, with advancements such as those leveraging wireless signals [9, 10, 11] (e.g., Wi-Fi, infrared, or radar) and deep learning-based multi-modal recognition techniques. These methods alleviate problems caused by lighting variations and occlusions to some extent. However, in complex traffic scenarios, wireless signals are susceptible to interference, significantly impacting recognition accuracy.

With the rapid development of deep learning technologies, skeleton-based recognition methods [12, 13, 14, 15] have become a research focus. By extracting precise pose features, these methods have achieved notable results in gesture recognition. However, existing research mainly targets skeletal data at close distances, leaving gaps in recognition performance in open traffic scenarios with long-range, multi-target, and complex environments.

Effectively identifying traffic control gestures from a crowd under various lighting conditions and positions in open traffic environments remains a challenge. This study addresses the issue by optimizing methods for traffic control gesture recognition, eliminating the need for depth information or skeletal data, thereby significantly reducing processing time. The main contributions of this paper are as follows:

1) Based on publicly available traffic videos in complex scenarios, a new dataset was constructed. It contains eight types of traffic control gestures, characterized by small targets and complex backgrounds, providing a solid data foundation for future research.

2) To address challenges like significant scale variations and complex backgrounds in traffic control scenarios, we designed the C3k2RFCA and C2DA modules. These effectively preserve multi-scale information and detail features. Additionally, multi-scale deep convolution and the Sobel operator were integrated into the Neck part, achieving deep fusion of edge and scale information.

3) The combination of WIoU (Wise-IoU) and an improved SlideLoss was employed as the loss function. WIoU improves the precision of bounding box localization, while SlideLoss optimizes sample weight allocation. Their synergy significantly enhances the performance of traffic control gesture detection in open scenarios.

4) Comprehensive experiments on the custom dataset and two public datasets demonstrate the proposed method's high accuracy and strong generalization capability. It effectively handles the challenges of small targets and complex backgrounds.

The structure of this paper is as follows: Section II introduces the related work, Section III describes the improved YOLO method and its application algorithm for traffic police

*Corresponding Author

gesture recognition, Section IV reports the experimental results, and Section V provides the conclusions.

II. Related Work

A. YOLOv11 Algorithm

First, YOLOv11, as a groundbreaking advancement in the field of object detection, achieves significant improvements in detection efficiency and accuracy through a series of innovative modules and optimized designs. The core innovations of this method include the newly introduced C3k2 module and C2PSA module. C3k2 is an improved version of the CSP (Cross Stage Partial) bottleneck structure. By integrating a two-layer convolutional structure, it significantly enhances feature extraction capabilities while maintaining low computational overhead. This optimization makes the method particularly effective in handling complex scenarios, multi-scale targets, and detailed features. Meanwhile, the C2PSA module incorporates a position-sensitive attention mechanism that dynamically focuses on the critical spatial regions where targets are located, enhancing the capture and processing of spatial information. This is especially effective in tasks involving small object detection and precise localization.

Additionally, YOLOv11 introduces lightweight optimizations in the detection head. By simplifying the network structure and reducing unnecessary computational processes, the parameter count has been reduced by approximately 20% compared to the previous generation. This significantly lowers computational costs while maintaining high accuracy. This lightweight design not only enhances runtime speed but also supports the efficient deployment of the method on resource-constrained devices.

B. Gesture Recognition

For traffic police gestures, it has been observed that each gesture can be represented by key hand movements. Therefore, static gesture recognition becomes an effective implementation method. The development of object detection algorithms has provided new perspectives for this field. Zhang et al. [16] improved the YOLOv3 algorithm to achieve gesture recognition; Wang et al. [17] combined YOLO algorithm and Kalman filter to realize high-precision gesture recognition and tracking; Saxena et al. [18] utilized YOLOX and YOLOv5 methods to achieve high-accuracy and high-efficiency gesture recognition, facilitating communication for individuals with hearing and speech impairments; Helal et al. [19] combined CNN and YOLO techniques to recognize sign language alphabets, achieving high accuracy while optimizing training time.

The YOLO algorithm is widely applied due to its accuracy and speed, but it still requires adaptations for different scenarios. For instance, Yang et al. [20] proposed the YOLO-LRHG method, combining YOLO with attention mechanisms to address long-distance human-machine interaction gesture recognition problems. Zhou et al. [21] designed the PEA-YOLO lightweight network, improving recognition performance through adaptive spatial feature pyramids and multi-path feature fusion.

Most of the previous research methods have been limited to improving the accuracy of gesture recognition at close range and have not addressed distant scenes with increased interference. This paper builds upon the YOLOv11 network and introduces enhancements to tackle challenges such as the loss of detailed features and the presence of complex backgrounds in far-distance open traffic scenarios. The goal is to improve the detection of key gestures in complex environments, providing an efficient and accurate gesture recognition solution for intelligent traffic management and autonomous driving.

III. Model

In this study, the RFCAConv module was utilized for sampling operations, and the C3k2 module was restructured and optimized. Additionally, the original C2PSA module was replaced with the C2DA module, and an edge-enhanced multi-branch fusion strategy was introduced in the Neck section. To further improve the performance and efficiency of the method, WIoU [22] and Slide-Loss [23] were newly proposed. The TPGR-YOLO overall structure is shown in Fig. 1.

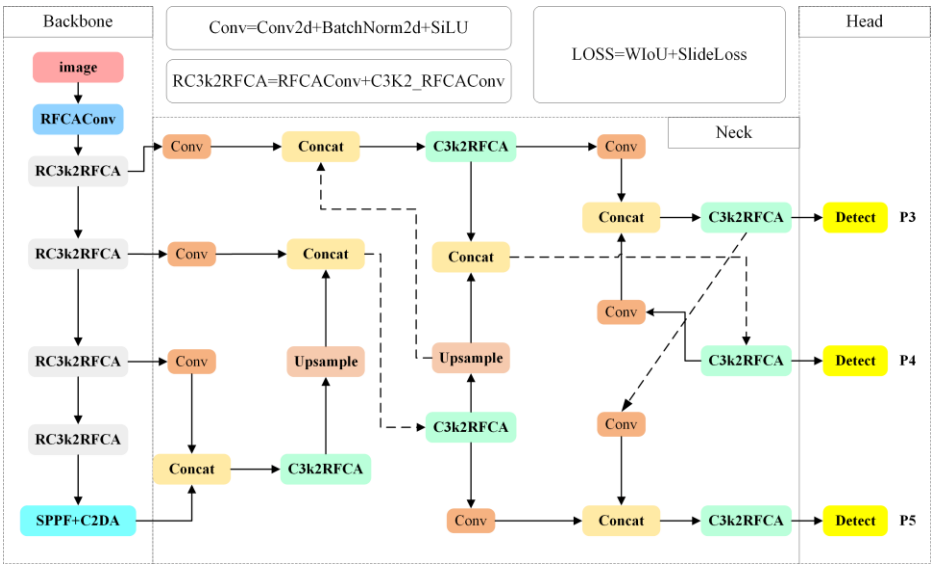


Fig. 1. TPGR-YOLO overall structure.

A. C3K2RFCA

RFCAConv [24] integrates receptive field generation convolution and coordinate attention mechanism (CA), effectively addressing the limitations of traditional convolution, where shared convolution kernels result in insufficient sensitivity to feature locations. It also overcomes the inadequacy of conventional spatial attention mechanisms in fully extracting regional features from contextual information. RFCAConv generates receptive field features, applies weight mapping to them, and then employs standard convolution operations to produce output features. This process combines the advantages of local feature enhancement and global positional attention, resulting in finer-grained and more effective attention weights. As shown in Fig. 2, its structure has been optimized and improved.

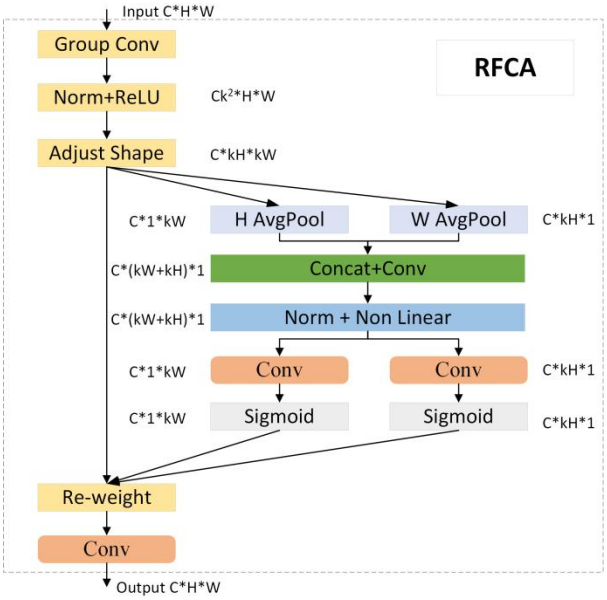


Fig. 2. RFCAConv structure.

For the input feature map $X \in R^{B\times C\times H\times W}$, the RFCAConv module first generates features $G \in R^{B\times(C\cdot k^2)\times H\times W}$ through the receptive field generation module. In the receptive field features G, the feature of each pixel is expanded into a high-dimensional feature map containing the local receptive field. Subsequently, the receptive field features G are unfolded into the spatial dimension.

$$G = \text{Conv}(X, \text{kernel\_size} = k, groups = C) \tag{1}$$

$$\begin{aligned} &G = \text{reshape}(G, (B, C, k^2, H, W)) \\ &\rightarrow G = rearrange(G, (H \cdot k, W \cdot k)) \end{aligned} \tag{2}$$

Next, the receptive field features G undergo global pooling along the horizontal and vertical directions, resulting in features X_h and X_w, which represent the global features in the horizontal and vertical directions, respectively. These two features are concatenated and processed through a 1×1 convolution, batch normalization, and the h-swish activation function to generate intermediate features Y. Y is then split into two parts: the horizontal features $X_h^{'}$ and the vertical features $X_w^{'}$. These parts are further processed through two separate 1×1 convolutions to generate attention weights in the two directions: horizontal weight A_h and vertical weight A_w, each normalized using a Sigmoid activation function.

$$X_h = Pool_h(G), X_w = Pool_w(G) \tag{3}$$

$$Y = \text{Act}(BN(Conv1(X_h \oplus X_w))) \tag{4}$$

$$X_h', X_w' = split(Y) \tag{5}$$

$$A_h = \sigma(Conv(X_h')), A_w = \sigma(Conv(X_w')) \tag{6}$$

Finally, the receptive field features G are pixel-wise weighted by multiplying A_h and A_w, resulting in enhanced feature representations. The weighted features are then passed through a convolutional layer to produce the final output feature map O. The principle is given in Eq. (1) – Eq. (6). This output feature map not only incorporates local receptive field information but also integrates global positional attention in the horizontal and vertical directions.

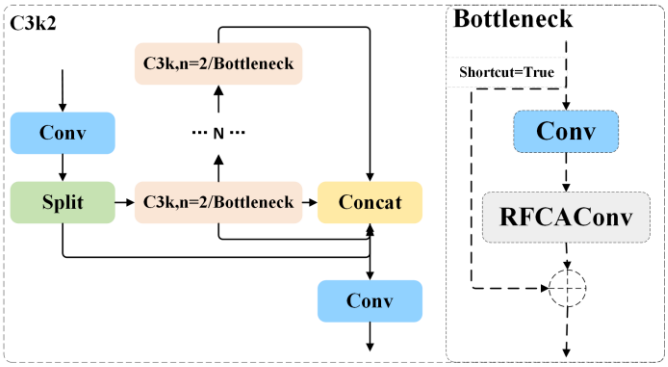


Fig. 3. C3k2RFCA structure.

At the same time, by introducing the global position sensitivity and local receptive field enhancement capabilities of RFCAConv, the C3k2 module has been structurally redesigned. This addresses the problem of reduced feature map resolution caused by multiple downsampling and stride convolutions in the C3k2 module, while compensating for the limited receptive field of the 3×3 convolution kernel, which hampers the ability to capture small objects and fine-grained features. The structure of the redesigned C3k2 module is shown in Fig. 3.

The input features first pass through RFCAConv, undergoing local receptive field enhancement and directional attention weighting to obtain feature representations with a stronger global receptive field. Subsequently, RFCAConv replaces traditional stride convolution for downsampling, further reducing resolution loss while enhancing focus on target regions. Finally, the output feature map retains more detailed information, providing stronger small-object detection capabilities and improved spatial perception.

B. C2DA

The C2PSA module, based on the global feature channel distribution attention mechanism, has limited local feature modeling capabilities, which can dilute spatial detail information and lead to reduced detection accuracy. Drawing inspiration from the concept of Deformable Attention [25], this paper proposes the C2DA module, introducing adaptive local feature learning based on deformable sampling to address the shortcomings of the C2PSA module. Its structure is shown in Fig. 4.

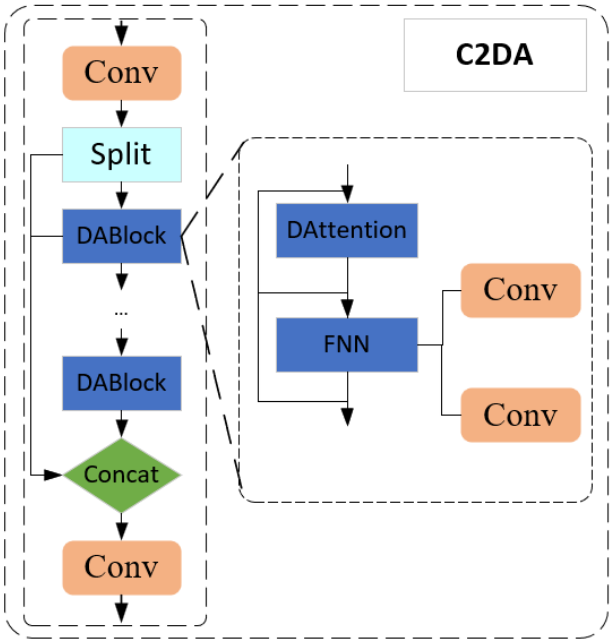


Fig. 4. C2DA structure.

First, the input features undergo dynamic offset sampling. The offset generation network in the C2DA module generates adaptive sampling positions based on the input features, enabling precise capture of critical features and detailed significant regions of small objects while avoiding interference from irrelevant areas. Second, a local sampling range is introduced by restricting dynamic sampling to a local area near the reference points. The offset magnitude is controlled using the tanh function and a scaling factor, ensuring the smoothness and local consistency of the sampling process. This approach enhances the modeling of object boundaries and details. In addition, C2DA retains the global characteristics of the attention mechanism. By leveraging query-key-value matching, it explores global relationships between pixels, allowing the capture of contextual information relevant to small objects and preventing their features from being overshadowed by complex backgrounds.

The generation of dynamic offsets begins by grouping the input feature map $X \in R^{B\times C\times H\times W}$. For each group of features $X_{group} \in R^{B\times \frac{C}{groups}\times H\times W}$, a convolutional network is used to generate the offset values $\Delta p = Conv_{offset}(X_{group})$. The offset values are $\Delta p \in R^{B\cdot g\times 2\times H_k\times W_k}$ then normalized and added to the reference points to obtain the offset positions p, where p represents the grid reference points. The principle is Eq. (7) – (8).

Using the offset positions p, the input feature map x is dynamically sampled through bilinear interpolation, producing sparsely sampled features $X_{sampled} \in R^{B\times C\times 1\times N}$.

$$p = \tanh(\Delta p)\cdot offset\_range + P_{ref} \tag{7}$$

$$X_{sampled} = GridSample(x, p) \tag{8}$$

The features obtained from dynamic offset sampling are then fed into the multi-head attention module. First, a convolutional layer is used to compute the values of Query(q), Key(k), and Value(v), where q, k, and v are the respective feature representations. Next, the similarity between the Query and Key is computed to obtain the attention weights A. The attention weights A are used to perform a weighted aggregation of v. The aggregated result z is then passed through the projection output layer W_{proj} to generate the final features. The input features x are added to the final features via a residual connection, resulting in the output y, where W_{proj} is the projection output layer, and X_{res} is914 the residual connection. The principle is given in Eq. (9) - Eq. (13).

$$L = H\cdot W, C_h = \frac{C}{heads} \tag{9}$$

$$q = W_q x, k = W_k x_{sampled}, v = W_v x_{sampled} \tag{10}$$

$$A = Soft\max(\frac{q\cdot k^T}{\sqrt{C_h}} + PE(q,k)) \tag{11}$$

$$z = A\cdot v \tag{12}$$

$$y = W_{proj}(z) + x_{res} \tag{13}$$

The C2DA module incorporates the powerful characteristics of the dynamic attention mechanism, leveraging deformable convolution and self-attention to enhance the ability to focus on critical spatial regions in visual tasks. It demonstrates strong adaptability while effectively balancing local details and global context.

C. SAFPN

Traffic police gesture recognition is often challenged by various complex background interferences, such as lighting conditions, vehicles, and pedestrians, which significantly increase the difficulty of gesture recognition. In scenarios with lighting variations or occlusions, edge features often provide more information than content features. However, the neck structure of the original network primarily focuses on high-level semantic features, lacking effective preservation of edge

and local details [26], which can easily lead to the loss of target edge information.

To address this issue, this paper proposes an edge-enhanced multi-branch fusion structure (SAFPN), which specifically includes the SSAF (Shallow Edge-Assisted Fusion) module and the AAF(Advanced Assisted Fusion) module. These modules are designed to strengthen edge information extraction and the comprehensive utilization of multi-scale features.

The SSAF module combines the capabilities of the Sobel operator and SAF, fusing the shallow extracted target edge information, high-resolution shallow features, information from the same level of the backbone network, and deep features. This approach significantly preserves critical details in shallow layers. The structure is illustrated in Fig. 5 of the paper.

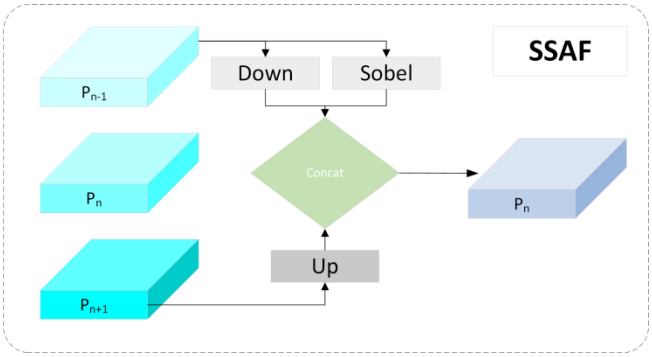


Fig. 5. SSAF structure.

SSAF demonstrates remarkable advantages in small object detection and handling complex backgrounds. The Sobel operator extracts the gradient information of images, effectively capturing the edge features of targets. It also suppresses interference from smooth areas in complex backgrounds, retaining only the edge regions with significant changes, thereby reducing the impact of background noise. Additionally, as a fixed edge extractor, the Sobel operator requires no additional training parameters, adds negligible computational complexity, and yet significantly enhances the network's sensitivity to target regions. The output results after applying SSAF are as follow [see Eq. (14)]

$$P'_n = concat\ (S(P_{n-2}), P_{n-1}, P_n, P'_{n+1}) \tag{14}$$

The AAF module aggregates shallow high-resolution features, shallow low-resolution features, information from sibling layers at the same level, and comprehensive information from the previous layer. This enables the transfer of more diverse gradient information to the network's output layer. As shown in Fig. 6, the final output layer combines information from multiple different layers, significantly enhancing the network's ability to detect targets at various scales. The output results after applying AAF are as follows. Additionally, the AAF module utilizes 1×1 convolutions to adjust the channel information of each layer, optimizing the contribution of features from each layer to the final result [see Eq. (15)]

$$P''_n = concat\ (P'_{n-1}, P''_{n-1}, P'_n, P'_{n+1}) \tag{15}$$

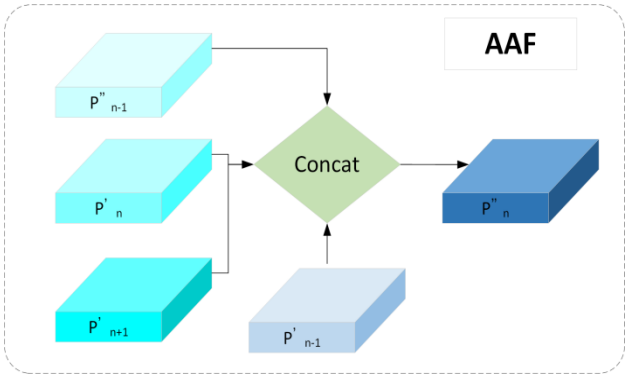


Fig. 6. AAF structure.

The SSAF module effectively reduces background noise interference, while the multi-scale fusion and dynamic weighting mechanism of the AAF module further integrate effective features. The overall design not only enhances the detection capability for small objects and prominent regions but also improves the method's robustness and detection accuracy in complex backgrounds.

D. Loss

In traffic police gesture recognition tasks in open scenarios, training data inevitably contains low-quality samples [27] due to issues with the images themselves or their annotations. Additionally, this task faces the challenge of imbalanced sample distribution, with an uneven distribution of hard and easy samples. These factors significantly limit the effectiveness of traditional loss functions in training.

To address this issue, we creatively introduce WIoU and Slide Loss, tackling the problem from two perspectives: target box optimization and sample weight allocation. Together, these approaches effectively mitigate the critical challenges in traffic police gesture recognition tasks.

1) WIoU: Since training data inevitably contains low-quality samples, geometric factors such as distance and aspect ratio exacerbate the penalties imposed on these samples, thereby reducing the method's generalization performance. When the anchor box aligns well with the target box, a good loss function should attenuate the penalties from geometric factors. Reduced training intervention can enhance the method's generalization capability. Based on this principle, Tong developed distance attention and introduced WIoU v1 with a two-layer attention mechanism, The principle is given in Eq. (16) – Eq. (18).

$$L_{WIoUv1} = R_{WIoU} L_{IoU} \tag{16}$$

$$R_{WIoU} = \exp\left(\frac{(x-x_{gt})^2 + (y-y_{gt})^2}{(W_g^2 + H_g^2)^*}\right) \tag{17}$$

$$L_{IoU} = 1 - IoU = 1 - \frac{W_i \cdot H_i}{S_u} \tag{18}$$

Here, W_g and H_g represent the size of the minimum enclosing box, x and y are the center coordinates of the anchor box, and X_{gt} and Y_{gt} are the center coordinates of the ground truth box. The symbol $*$ indicates that W_g and H_g are detached from the computation graph to prevent gradients that hinder convergence. W_i and H_i denote the width and height of the predicted box, while S_u represents the union area.

WIoU v3 builds upon WIoU v1 by introducing a dynamic non-monotonic focusing mechanism, using outlier degree instead of IoU for anchor box quality evaluation. It also provides a more flexible gradient gain allocation strategy. This strategy reduces the competitiveness of high-quality anchor boxes while mitigating harmful gradients from low-quality samples, allowing the method to focus on optimizing medium-quality anchor boxes. Consequently, it enhances overall detection performance. The principle is given in Eq. (19) - Eq. (20)

$$L_{WIoUv3} = r \cdot L_{WIoUv1}, r = \exp(-\beta \cdot \delta) \tag{19}$$

$$\beta = \frac{L^*_{IoU}}{\overline{L}_{IoU}} \in [0, +\infty) \tag{20}$$

Here, β represents the outlier degree of the anchor box, and L^*_{IoU} is the exponentially weighted moving average of L_{IoU}. L^*_{IoU} is a dynamic value used to measure the quality of the anchor box. δ is a hyperparameter that controls the mapping relationship between the outlier degree β and the gradient gain r. When $\beta = \delta$, $r = 1$, meaning the anchor box receives the highest gradient gain.

In scenarios involving long distances and small targets, WIoU can provide more accurate target localization by applying weighted corrections to the shape, size, and offset position of the bounding box. It effectively alleviates the issue of bounding box displacement in complex backgrounds.

2) SlideLoss: Slide Loss introduces a dynamic modulation factor to segmentally adjust sample weights based on their IoU values. Additionally, Slide Loss dynamically adjusts the IoU threshold to ensure stability in loss calculation.

It categorizes samples into three IoU intervals: low, medium, and high. Different weights are assigned to each interval: low-IoU samples retain their original weights, medium-IoU samples are assigned amplified weights to enhance focus on hard-to-classify samples, and high-IoU sample weights are exponentially decayed to reduce interference from easily classifiable samples.

This mechanism enables the model to automatically assign higher optimization weights to challenging samples in gesture detection tasks under long-distance and complex background conditions. It addresses the shortcomings of Binary Cross-Entropy (BCE) loss, including imbalanced positive-negative sample weighting, insufficient optimization for hard-to-classify samples, and inadequate focus on low-confidence samples. The principle is given in Eq. (21) – Eq. (25).

$$L_{Slide} = L_{base} \cdot \alpha \tag{21}$$

$$\alpha_1 = 1.0, \text{ if true} \leq \beta - 0.1 \tag{22}$$

$$\alpha_2 = \exp(1.0 - \beta), \text{if } \beta - 0.1 < \text{true} < \beta \tag{23}$$

$$\alpha_3 = \exp(-(\text{true} - 1.0)), \text{ if true} \geq \beta \tag{24}$$

$$\text{auto\_iou} = \max(\beta, 0.2) \tag{25}$$

β represents the dynamic IoU threshold, constrained to a minimum value of 0.2. Low-IoU samples $true \leq \beta - 0.1$ retain their loss unchanged $\alpha_1 = 1.0$, medium-IoU samples $\beta - 0.1 < true < \beta$ have their loss amplified to $\exp(1.0 - \beta)$ to enhance focus on hard-to-classify cases, and high-IoU samples $true \geq \beta$ have their loss exponentially decayed to $\exp(-(\text{true} - 1.0))$, reducing the interference of easily classified samples on the overall loss. By assigning adaptive weights to samples at different IoU levels using the modulation factor α, Slide Loss optimizes hard-to-classify samples, preserves low-IoU samples, and minimizes the influence of high-IoU samples. This design makes Slide Loss particularly effective for traffic police gesture detection in long-distance and complex background scenarios.

WIoU enhances the geometric information modeling capability of bounding boxes through dynamic weighting, while Slide Loss adjusts sample weights based on IoU differences, prioritizing the optimization of rare and challenging samples with weaker gradients. The ingenious combination of these two methods effectively addresses the critical challenges in traffic police gesture recognition tasks, significantly improving the detection performance and generalization ability of the approach.

IV. Experience

To evaluate the effectiveness of the proposed method in traffic police gesture recognition tasks with long distances and complex backgrounds, this study conducted comprehensive testing and assessment across various scenarios. The proposed method was trained and validated on a custom-built traffic police gesture dataset and compared with several mainstream object detection methods. To further demonstrate the algorithm's generalization ability, performance comparisons were also made with leading algorithms on the WiderPerson and VOC datasets. Additionally, the study performed ablation experiments to analyze the contributions of different components of the method in detail, and visual methods were employed to conduct an in-depth discussion of the experimental results.

A. Setting

The improved YOLOv11 method was used for training and testing during the experiments. The experimental system was based on Ubuntu 22.04. Table I presents the experimental environment and the configurations of some hyperparameters.

TABLE I. EXPERIMENTAL ENVIRONMENT CONFIGURATION

Project Environment	Parameter Setting
Framework	Pytorch
CPU	Xeon(R) Platinum 8352V
GPU	3080*2
Batch Size	64
Epoch	300
Workers	8
CUDA	12.1
Optimizer	SGD

B. Datasets and Evaluation Metrics

The datasets used in this study include the self-collected CTPG dataset, the WiderPerson dataset, and the VOC dataset. Accuracy is evaluated using metrics such as Precision (P), Recall (R), and Mean Average Precision (mAP).

The self-collected CTPG dataset is sourced from Beijing University of Technology's public transportation control gestures [28]. It includes difficult samples from midday and evening scenarios, captured from three different viewing angles: left, center, and right. During the experiment, frames were randomly sampled from the provided video samples and annotated using the makesense platform, resulting in 5,857 training images, 1,200 validation images, and 1,200 test images. The gestures in the dataset represent real-world road scenarios, characterized by long distances, complex backgrounds, and occlusion. The gestures consist of eight categories: "Stop," "Straight Ahead," "Left Turn," "Left Turn and Wait," "Right Turn," "Lane Change," "Slow Down," and "Pull Over." Example samples are shown in Fig. 7.

The WiderPerson dataset [29] is a large-scale dataset for pedestrian detection, containing 13,382 images and over 400,000 high-quality annotated pedestrian instances. It covers complex scenarios such as occlusion, dense crowds, small targets, and various pedestrian poses. The dataset is sourced from real-world diverse scenes and provides precise bounding box annotations, supporting training and evaluation for pedestrian detection tasks. Its scene diversity and challenges make it an important benchmark dataset for pedestrian detection research.

The PASCAL VOC dataset [30] is one of the most widely used standard datasets in the field of computer vision, focusing on tasks such as object detection, image classification, and semantic segmentation. The dataset includes 20 common object categories and provides precise annotation information, including bounding boxes, category labels, and pixel-level segmentation masks. The VOC dataset features diverse scenes and object variations, and its clear evaluation metrics and rich challenges make it a key benchmark for object detection and image segmentation research.

C. Comparison Experiments

To demonstrate the superiority of the proposed method, this study compares several mainstream algorithms on the aforementioned datasets, including single-stage and two-stage conventional CNN algorithms, lightweight algorithms, and Transformer-based methods.

As shown in Table II, a comprehensive comparison is made between multiple object detection methods in terms of parameter count, computational cost, model size, performance metrics, and inference speed. The proposed TPGR-YOLO outperforms other methods across several key indicators. TPGR-YOLO has only 2.8 M parameters, a computational cost of 7.6 GFLOPs, and a model size of just 6 M. It achieves superior performance in accuracy (P = 94.4%), recall (R = 83.4%), mAP@50 (93.6%), mAP@50:95 (72.7%), and inference speed (FPS = 72.7), demonstrating its excellence in lightweight design and high-performance object detection. In contrast, traditional methods such as Faster R-CNN have a parameter count as high as 60 M, a computational cost of 255 GFLOPs, and while they show decent accuracy (P = 79.7%, mAP@50 = 86.2%), their inference speed is only 58.5 FPS, making them unsuitable for real-time applications. ASF-YOLO achieves a high FPS while achieving good accuracy due to its lightweight treatment; RT-DETR balances accuracy (P = 90.2%, R = 77.8%) and speed (FPS = 65.8), but its parameter count is 42 M, with a computational cost of 136 GFLOPs and a model size of 166 M, which is not suitable for resource-constrained scenarios. The YOLO series shows good performance in terms of both lightweight design and performance, but still lags behind TPGR-YOLO. For example, YOLOv11n has 2.6 M parameters and an accuracy of 90.1%, but its mAP and recall are lower than TPGR-YOLO. LD-YOLOv10 shows improvements in accuracy (P = 89.7%, R = 77.1%) and speed (FPS = 65.2), but its parameter count and computational cost are 8.1 M and 24.7 GFLOPs, respectively, making it less efficient than TPGR-YOLO.

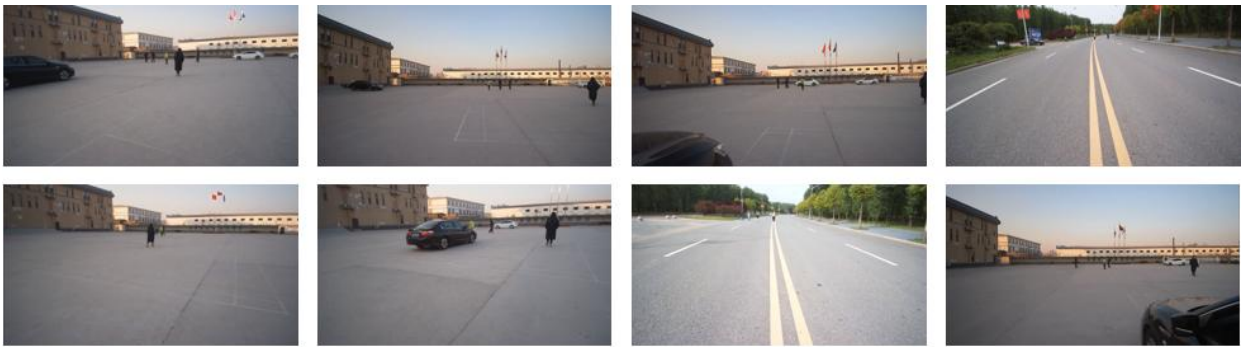


Fig. 7. Several types of action diagram.

TABLE II. COMPARISON OF MAINSTREAM ALGORITHMS ON THE CTPG DATASET

Models	Para(M)	GFLOPs(G)	Model size(M)	P	R	mAP@50	mAP@50:95	FPS
Faster R-CNN[31]	60	255	109	79.7	72.3	86.2	58.5	30
YOLOv5[32]	7.2	16.5	14	82.1	70.4	88.0	60.5	108
YOLOv7	36.9	104.7	72.5	85.3	73.2	86.3	63.4	70
ASF-YOLO[33]	15.9	79.1	25	88.5	74.5	86.0	64.7	137
YOLOv9s	9.7	37.6	15.2	87.6	75.2	85.6	63.4	100
LD-YOLOv10[34]	8.1	24.7	17	89.7	77.1	88.3	65.2	102
RT-DETR[35]	42	136	166	90.2	77.8	84.9	65.8	40
YOLOv11n	2.6	6.3	5.5	90.1	78.2	88.2	67.4	140
TPGR-YOLO	2.8	7.6	6.0	94.4	83.4	93.6	72.7	135

To evaluate the generalization ability of the algorithm, this study compares the proposed method with mainstream algorithms on the WiderPerson dataset. The results are shown in Table III. The experimental results demonstrate that the proposed TPGR-YOLO outperforms other object detection methods in multiple metrics, including precision, recall, and mAP. TPGR-YOLO achieves a precision of 94.5%, recall of 81.0%, mAP@50 of 89.2%, and mAP@50:95 of 69.6%, leading across all performance indicators.

In comparison, YOLOv5s and YOLOv7-tiny show weaker performance, with precisions of 82.2% and 83.2%, mAP@50 of 78.9% and 82.7%, and mAP@50:95 of only 59.7% and 61.7%. As the methods evolve, their performance gradually improves. When compared to the state-of-the-art YOLO detector, YOLOv11n, the proposed method surpasses its precision, achieving 91.8%, with an mAP@50 of 87.8%. TPGR-YOLO demonstrates significant advantages in balancing detection accuracy and performance, fully proving the superiority of the improved algorithm.

TABLE III. COMPARISON EXPERIMENTS OF THE WIDERPERSON DATASET

Models	P	R	mAP@50	mAP@50:95
YOLOv5s	82.2	66.7	78.9	59.7
YOLOv7-tiny	83.2	71.7	82.7	61.7
YOLOv8n[36]	86.4	76.4	85.3	65.9
Yolov9s	89.4	76.8	86.0	65.6
YOLOv10s	90.7	78.2	86.7	67.8
Yolov11n	91.8	79.3	87.8	68.4
Ours	94.5	81.0	89.2	69.6

The experimental results on the VOC dataset show that the proposed TPGR-YOLO performs the best across all key metrics. The results are shown in Table IV. TPGR-YOLO achieves a precision of 71.9%, recall of 64.8%, mAP@50 of 70.7%, and mAP@50:95 of 54.8%, significantly outperforming other methods.

In comparison, the weaker-performing YOLOv5s achieves only 55.7% precision and 53.5% mAP@50, while YOLOv7-tiny reaches precision and mAP@50 of 59.4% and 57.2%, respectively. As the methods evolve, performance improves. For example, YOLOv10s and YOLOv11n achieve mAP@50 values of 63.1% and 67.9%, respectively. However, TPGR-YOLO still leads across all metrics, particularly in mAP@50:95, where it surpasses YOLOv11n by approximately 3.1 percentage points. These results demonstrate that TPGR-YOLO has a clear advantage in balancing detection accuracy and performance, making it one of the best-performing methods on the VOC dataset.

TABLE IV. COMPARISON EXPERIMENTS OF THE VOC DATASET

Models	P	R	mAP@50	mAP@50:95
YOLOv5s	55.7	53.6	53.5	46.7
YOLOv7-tiny	59.4	55.2	57.2	45.0
YOLOv8n	62.7	56.3	61.8	46.0
Yolov9s	65.4	58.4	62.2	49.7
YOLOv10s	68.2	59.5	63.1	50.5
Yolov11n	71.2	61.7	67.9	51.7
Ours	71.9	64.8	70.7	54.8

The performance comparison between WIoU combined with Slide Loss and the original model loss (Table V) reveals that: The sliding window mechanism in Slide Loss refines the regression process of prediction boxes, significantly enhancing the model's adaptability to small targets and complex scenarios, demonstrating marked advantages over traditional BCE loss in small object detection and regression accuracy. While both WIoU and CIoU improve model performance when combined with Slide Loss, WIoU achieves superior comprehensive performance through its focus on medium-quality anchor boxes and suppression of low-quality samples interference, thereby forming complementary advantages with Slide Loss.

TABLE V. COMPARISON OF DIFFERENT LOSS FUNCTIONS

LOSS	P	R	mAP@50	mAP@50:95
CIoU+BCE	90.1	78.2	88.2	67.4
CIoU+SlideLoss	90.6	78.4	88.7	67.9
WIoU+BCE	90.7	78.3	88.6	68.2
WIoU+SlideLoss	91.4	78.9	89.5	68.5

D. Ablation Experiments

To validate the performance improvement contributed by each module, this study conducts ablation experiments by progressively introducing key modules and evaluating their impact on performance. The results are shown in Table VI.

TABLE VI. ABLATION ANALYSIS OF CTPG DATASET

Modules	Para/M	GFLOPs	Model size	P	R	mAP@50	mAP@50:95
YOLOv11n	2.6	6.3	5.5	90.1	78.2	88.2	67.4
+RFCAConv	2.7	6.9	5.8	90.7	79.4	89.2	68.1
+C3k2RFCA+C2DA	2.7	6.9	5.8	91.2	80.8	90.7	69.5
+C3k2RFCA+C2DA+SAFPN	2.8	7.6	5.8	92.5	81.5	91.4	70.2
+C3k2RFCA+C2DA+SAFPN+WIoU	2.8	7.6	6.0	93.9	82.6	92.8	71.1
TPGR-YOLO	2.8	7.6	6.0	94.4	83.4	93.6	72.7

The baseline method, YOLOv11n, has 2.6 M parameters, 6.3 GFLOPs, and a model size of 5.5 M, achieving a precision of 90.1%, recall of 78.2%, mAP@50 of 88.2%, and mAP@50:95 of 67.4%, with relatively limited performance. Upon introducing the RFCAConv module, the number of parameters slightly increases to 2.7 M, the computational cost rises to 6.9 GFLOPs, and the method's precision improves to 90.7%, recall increases to 79.4%, mAP@50 reaches 89.2%, and mAP@50:95 rises to 68.1%, indicating that RFCAConv effectively enhances feature extraction capabilities.

Further adding the C2DA module leads to continued performance improvement, with precision reaching 91.2%, recall rising to 80.8%, and mAP@50 and mAP@50:95 increasing to 90.7% and 69.5%, respectively, showcasing the significant potential of C2DA's dynamic offset and local feature modeling in object detection.

When combining C2DA and SAFPN modules, both precision and recall further improve to 92.5% and 81.5%, respectively, with mAP@50 reaching 91.4% and mAP@50:95 increasing to 70.2%, demonstrating the significant effect of these modules on multi-scale feature fusion and contextual modeling.

Finally, introducing the WIoU loss function to optimize the method's localization capability results in a precision increase to 93.9%, recall reaching 82.6%, and mAP@50 and mAP@50:95 achieving 92.8% and 71.1%, respectively, highlighting the importance of WIoU for optimizing small object localization strategies.

The complete TPGR-YOLO method, integrating all modules, has 2.8M parameters, 7.6 GFLOPs, and a model size of 6.0 M. Its final precision reaches 94.4%, recall is 83.4%, and mAP@50 and mAP@50:95 are 93.6% and 72.7%, respectively. The experimental results confirm that each module contributes positively to performance improvement, and TPGR-YOLO demonstrates exceptional performance in lightweight design and high-performance object detection.

E. Research Result

The study proposes an improved TPGR-YOLO for traffic police gesture recognition by introducing the RFCAConv module, the C2DA module, the edge-enhanced multi-branch fusion strategy (SAFPN), as well as the WIoU and SlideLoss loss functions. The proposed method demonstrates excellent performance in feature extraction, small object detection, and complex background handling. Experimental results show that TPGR-YOLO achieves high precision, recall, and mAP scores on the self-built CTPG dataset, the WiderPerson dataset, and the VOC dataset, while maintaining real-time inference speed. Furthermore, the ablation study validates the positive contributions of each improved module to the overall performance.

F. Visual Analysis

1) Dataset visualization: The label distribution after training on the CPRG dataset is shown in Fig. 8. From the figure, it can be observed that the "Straight Ahead" category has the largest number of samples, exceeding 1,000, while the "Pull Over" category has the fewest samples. It can also be noted that the labels are primarily distributed in the central region of the images, with most of them having a relatively uniform shape, exhibiting a clear concentration trend.

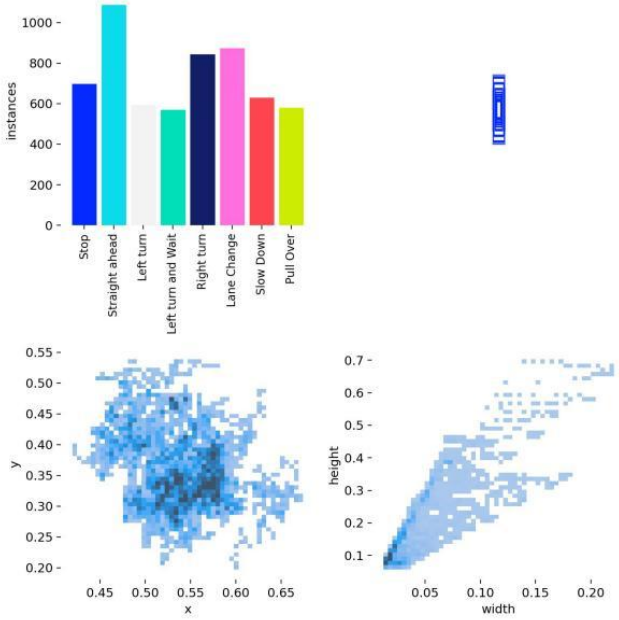


Fig. 8. Dataset visualization.

2) Results visualization: The visualization of the experimental results is shown in Fig. 9. The figure contains 8 rows, corresponding to the categories "Stop," "Straight Ahead," "Left Turn," "Left Turn and Wait," "Right Turn," "Lane Change," "Slow Down," and "Pull Over." The left column displays the original images, the middle column presents the visualization results after training with the original method, and the right column shows the visualization results after applying the improved method.

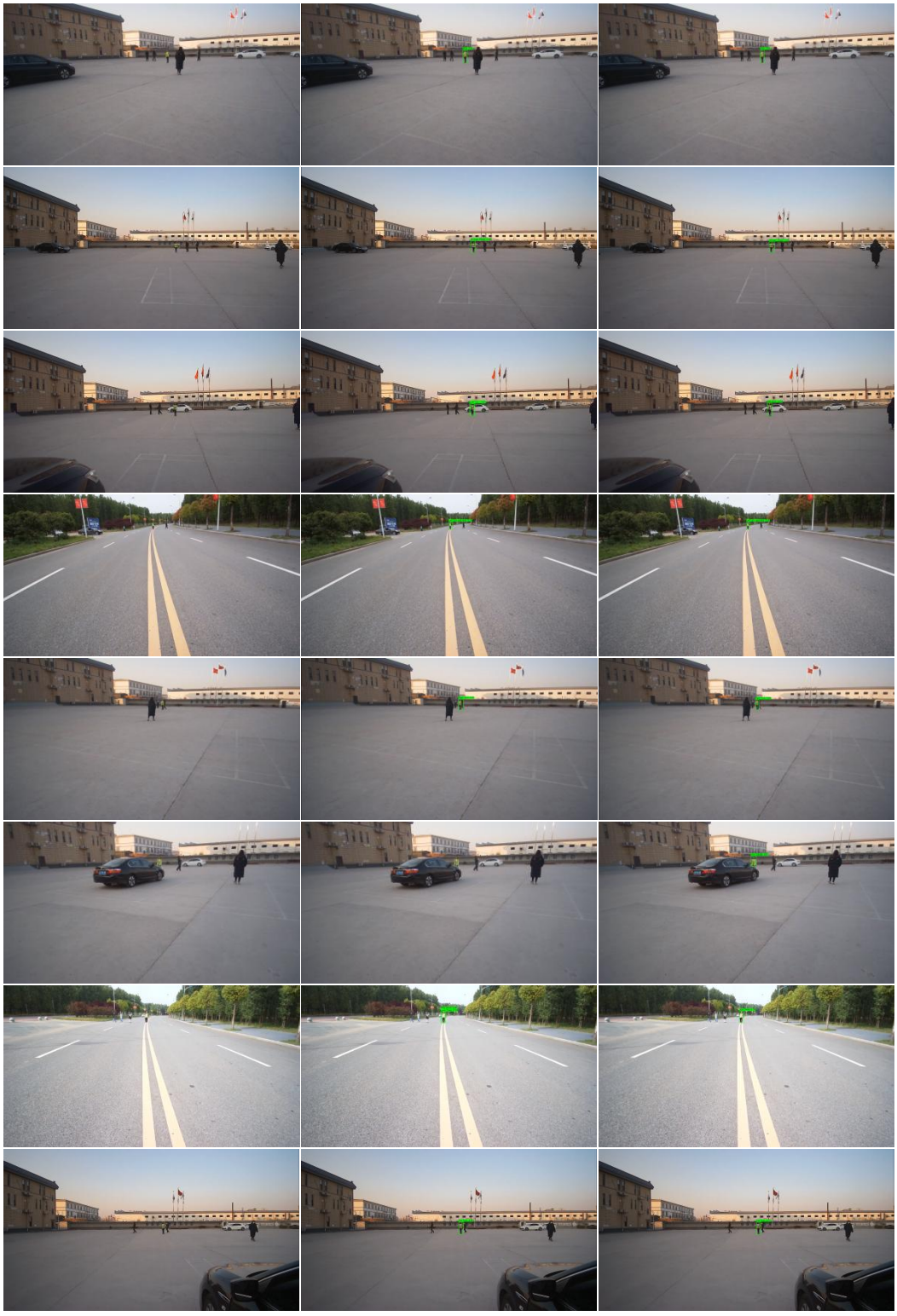


Fig. 9. Result visualization.

From the figure, it can be observed that in the "Line Change" scenario, the original model failed to correctly recognize the traffic officer's gesture due to the vehicle obstructing the lower part of the officer's body, whereas the improved model accurately identified the gesture. In the "Slow Down" and "Left Turn" scenarios, interference from a white vehicle in the background led the original model to incorrectly recognize two actions, while the improved model, by enhancing the recognition of target edges, demonstrated ideal performance in these scenes.

For the "Pull Over" and "Right Turn" gestures, the original method failed to accurately identify the edges of the body and gestures during inference due to background interference, resulting in larger bounding boxes or failure to properly frame the hand. This neglect of detailed information directly affected the inference accuracy. The improved method successfully overcame these issues, showing superior performance.

V. CONCLUSION

In order to address the challenges of small-scale traffic police gestures and complex backgrounds in open traffic scenes, this study improves the backbone network, neck network, and loss function of the YOLOv11 model. The modified algorithm demonstrates strong generalization ability across multiple public datasets. However, since the algorithm relies on single-frame image classification, misjudgments on boundary gestures are inevitable.

Although this research effectively addresses the issues of small-scale objects and complex backgrounds, future work should focus on overcoming the challenges of real-time deployment and exploring the integration of additional sensor data. Furthermore, the model should be improved to accommodate misjudgments of boundary gestures and non-standard traffic police signals. This will enhance the model's applicability and reliability in real-world applications.

ACKNOWLEDGMENT

This work was supported, the National Natural Science Foundation of China (Grant No. 62102033, U24A20331, 62371013, 61931012), the Beijing Natural Science Foundation (No. L247007), the R&D Program of Beijing Municipal Education Commission (Grant No. KZ202211417048), The Project of Construction and Support for high-level Innovative Teams of Beijing Municipal Institutions (Grant No. BPHR20220121), the Academic Research Projects of Beijing Union University (No. ZKZD202302, ZK20202403).

REFERENCES

[1] Mengying Chang, Huizhi Xu, Yuanming Zhang. Low light recognition of traffic police gestures based on lightweight extraction of skeleton features[J], Neurocomputing. 2025, 617: 129042-129042.

[2] Mingde Zheng, Michael S. Crouch, Michael S. Eggleston. Surface Electromyography as a Natural Human-Machine Interface: A Review[J], IEEE sensors journal, 2022, 22(10): 9198-9214.

[3] Muhammad A, Dong S K, Muhammad O, Kang R P, et al. OR-Skip-Net: Outer Residual Skip Network for Skin Segmentation in Non-Ideal Situations[J], Expert Systems with Applications, 2020, 141: 112922-112922.

[4] Hao Z, Dongzhi Z, Bao Z, Dongyue W, Mingcong T, et al. Wearable Pressure Sensor Array with Layer-by-Layer Assembled MXene Nanosheets/Ag Nanoflowers for Motion Monitoring and Human-Machine Interfaces[J], Acs Applied Materials&Interfaces, 2022, 14(43): 48907-48916.

[5] Hedan Bai et al. Stretchable distributed fiber-optic sensors. Science, 2020, 370, 848-852.

[6] LIUMY, HANGCZ, WUXY, et al. Investigation of stretchable strain sensor based on CNT/AgNW applied in smart wearable devices[J]. Nanotechnology, 2022, 33(25): 255501.

[7] LI XT, KOHKH, FARHANM, et al. An ultraflexible polyurethane yarn-based wearable strain sensor with a polydimethylsiloxane infiltrated multilayer sheath for smart textiles[J]. Nanoscale, 2020, 12(6): 4110-4118.

[8] Wenxuan M, Qingtian Z, Ge S, Minghao Z, et al. Design and Implementation of Traffic Police Hand Gesture Recognition System Based on Surface Electromyographic Signals[J], 2022 IEEE 5th Advanced Information Management, Communicates, Electronic and Automation Control Conference(IMCEC), 2022, 5: 1888-1894.

[9] Wenxuan Ma;Ge Song;Qingtian Zeng;Hongxin Zhang;Minghao Zou;Ziqi Zhao. FFCSLT: A Deep Learning Model for Traffic Police Hand Gesture Recognition Using Surface Electromyographic Signals[J]. 2024, 15(8): 13640-13655.

[10] ZHANG Weidong, WANG Zexing, WU Xuangou. WiFi Signal-Based Gesture Recognition Using Federated Parameter-Matched Aggregation [J]. Sensors, 2022, 22(6): 2349.

[11] ZHOU Chengwei, GU Yujie, SHI Zhiguo, et al. Structured Nyquist correlation reconstruction for DOA estimation with sparse arrays[J]. IEEE Transactions on Signal Processing, 2023, 71: 1849-1862.

[12] Feiyi X, Feng X, Jiucheng X, Chi-Man P, Huimin L, Hao G, et al. Action Recognition Framework in Traffic Scene for Autonomous Driving System. [J], IEEE Transactions on Intelligent Transportation Systems, 2022, 23(11): 22301-22311.

[13] Zheng F, Junjie C, Kun J, Sijia W, Junze W, Mengmeng Y, Diange Y, et al. Traffic Police 3D Gesture Recognition Based on Spatial-Temporal Fully Adaptive Graph Convolutional Network[J], IEEE Transactions on Intelligent Transportation Systems, 2023, 24: 9518-9531.

[14] Nan M, Zhixuan W, Yifan F, Cheng W, Yue G, et al. Multi-View Time-Series Hypergraph Neural Network for Action Recognition[J], IEEE Transactions on Image Processing, 2024, 33: 3301-3313.

[15] Xiaofeng G, Qing Z, Yaonan W, Yang M, et al. MG-GCT: A Motion-Guided Graph Convolutional Transformer for Traffic Gesture Recognition[J], IEEE Transactions On Intelligent Transportation Systems, 2024. 25: 14031-14039.

[16] Ziwei Zhang, Bingbing Wu, Yulian Jiang. Gesture Recognition System Based on Improved YOLO V3[J], 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), 2022: 1540-1543.

[17] Shichao W, Chenxia G, Ruifeng Y, Qianchuang Z, Haoyu R, et al. A Lightweight Vision-Based Measurement for Hand Gesture Information Acquisition[J], IEEE sensors journal, 2023, 23(5): 4964-4973.

[18] Saloni S, Anushka P, Prachi J, Ahbaz M, Varsha N, et al. Hand Gesture Recognition Using YOLO Models for Hearing and Speech Impaired People[J], 2022 IEEE Students Conference on Engineering and Systems(SCES), 2022: 1-6.

[19] Maha H, Wesam S, Mohammed Z, Tariq K, et al. Hand Gesture Recognition Based on CNN and YOLO Techniques[J], Decision Science Letters, 2024, 13(4): 977-990.

[20] Yu-Yu Yang, Hsu-Han Yang, Jung-Kuei Yang. YOLO-LRHG: Long Range Hand Gesture Detection Using YOLO with Attention Mechanism[J], 2024 IEEE 7th International Conference on Electronic Information and Communication Technology(ICEICT), 2024: 985-990.

[21] Weina Zhou, Xile Li. PEA-YOLO: a Lightweight Network for Static Gesture Recognition Combining Multiscale and Attention Mechanisms[J], Signal Image and Video Processing, 2024, 18(1): 597-605.

[22] Jichi Liu, Wei Li, Houkun Lyu, Feng Qi. YOLO-based microglia activation state detection[J]. The Journal of Supercomputing. 2024: 24412-24434.

[23] Jian X, Chenglong Z, Jiaqiang M, Zibo C, Ying L, et al. Lightweight Detection Model for Safe Wear at Worksites Using GPD-YOLOv8 Algorithm[J], Scientific Reports, 2025, 15(1): 1-13.

[24] Zhaohui L, Wenshuai H, Wenjing C, Jiaxiu C, et al. The Algorithm for Foggy Weather Target Detection Based on YOLOv5 in Complex Scenes[J], Complex & Intelligent Systems, 2024, 11(1): 1-18.

[25] Renxiang Z, Guangyun Z, Rongting Z, Xiuping J, et al. A Deformable Attention Network for High-Resolution Remote Sensing Images Semantic Segmentation[J], IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-14.

[26] Tengfei Ma, Haitao Yin. MAFPN: a Mixed Local-Global Attention Feature Pyramid Network for Aerial Object Detection[J], Remote Sensing Letters, 2024, 15(9): 907-918.

[27] Xiaoqiang D, Hongchao C, Zenghong M, Wenwu L, Mengxiang W, Zhichao M, et al. DSW-YOLO: A Detection Method for Ground-Planted Strawberry Fruits under Different Occlusion Levels[J], Computers and Electronics in Agriculture, 2023, 108304.

[28] Nan M, Zhixuan W, Yifan F, Cheng W, Yue G, et al. Multi-View Time-Series Hypergraph Neural Network for Action Recognition[J], IEEE Transactions on Image Processing, 2024, 33: 3301-3313.

[29] Lihu P, Jianzhong D, Zhengkui W, Shouxin P, Cunhui Z, et al. HF-YOLO: Advanced Pedestrian Detection Model with Feature Fusion and Imbalance Resolution[J], Neural Processing Letters, 2024, 56(2).

[30] Jinsheng X, Haowen G, Jian Z, Tao Z, Qiuze Y, Yunhua C, Zhongyuan W, et al. Tiny Object Detection with Context Enhancement and Feature Purification[J], Expert Systems with Applications, 2023, 211: 118665.

[31] Kyungseo Min, Gun-Hee Lee, Seong-Whan Lee. Attentional Feature Pyramid Network for Small Object Detection. [J], Neural Networks, 2022, 155: 439-450.

[32] Xudong Dong, Shuai Yan, Chaoqun Duan. A Lightweight Vehicles Detection Network Model Based on YOLOv5[J], Engineering Applications of Artificial Intelligence, 2022, 113: 104914-104914.

[33] Shuai Yuan, Xiangjie Kong, Shuai Zhang. Research on Enhanced YOLOv8 Gesture Recognition Method for Complex Environments*[J], 2024 Wrc Symposium on Advanced Robotics and Automation, Wrc Sara, 2024: 141-146.

[34] Qiu, Xiaoyang; Chen, Yajun; Cai, Wenhao; Niu, Meiqi; Li, Jianying. LD-YOLOv10: A Lightweight Target Detection Algorithm for Drone Scenarios Based on YOLOv10. Electronics; Basel , 2024: 3269.

[35] Huanyu Y, Jun W, Yuming B, Jiacun W, et al. Istd-Detr: A Deep Learning Algorithm Based on Detr and Super-Resolution for Infrared Small Target Detection[J], Neurocomputing, 2025: 129289.

[36] Wong Min On, Nirase Fathima Abubacker. YOLO-Driven Lightweight Mobile Real-Time Pest Detection and Web-Based Monitoring for Sustainable[J]. International Journal of Advanced Computer Science and Applications, 2024, 15: 658-673.

A Hybrid SETO-GBDT Model for Efficient Information Literacy System Evaluation

Jiali Dai[1], Hanifah Jambari[2], Mohd Hizwan Mohd Hisham[3]
College Office, Wenzhou Polytechnic, Wenzhou 325000, Zhejiang, China[1]
Faculty of Social Sciences and Humanities, Universiti Teknologi Malaysia, Johor, 81310, Malaysia[1, 2, 3]

***Abstract*—Information literacy (IL) is essential for vocational education talents to thrive in the modern information age. Traditional assessment methods often lack quantitative precision and systematic evaluation models, making it difficult to accurately measure IL levels. This paper aims to develop a robust, data-driven model to assess information literacy in vocational education talents. The goal is to improve the accuracy and efficiency of IL evaluations by combining machine learning techniques with optimization algorithms. The proposed method integrates the Stock Exchange Trading Optimization (SETO) algorithm with the Gradient Boosting Decision Tree (GBDT) to construct the SETO-GBDT model. This model optimizes parameters such as the number of decision trees and tree depth. A comprehensive evaluation index system for IL is built, focusing on learning attitude, process, effect, and practice. The SETO-GBDT model was trained and tested using real-world data on IL indicators. The SETO-GBDT model outperformed traditional models such as Decision Tree, Random Forest, and GBDT optimized by other algorithms like SCA and SELO. Specifically, it achieved an RMSE of 0.13, an R^2 of 0.98, and reduced evaluation time to 0.092 s, demonstrating superior accuracy and efficiency. The research concludes that the SETO-GBDT model offers a significant improvement in evaluating IL for vocational education talents. The model's high accuracy and reduced evaluation time make it an effective tool for assessing and enhancing information literacy, aligning with the educational goals of developing well-rounded, information-savvy professionals.**

***Keywords*—Vocational education; talent; information literacy; system building; educational evaluation; gradient augmentation; decision tree**

I. Introduction

China has implemented various programs to enforce the systematic advancement of educational evaluation reform. The primary focus of the current curriculum reform is to develop scientific core literacy. The goal is to transform the curriculum to enhance students' comprehensive ability to use information technology to solve problems. This will enable students to become well-rounded individuals with high-quality technical skills and moral, intellectual, physical, social and aesthetic development [1]. Educational talent assessment plays a crucial role in talent education. It involves designing an evaluation index system and educational activities to determine the reasonableness of the educational process, appropriateness of the educational methods used, and whether the expected educational outcomes are achieved [2]. In vocational education, the rapid advancement of information technology has led to the constant flow of information. In this vast and complex information landscape, the ability to distinguish, summarize, and synthesize information has become an increasingly important challenge [3]. For college students, the skill of efficiently and accurately finding the information they need within a limited time frame has become an essential foundational skill. The enhancement of information literacy among vocational education talents involves developing their vocational information literacy skills, including vocational skills, information skills, comprehensive skills, and other vocational information literacy skills. This enables them to acquire the necessary knowledge and adapt to the rapid development of the information society [4].

The evaluation of vocational education talent information literacy is a crucial tool to enhance the information literacy of vocational education talent. Employing appropriate assessment methods may facilitate students' growth and ensure that their information literacy meets the intended objectives [5]. The present research on talent information literacy focuses mostly on three areas: defining information literacy, constructing talent information literacy systems, and assessing talent information literacy [6]. Fei and Erjun [7] studied the eight aspects of information literacy, i.e., skillful use of information tools, correct access to information, proper handling of information, timely generation of information, good at creating information, maximizing the benefits of information, strengthening information collaboration, enhancing information immunity, etc.; Esfandiari and Arefian [8] investigated four information literacy evaluation indexes for undergraduate-type students, i.e., information awareness, information competence, information evaluation, and information ethics; Riithi and Kimani [9] analyzed the academic attention to information literacy education and found that 2012 to 2014 was a period of rapid growth in information literacy research; Ganesan and Gunasekaran [10] elaborated on the content of information literacy education in applied schools and gave strategies and methods for cultivating students' information literacy; Vianna and Caregnato [11] used hierarchical analysis methods to construct a system of information literacy for talents and studied the corresponding cultivation methods; Faber [12] used a simple decision tree to fit the nonlinear mapping relationship between talent information literacy indicators and assessment values. The examination of the present research literature on information literacy reveals the existence of the following issues: a) The present assessment of talent information literacy is still in its early developmental stage, focusing mostly on qualitative analysis of the meaning of information literacy and the importance of the study, while lacking quantitative analysis [13]; b) The present assessment methods are not suitable for constructing the information literacy assessment model due to

the large dimensionality of the information literacy evaluation index system. The selection of the talent information literacy evaluation index is insufficiently thorough and lacks a systematic approach [14].

The advancement of integrated learning technology has led to the adoption of efficient integrated decision trees to improve assessment models. This has become a prominent area of future development and research. However, the performance of the integrated learning algorithm is limited by the parameter settings. Therefore, optimizing the algorithm through hyper-parameter search has emerged as a method to enhance the assessment algorithm [15]. This paper presents a method for assessing the information literacy of vocational education talents. The method combines literature analysis and the development of intelligent algorithms, specifically using an intelligent optimization algorithm-integrated decision tree framework. This paper presents a talent information literacy assessment model for vocational education talents by analyzing the problem of constructing and evaluating an information literacy system. The model is designed using a framework that combines the gradient enhancement decision tree and the stock market trading optimization algorithm. It is then applied to the development of information literacy in vocational education talents. By doing rigorous experimental research, this study confirms that the proposed strategy is indeed superior.

II. Information Literacy System

A. Analysis of the Process

The purpose of developing an information literacy system for vocational education skills is to enhance students' overall abilities in the field of information technology. This includes fostering their awareness of information, understanding of information, competence in handling information, and adherence to ethical standards related to information (Fig. 1). Developing an information literacy system for vocational education can enhance students' capacity to adapt to the demands of work and life in the digital age, while also improving their vocational competitiveness and lifelong learning skills [16].

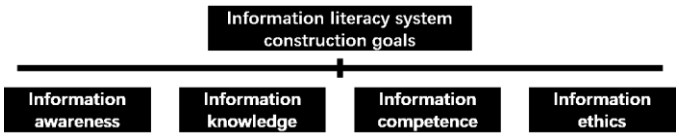


Fig. 1. Objectives of information literacy system construction for vocational education talents.

The process of constructing a vocational education talent literacy system involves various components such as curriculum design, teaching methods, teaching resources, teacher training, assessment, and feedback [17], as depicted in Fig. 2. This process is driven by the goal and importance of developing vocational education talent literacy.

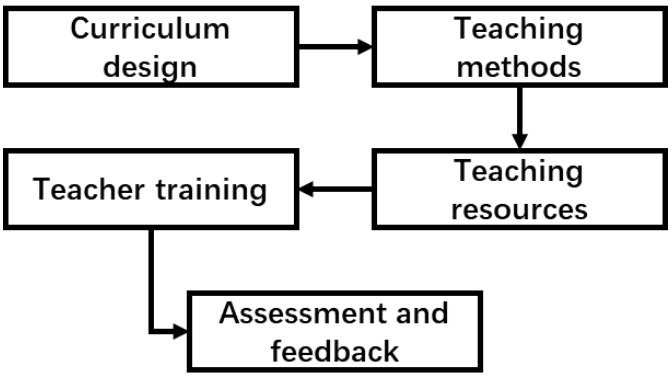


Fig. 2. Construction process of information literacy system for vocational education talents.

B. Constructive Thinking Refers to the Process

The construction process of the vocational education talent information literacy system involves extracting talent literacy assessment indexes from four aspects: learning attitude, learning process, learning effect, and final practice. These indexes serve as the first-level indexes. Through the refinement of the evaluation process, second-level indexes are extracted to construct the vocational education talent literacy system. Fig. 3 illustrates the concept of this construction process.

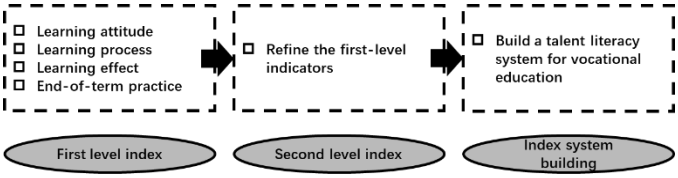


Fig. 3. Constructing information literacy system for vocational education talents.

C. Construction of Talent Information Literacy System

This paper adheres to the principles of objectivity, comprehensiveness, focus, and practicality in constructing an information literacy system for vocational education talents (Fig. 4). It selects relevant indicators from four aspects: information literacy learning attitude, learning process, learning effect, and final practice, as illustrated in Fig. 4. The indicators for learning attitude encompass class attendance and the number of assignments submitted. The indicators for learning process involve the organization of information literacy materials, the design of information literacy papers, information retrieval, information management, and display presentation production. The indicators for learning effects consist of teacher information literacy training and student information feedback. The final practice indicators encompass information literacy application and examination results [14].

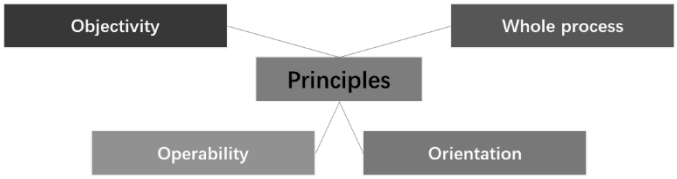


Fig. 4. Principles for the selection of evaluation system indicators.

Diagram depicting the creation of the evaluation system, as seen in Table I.

TABLE I. EVALUATION OF SYSTEM CONSTRUCTION

No.	First level	Var.	Second level	Var.
1	Learning attitude	A	Class attendance	A1
			Number of assignments turned in	A2
2	Learning process	B	Material arrangement	B1
			Information literacy paper design	B2
			Information retrieval	B3
			Information management	B4
			Display and presentation production	B5
3	Learning effect	C	Literacy training for teachers	C1
			Feedback on students' information	C2
4	End-of-term practice	D	Information literacy application	D1
			Test scores	D2

III. EVALUATING THE INFORMATION LITERACY SKILLS

A. Talent Information Literacy Assessment Framework

The method for assessing the information literacy of vocational education talents involves using the information literacy assessment index as input for the evaluation model. The output is a comprehensive assessment score. The assessment model is constructed using the integrated learning method based on the gradient boosting decision tree algorithm. The hyperparameters of the gradient boosting decision tree are optimized using a stock market trading optimization algorithm. The specific framework is illustrated in Fig. 5.

The talent information literacy evaluation model construction framework analysis reveals that the research on constructing and assessing the talent information literacy system in vocational education is a crucial technology. This research utilizes an enhanced integrated learning algorithm to establish a mapping relationship between the assessment index value of information literacy and the assessment scores, as depicted in Fig. 6. This research utilizes gradient boosting decision tree to establish the mapping connection and enhances the assessment accuracy of the talent information literacy evaluation model by optimizing the hyperparameters of the GBDT technique using a stock market trading optimization algorithm.

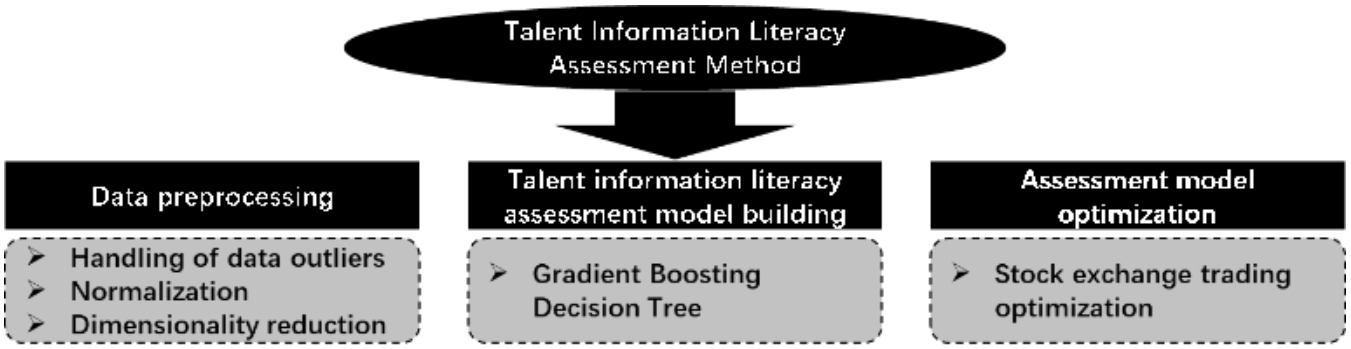


Fig. 5. Talent information literacy evaluation model construction framework.

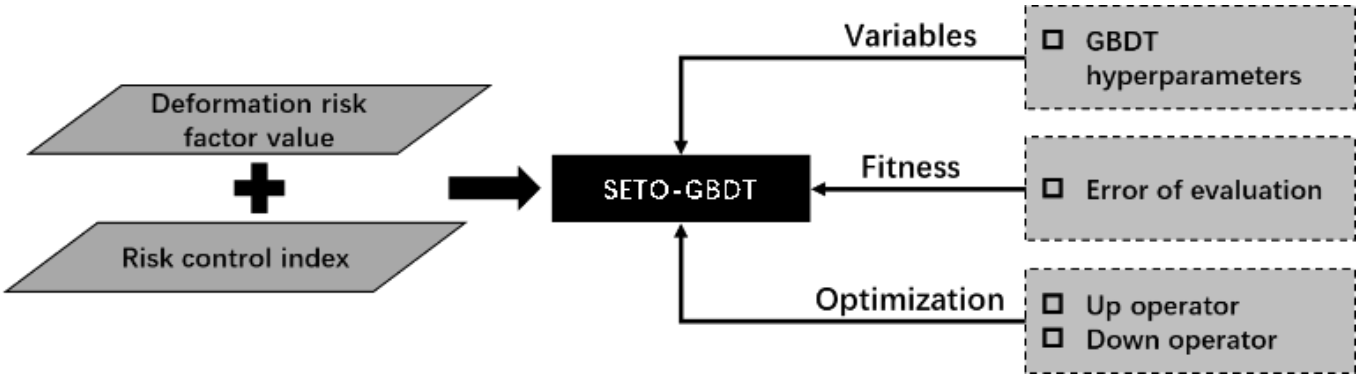


Fig. 6. Information literacy assessment method for vocational education talents based on the improved integrated learning model.

B. Gradient Boosting Decision Tree

1) The theory of gradient enhancing decision trees: The Gradient Boosting Decision Tree (GBDT) [18] is a popular machine learning approach that is commonly employed for solving regression and classification issues. It enhances the precision of predictions by repeatedly creating several decision trees and decreasing the loss function via gradient descent. The primary objective of each new decision tree is to rectify the residuals of the preceding tree, which refers to the disparity

between the actual value and the current model's prediction. The fundamental concept behind Gradient Boosting Decision Trees (GBDT) is to construct a robust prediction model by amalgamating numerous feeble learners, typically decision trees. The structure of this model is illustrated in Fig. 7.

In this paper, the GBDT algorithm is chosen to build an information literacy assessment model for vocational education talents in the form of regression tree. The specific form of decision tree is as follows:

$$T\left(x;c,R\right)=\sum_{v=1}^{M'}c_v I\left(x\in Rv\right) \tag{1}$$

$$I=\begin{cases}0 & x\notin R_v\\ n & x\in R_v\end{cases} \tag{2}$$

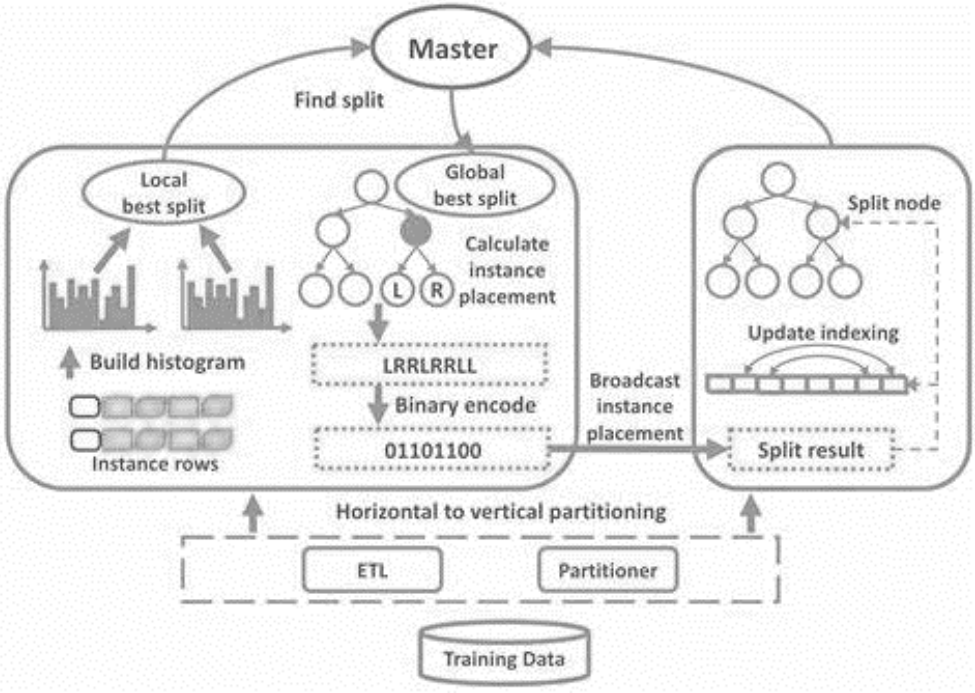


Fig. 7. GBDT structure.

Where, $R=\left\{R_1,R_2,\cdots,R_M^t\right\}$ is the decision tree leaf; M^t is the number of leaf nodes; I is the difference function; $c=\left\{c_1,c_2,\cdots,c_{Mt}\right\}$, $c_v=mean\left(y\middle|x\in R_v\right)$ represent the output characteristic mean of the samples in the leaf space. The GBDT model is a combination of multiple decision trees with the following structure:

$$F\left(x\right)=\sum_{ix=1}^{N'}T_{ix}\left(x;c,R\right) \tag{3}$$

where T_{ix} represents the ixth tree and $ix=1,2,\cdots,N^t$.

During each iteration, a decision tree T_m is added to the decision model based on the previous iteration number in the following form:

$$T_m=\arg\min_T\sum_{k=1}^{n^t}L\left(y_k,\sum_{k=1}^{m-1}T_j\left(x_k\right)+T\left(x_k\right)\right) \tag{4}$$

Where $L\left(\cdot\right)$ denotes the loss function, j is the number of iterations, k is the number of samples in the training set, x_k and y_k denote the training samples. The loss function is set to the least squares function, and the m^{th} tree is built on the residuals of the sum of the decision trees in the previous iteration, and the GBDT model is constructed as follows after m iterations:

$$F_m\left(x\right)=F_{m-1}\left(x\right)+\arg\min_T\sum_{k=1}^{n^t}L\left(y_k,F_{m-1}\left(x_k\right)+T\left(x_k\right)\right) \tag{5}$$

The minima of the GBDT model are calculated by the gradient descent method and the negative gradient direction of the loss function at the current F_{m-1} is set to be the direction of the maximum descent gradient:

$$F_m\left(x\right)=F_{m-1}\left(x\right)+r_m\sum_{k=1}^{n^t}\nabla F_{m-1}L\left(y_k,F_{m-1}\left(x_k\right)\right) \tag{6}$$

$$r_m = \arg\min_r \sum_{k=1}^{n^t} L\left(y_k, F_{m-1}(x_k) - r\frac{\partial L\left(y_k, F_{m-1}(x_k)\right)}{\partial F_{m-1}(x_k)}\right) \quad (7)$$

In order to avoid the fitting phenomenon of GBDT model, the model learning rate was used to determine the model:

$$F_m(x) = F_{m-1}(x) + \nu r_m T_m(x) \quad (8)$$

where ν is the learning rate of the GBDT model.

2) GBDT process steps: The GBDT principle involves a sequential process outlined in a flowchart (Table II). The steps are as follows: 1) initialize the model, 2) calculate the residuals, 3) build the decision tree, 4) update the model, and 5) output the model until a predetermined number of iterations is reached or the stopping condition is met [19].

TABLE II. GBDT FLOW CHART

Algorithm 1: GBDT Model
1. Initialize model.
2. Calculate residuals.
3. Construct a decision tree.
4. Update the model.

3) GBDT Advantage: The Gradient Boosting Decision Tree (GBDT) has several advantages: 1) It exhibits a high level of precision in its prediction capacity; 2) It is capable of directly handling various sorts of data; 3) It demonstrates resilience and generalization ability; 4) It efficiently handles large-scale data; and 5) It can effectively train unbalanced data, as seen in the accompanying Fig. 8.

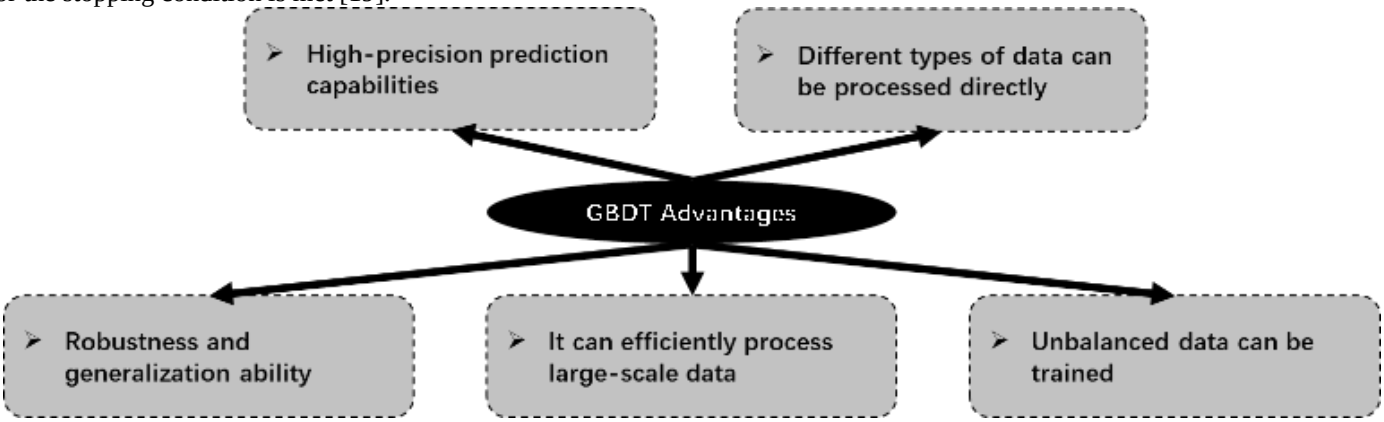


Fig. 8. GBDT advantages

4) Applications of gradient boosting decision trees: Gradient Boosting Decision Trees (GBDT) demonstrates exceptional performance in several practical applications (Fig. 9), such as financial risk management, stock market forecasting, medical diagnostics, and natural language processing [20]. Due to its strength and versatility, this tool is highly used by data scientists and machine learning engineers for addressing intricate prediction issues (Fig. 10) [19].

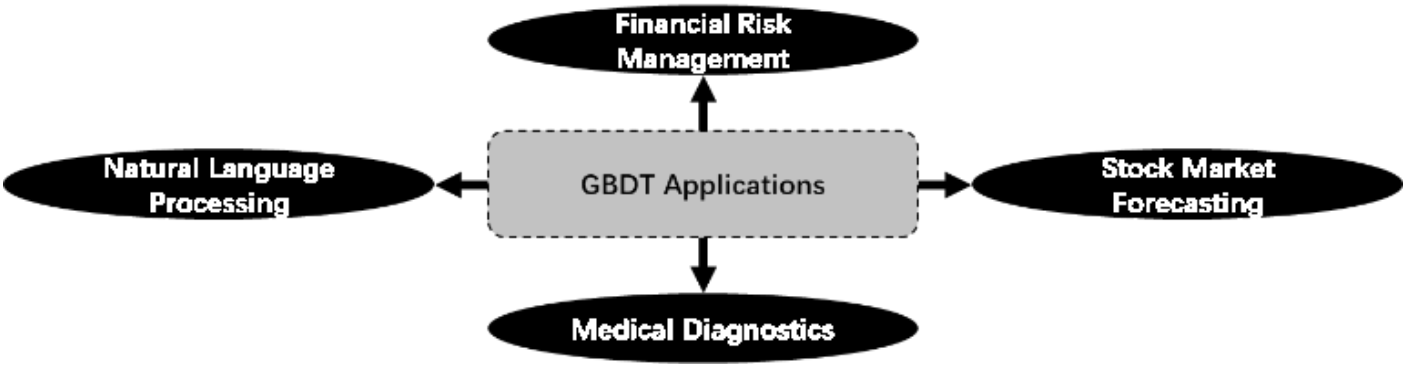


Fig. 9. GBDT application.

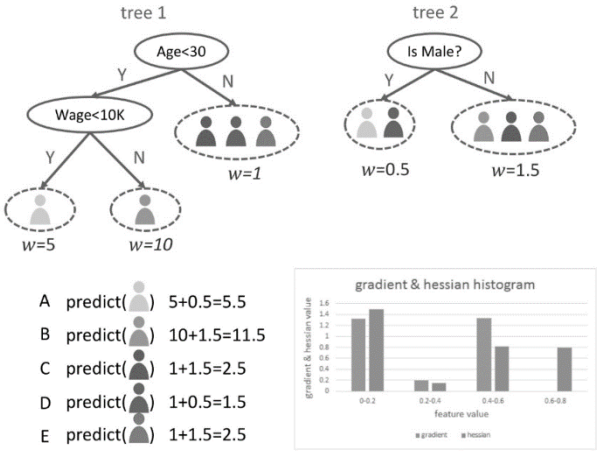


Fig. 10. GBDT problem solving approach.

C. The SETO Optimized Gradient Boosting Decision Tree (GBDT) Model

1) SETO algorithm: A swarm intelligence optimization algorithm known as Stock Exchange Trading Optimization (SETO) [21] takes its cues from the ever-changing stock market and its trading patterns to determine which stock is most likely to maximize profit. Here, each stock is seen as a possible solution to the problem. The method repeatedly optimizes operators such as rise, fall, and exchange operations. Ultimately, the share that yields the highest profit is identified as the ideal answer.

a) Initialization of the population

$$s_{ij} = l_{ij} + \phi_{ij} \cdot \left(u_{ij} - l_{ij}\right) \tag{9}$$

ϕ_{ij} represents a random integer in the interval [0, 1], u_{ij} and l_{ij} signify the upper and lower bounds of the search space, respectively, and s_{ij} symbolizes the j^{th} dimension associated with the i^{th} person.

The following is the equation for calculating the earnings value of a stock, which is used to analyze individual stocks:

$$f_i = f\left(S_i\right) = f\left\{s_{i1}, s_{i2}, \cdots, s_{iD}\right\} \tag{10}$$

where f_i denotes the fitness value of the i^{th} stock individual S_i.

There are specific numbers of sellers and buyers for every stock in the stock market. The first traders are defined using a random initialization process. Here nf_i is how to compute the normalized fitness value in order to accomplish this mechanism:

$$nf_i = \frac{f_i - \min\left(M\right)}{\sum_{k=1}^{N}\left(f_k - \min\left(M\right)\right)}, M = \left\{f_k \middle| k = 1, 2, \cdots, N\right\} \tag{11}$$

S_i The number of traders is calculated as follows:

$$T_i = \left[nf_i \times T\right] \tag{12}$$

T represents the total number of traders, whereas T_i denotes the trading volume of stock S_i. Here S_i is how to figure out how many people are buying and selling stocks:

$$b_i = \left[r \times T_i\right] \tag{13}$$

$$s_i = T_i - b_i \tag{14}$$

b_i and s_i represent the quantities of buyers and sellers, respectively, whereas r is a random variable uniformly distributed between $\left[0,1\right]$.

b) Ascent operation operator: The upward action mostly emulates the appreciation of the stock price. At this juncture, the stock may ascend to a greater valuation, and the peak price can attain the ideal threshold. The equation for replicating the upward action operator is articulated as follows:

$$S_i\left(t+1\right) = S_i\left(t\right) + R \times \left(S^g\left(t\right) - S_i\left(t\right)\right) \tag{15}$$

$S_i\left(t\right)$ represents the i^{th} stock person for the t^{th} iteration, $S^g\left(t\right)$ signifies a D-dimensional random vector, and $r_j \in R$ indicates the ideal solution for the t^{th} iteration.

The parameter R enhances the degree of random variation to assist individuals in evading local optima and exploring broader geographical areas, while $r_j \in R$ is specified as follows:

$$r_j = U\left(0, pc_i \times d_1\right) \tag{16}$$

U produces evenly dispersed random numbers within the range of $\left[0, pc_i \times d_1\right]$. pc_i is the bid-ask ratio of the stock S_i, and d_1 denotes the normalized distance between $S_i\left(t\right)$ and $S^g\left(t\right)$:

$$d_1 = \frac{\sqrt{\sum_{j=1}^{D}\left(S_j^g\left(t\right) - S_{ij}\left(t\right)\right)^2}}{ub - lb} \tag{17}$$

ub and lb represent the upper and lower limits of the search space, respectively. Typically, an increase in demand for a stock correlates with an appreciation in its value. The parameter pc_i mimics the effect of stock growth demand and delineates stock demand based on the overall number of purchasers, calculated as follows:

$$pc_i = \frac{b_i}{s_i + 1} \tag{18}$$

In order to avoid pc_i crossing the boundary, the parameter pc_i is limited to the range $[0, 2]$, which is calculated as follows:

$$pc_i = \min\left(\frac{b_i}{s_i + 1}, 2\right) \tag{19}$$

In the ascending phase, stock demand escalates, resulting in a rise in buyers and a reduction in sellers:

$$b_i = b_i + 1 \tag{20}$$

$$s_i = s_i - 1 \tag{21}$$

c) Falling operation operator: Most of the time, the decline operation operator will mimic a falling stock price using the following equation:

$$S_i(t+1) = S_i(t) - W \times \left(S_i^l(t) - S_i(t)\right) \tag{22}$$

$S_i^l(t)$ represents the current local optimal solution of the ith stock, W signifies a D-dimensional random vector, and $w_j \in W$ is defined as follows:

$$w_j = U\left(0, nc_i \times d_2\right) \tag{23}$$

Where U generates uniformly distributed random numbers in the range $[0, nc_i \times d_2]$. nc_i is the sell-buy ratio of the stock S_i , and d_2 denotes the normalized distance between $S_i(t)$ and $S_i^l(t)$:

$$d_2 = \frac{\sqrt{\sum_{j=1}^{D}\left(S_{ij}^l(t) - S_{ij}(t)\right)^2}}{ub - lb} \tag{24}$$

$$nc_i = \min\left(\frac{s_i}{b_i + 1}, 2\right) \tag{25}$$

In the decline phase, the supply of stocks escalates. In each repetition, the number of vendors rises while the number of customers diminishes during the decline phase:

$$s_i = s_i + 1 \tag{26}$$

$$b_i = b_i - 1 \tag{27}$$

d) Exchange operation operator: During the trading phase, the trader employs the most lucrative stock to substitute the least expensive stock. During this phase, the trader divests from the least performing stock and acquires the highest performing stock. This operational method enables traders to attract stocks. The least favorable stock is obtained as follows:

$$S_{worst} = S_w \; where \quad f(S_w) < f(S_j)$$
$$\forall j = 1, 2, \cdots, N, w \neq j \tag{28}$$

Subsequently, the least favorable stock queue eliminates one seller and incorporates it into the most favorable stock queue and the ideal stock. S_{best} definition is derived as follows:

$$S_{best} = S_b \; where \quad f(S_b) < f(S_j)$$
$$\forall j = 1, 2, \cdots, N, b \neq j \tag{29}$$

The exchange operation operator augments the population size. The operation reduces the quantity of suppliers while augmenting the quantity of purchasers. Consequently, the buyer-seller ratio escalates, thereby enhancing the probability of the stock appreciating.

e) RSI calculation: We use the RSI indicator to recognize when a stock is rising or falling. As the RSI value increases, SETO behaves up or down modeled as follows:

$$\begin{cases} \text{rising} & RSI \leq 30 \\ \text{falling} & RSI \geq 70 \\ p \times \text{rising} + (1-p) \times \text{falling} & 30 < RSI < 70 \end{cases} \tag{30}$$

When p represents a binary random variable, and $p \in \{0, 1\}$ is calculated as follows:

$$p = \begin{cases} 1 & rand \geq 0.5 \\ 0 & else \end{cases} \tag{31}$$

$rand$ represents a random number inside the interval [0, 1]. The RSI for the ith stock is computed as follows:

$$RSI = 100 - \frac{100}{1 + RS} \tag{32}$$

Relative intensities were computed using the simple moving average method:

$$RS = \frac{\sum_{i=1}^{K} P_i}{\sum_{i=1}^{K} N_i} \tag{33}$$

P_i and N_i represent upward and downward price fluctuations, respectively. K represents the RSI trading time frame. The equations for P_i and N_i are as follows:

$$P_i = \begin{cases} 1 & if\left(f_i(t) - f_i(t-1)\right) > 0 \\ 0 & otherwise \end{cases} \tag{34}$$

$$N_i = \begin{cases} 1 & if\left(f_i(t-1) - f_i(t)\right) > 0 \\ 0 & otherwise \end{cases} \tag{35}$$

where, $f_i(t)$ and $f_i(t-1)$ denote the fitness values for the current versus previous iteration counts, respectively.

f) Algorithmic steps: The SETO algorithm's location updating approach is shown in the pseudo-code included in Table III.

TABLE III. SETO ALGORITHM PSEUDO-CODE

Algorithm 2: SETO Algorithm
1. Initialize AOA parameters;
2. Initialize population of shares;
3. Evaluate initial population and update best share with best value;
4. While t <= tmax do
5. For each share do
6. If RSI <= 30
7. Carry out rising operator;
8. Elseif RSI >= 70
9. Carry out falling operator;
10. Else
11. Carry out rising and falling phase;
12. End
13. Carry out exchange phase;
14. Calculate RSI;
15. End
16. Evaluate object and update best object;
17. t = t + 1;
18. End
19. Output best solution.

2) SETO-GBDT: This paper utilizes the SETO algorithm to optimize the parameters of the GBDT model [22]. These parameters include the number of decision trees (Para1), the maximum depth of the tree (Para2), the minimum number of samples required for internal nodes (Para3), the minimum number of samples required for leaf nodes (Para4), and the optimal number of segmented features (Para5). The optimization process aims to minimize the regression error of the GBDT model, as demonstrated in Table IV.

D. Application of SETO-GBDT Model in the Assessment of Information Literacy of Vocational Education Talents

This research applies the SETO algorithm optimization GBDT model to design a vocational education talent information literacy assessment model, aiming to tackle the problem of vocational education talent information literacy assessment. The information literacy assessment method for vocational education talents, based on the SETO-GBDT model, consists of two main components: the development of an information literacy assessment index system for vocational education talents, and the creation of an information literacy assessment model for vocational education talents. The specific steps for implementing this method are illustrated in Fig. 11.

Fig. 11 illustrates the process of constructing an information literacy assessment index system for vocational education talents. The first part involves analyzing the development of information literacy in vocational education talents and using this analysis to design the information literacy assessment index system. The second part focuses on standardizing the data of the information literacy assessment index for vocational education talents, with the index value serving as input and the information literacy assessment value as output. The literacy assessment indicator data is standardized using the SETO algorithm to optimize the GBDT parameters. This algorithm is used to train the mapping relationship between the indicator value and the assessment value of vocational education talents' information literacy assessment.

TABLE IV. PSEUDO-CODE OF SETO-GBDT ALGORITHM

Algorithm 3: GBDT based on SETO algorithm
1. Determine optimized variables, including Para 1–5;
2. Set SETO algorithm parameters;
3. Initialize stock population;
4. Calculate fitness of stock using Error, and update best stock;
5. While t <= tmax
6. Calculate RSI value;
7. If RSI <= 30
8. Carry out rising operator;
9. Else if RSI >= 70
10. Carry out falling operator;
11. Else Carry out rising and falling phase;
12. End
13. Carry out exchange phase;
14. End
15. Output best parameters of GBDT model;
16. Build SETO-GBDT model.

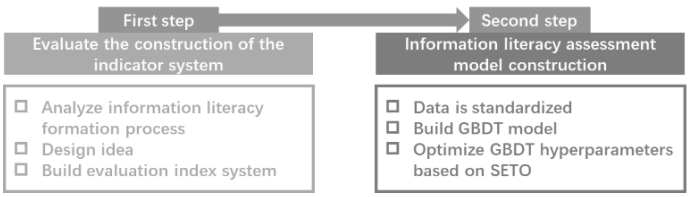


Fig. 11. Step-by-step diagram of the information literacy assessment model for vocational education talents combined with SETO-GBDT.

IV. Data Analysis

This paper aims to assess the effectiveness of the SETO-GBDT model in evaluating the information literacy of vocational education talents. We utilize a dataset consisting of information literacy assessment indexes of vocational education talents and compare and analyze the performance of the SETO-GBDT model with the GBDT model optimized by the SCA [23], SELO [24], HBO [25], and LFD [26] algorithms.

A. Environment, Data, and Algorithm Settings

The SETO-GBDT model is used to assess the information literacy of vocational education talents through a simulation experiment in the Windows 10 environment. The visualization software used is Matlab 2022a, the method programming software is Python 3.8, and the fundamental algorithm is implemented in C++.

The data set of indicators for assessing information literacy in vocational education was gathered by methods such as literature data analysis, case study analysis, comparison analysis, and questionnaire survey (Fig. 12). The data from the study subjects were randomly split into three groups: 70% for training, 15% for testing, and 15% for validation. The model's average evaluation indexes were then calculated using the ten-fold cross-validation method.

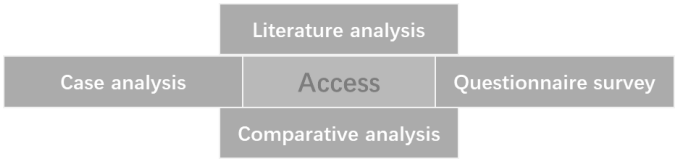


Fig. 12. Data access.

The information literacy evaluation algorithm for vocational education abilities based on the SETO-GBDT model utilizes several comparison algorithms, such as decision tree, RF, AdaBoost, GBDT, and GBDT algorithms optimized by SCA, SELO, HBO, and LFD. The particular parameter values may be found in Table V and Table VI. The algorithms SCA, SELO, HBO, and LFD each have 100 populations and a maximum iteration number of 1000.

TABLE V. Parameter Settings of the Contrast Evaluation Algorithm

No.	Algorithms	Parameter settings
1	Decision tree	Maximum number of splits is 4
2	Random forest	N_tree=500, m_try=floor(80.5)
3	AdaBoost	Regressors number is 15, Iteration is 50
4	GBDT	Decision tree number is 48, maximum depth of tree is 10, samples for internal nodes is 18, samples required for leaf nodes is 1, optimal segmentation features is 9.

TABLE VI. Comparison Optimization Algorithm Parameter Settings

No.	Algorithms	Parameter settings
1	SCA	a=2, r1=1-2t/G, r2=[0,2π], r3=[0,2], r4=[0,1]
2	SELO	P=2, O=3, rp=0.999, rk=0.1, prob=0.999
3	HBO	C=G/25, p1=1-t/G, p2=p1+(1-p1)/2
4	LFD	Threshold=2, CSV=0.5, β=1.5
5	SETO	T=100

B. GBDT Model Optimization Results

The average index value of each algorithm is statistically obtained through the ten-fold cross-validation method, and the specific results are shown in Table VII, Table VIII and Fig. 13.

TABLE VII. Optimization Results of different Optimization Algorithms to Optimize the GBDT Model

No.	Algorithms	Opti. value	Opti. time	Iter. num
1	SCA-GBDT	2.668	3.72	1000
2	SELO-GBDT	1.460	3.33	968
3	HBO-GBDT	0.510	3.45	755
4	LFD-GBDT	0.367	3.10	631
5	SETO-GBDT	0.125	2.71	400

TABLE VIII. Results of Different Optimization Algorithms to Optimize the Parameters of GBDT Model

No.	Algorithms	Para1	Para2	Para3	Para4	Para5
1	SCA-GBDT	30	10	10	2	12
2	SELO-GBDT	41	9	21	3	7
3	HBO-GBDT	49	10	10	3	12
4	LFD-GBDT	44	7	19	5	10
5	SETO-GBDT	50	10	25	5	8

Table VIII presents a comparison of the optimization accuracy, optimization time, and convergence number results of different optimization algorithms for optimizing the GBDT model. From Table VIII, it can be seen that in terms of optimization accuracy, the GBDT evaluation model based on SETO algorithm has the highest accuracy of 0.125, followed by LFD-GBDT, HBO-GBDT, SELO-GBDT, and SCA-GBDT models; in terms of optimization time, the SETO-GBDT model has the lowest evaluation time of 2.71 s; in terms of the number of times of convergence, the SETO-GBDT model converged to the optimal value in 400 times.

The results of optimizing GBDT model parameters with different optimization algorithms are given in Fig. 13. The results of optimizing GBDT model parameters based on SETO algorithm: number of decision trees Para1=50, maximum depth of tree Para2=10, minimum number of samples required for internal nodes Para3=25, minimum number of samples required for leaf nodes Para4=5, and optimum number of segmentation features Para5=8.

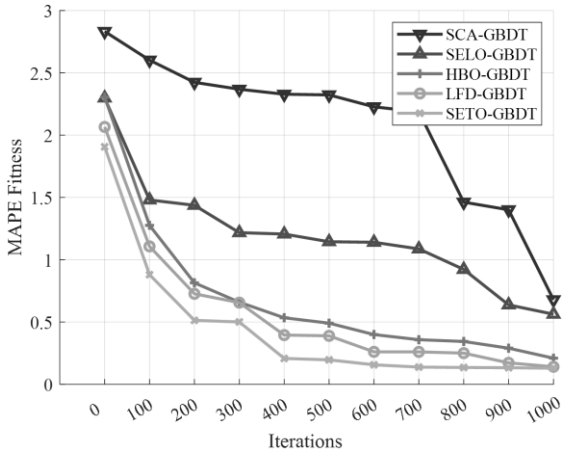


Fig. 13. Convergence curve of GBDT model optimized by different optimization algorithms.

Fig. 13 gives the convergence curve of GBDT model optimized by different optimization algorithms. In Fig. 13, it can be seen that the convergence curve of the optimized GBDT model based on SETO algorithm converges to 0.125 at the 400th iteration.

C. Evaluation of Model Comparison Results

In order to avoid unexpected results of the experiment, 10 independent tests were conducted and the RMSE, R2, training time, and evaluation time averages of the decision tree, RF, AdaBoost, GBDT, and SETO-GBDT algorithms were counted, as shown in Table IX.

TABLE IX. STATISTICS AND COMPARISON OF THE RESULTS OF THE ASSESSMENT INDICATORS FOR THE CONTRASTING ASSESSMENT ALGORITHMS

No.	Evaluation models	RMSE	R^2	Training T/s	Evaluation T/s
1	Decision tree	1.37	0.76	2.30	0.189
2	RF	0.88	0.86	3.94	0.166
3	AdaBoost	0.36	0.93	3.25	0.132
4	GBDT	0.31	0.96	4.45	0.104
5	SETO-GBDT	0.13	0.98	2.71	0.092

Table IX presents the statistics and comparisons of the evaluation index results for the decision tree, RF, AdaBoost, GBDT, and SETO-GBDT algorithms. The SETO-GBDT algorithm performs the best in terms of RMSE, achieving a value of 0.13. It also outperforms the other algorithms in terms of R2, with a value of 0.98. The decision tree algorithm has the shortest training time, taking only 2.30 seconds. On the other hand, the SETO-GBDT algorithm has the shortest evaluation time, which is 0.092 seconds.

V. CONCLUSION

This paper addresses the issue of assessing information literacy in vocational education. It proposes a model for assessing information literacy in vocational education based on SETO-GBDT. The model is validated through literature analysis, case study analysis, comparative analysis, and questionnaire survey. The findings are as follows:

- The SETO method enhances the convergence accuracy and decreases the optimization time of the GBDT model compared to other optimization techniques, while also accelerating the convergence process.
- The SETO-GBDT model outperforms other evaluation models in terms of evaluation error RMSE, R2 value, and evaluation time. The RMSE is 0.13, the R2 value is 0.98, and the evaluation time is 0.092s.
- The experimental findings confirm the accuracy of the SETO-GBDT model in evaluating the impact.

The superior performance of the SETO-GBDT model is evident in its higher convergence accuracy and shorter optimization time, making it a valuable tool for educational institutions seeking to assess and enhance the information literacy of their students. This model not only streamlines the evaluation process but also aligns with the goal of fostering well-rounded, information-savvy professionals in today's fast-paced, data-driven society.

REFERENCES

[1] Munavalli S B .Impact of information literacy skill assessment to explore learning resources[J].Journal of Library and Information Communication Technology, 2023.

[2] Juan Bartolomé, Garaizar P .Design and Validation of a Novel Tool to Assess Citizens' Netiquette and Information and Data Literacy Using Interactive Simulations[J].Sustainability, 2022, 14.

[3] Vitorino E V .Indicators for Information Literacy in Brazil: virtues, trends and possibilities[J].perspectivas em ciencia da informacao, 2022, 27 (2):7-36.

[4] Ferguson J .Flipping the (COVID-19) Classroom: Redesigning a First-Year Information Literacy Program during a Pandemic[J].portal: Libraries and the Academy, 2023, 23:145 - 168.

[5] Carless D .From teacher transmission of information to student feedback literacy: Activating the learner role in feedback processes:[J].Active Learning in Higher Education, 2022, 23(2):143-153.

[6] Paul M , Deja M ,Magorzata Kisilowska-Szurmińska, Gowacka E, Marzena W, Wojciechowska M.Understanding information literacy among doctoral students: an ILDoc model and assessment tool[J].The Journal of Academic Librarianship, 2024, 50(2).

[7] Fei D , Erjun S .How the ICT Development Level Influences Students' Digital Reading Literacy: A Multi-level Model Comparison Based on PISA 2018 Data[J].Frontiers of Education in China, 2022, 17(2):151-180.

[8] Esfandiari R , Arefian M H .Developing collective eyes for Iranian EFL teachers' computer-assisted language assessment literacy through internet- based collaborative reflection[J].Education and Information Technologies, 2024, 29(8):9473-9494.

[9] Riithi C W , Kimani G .Assessment of Literacy Levels of Small-Scale Poultry Farmers for improved information behaviour and production: case of Kabaa location, Kenya[J].International Journal of Current Aspects, 2022.

[10] Ganesan P , Gunasekaran M .Assessment of information literacy skills and knowledge-based competencies in using electronic resources among medical students[J].Digit. Libr. Perspect. 2022, 38:444-459.

[11] Vianna B I , Caregnato S E .Institutional diagnosis models for implementation of Information Literacy programs in academic libraries[J]. perspectivas em ciencia da informacao, 2022, 27(2):242-267.

[12] Faber C J .Information Literacy Modules for First-Year Engineering Students[J].AEE Journal, 2022.

[13] Yu C , Xu W .Writing assessment literacy and its impact on the learning of writing: a netnography focusing on Duolingo English Testexaminees[J]. Language Testing in Asia, 2024, 14(1).

[14] Schmidt L A J , Deschryver M .The Role of Digital Application Literacy in Online Assessment.[J].Journal of Educational Technology Systems, 2022, 50.

[15] Stellwagen Q H , Rowley K L , Otto J .Flip This Class: Maximizing Student Learning in Information Literacy Skills in the Composition Classroom through Instructor and Librarian Collaboration[J].Journal of library administration, 2022.

[16] Current M D .Tracking student learning outcome engagement at the reference desk to facilitate assessment[J].Reference services review, 2023.

[17] Ke D D , Suzuki K , Kishi H , Kurokawa Y, Shen S. Definition and assessment of physical literacy in children and adolescents: a literature review[J]. Journal of Physical Fitness and Sports Medicine, 2022, 11(3):149-159.

[18] Xiao Y L, Sehn H R, Xu Y H, Yu T G, Zheng Y J, Xie H Z, Ting J. Water quality prediction in Minjiang River basin based on GBDT-LSTM[J]. Journal of Ecology and Environment,2024,33(04):597-606.

[19] Nie X ,Hong Y. Long-term power load forecasting based on WOA-VMD-GBDT[J]. Science and Technology Wind,2024,(12):70-72.

[20] Chen H, Shang B.. Mathematical modeling and performance analysis of paper mill beating degree GBDT[J]. Paper Science and Technology,2024,43(04):73-76.

[21] Emami H .Stock exchange trading optimization algorithm: a human-inspired method for global optimization[J].Journal of supercomputing, 2022(2) :78.

[22] He J, Wang W, Liu C, Su Y. Research on the prediction and control method of tunnel surface settlement based on GBDT-PSO hybrid algorithm[J]. Transportation World,2024,(13):181-183.

[23] Sun B L, Sun B W. Entropy optimization with improved SCA for image segmentation and its application to image recognition[J]. Computer Engineering and Design,2024,45(05):1516-1524.

[24] Kumar M, Kulkarni A J, Satapathy S C. Socio evolution & learning optimization algorithm: a socio-inspired optimization methodology[J]. Future Generation Computing System, 2018, 81:252-272.

[25] Askari Q, Saeed M, Younas I. Heap-based optimizer inspired by corporate rank hierarchy for global optimization[J]. Expert Systems with Applications, 2022, 161: 113702.

[26] Houssein E H , Saad M R , Hashim F A , Shaban H, Hassaballah H. Lévy flight distribution: a new metaheuristic algorithm for solving engineering optimization problems[J].Engineering Applications of Artificial Intelligence, 2020, 94: 103731.

Bridging the Gap Between Industry 4.0 Readiness and Maturity Assessment Models: An Ontology-Based Approach

ABADI Asmae[1]*, ABADI Chaimae[2], ABADI Mohammed[3]
Euromed University of Fes, UEMF, Morocco[1]
ENSAM, Moulay Ismail University, Meknes, Morocco[2]
Team Optimization of Production Systems and Energy, Laboratory of Advanced Research in Industrial and Logistic Engineering (LARILE), Hassan II University of Casablanca, Morocco[3]

***Abstract*—The rapid evolution of Industry 4.0 technologies has created a complex and interconnected landscape of readiness and maturity assessment models. However, these models often fail to address the full spectrum of organizational readiness across strategic, technological, operational, and cultural dimensions, while also not accounting for emerging paradigms such as Industry 5.0. This paper proposes a conceptual model for an ontology that integrates all relevant domain knowledge into a unified framework, capturing strategic, technological, operational, and cultural readiness and maturity within a single comprehensive model. The ontology provides a systematic approach to understanding the interconnectedness of I4.0 and Industry 5.0 assessment models, facilitating a holistic view of an organization's preparedness for digital transformation. By bridging the gap between these two stages of industrial evolution, the model enables interoperability across diverse frameworks, promoting more informed decision-making and strategic planning. This research highlights the potential of the proposed ontology to support the ongoing shift from Industry 4.0 to Industry 5.0, offering a valuable tool for researchers, practitioners, and decision-makers navigating the complexities of next-generation industrial ecosystems. The paper further discusses the theoretical underpinnings and practical applications of the model in fostering a smooth transition toward a more human-centric, sustainable, and technologically advanced industrial future.**

***Keywords*—Industry 4.0; readiness assessment; maturity assessment; digital transformation; ontology development; conceptual model; knowledge engineering**

I. Introduction

The rapid advancement of Industry 4.0 technologies has brought transformative changes to manufacturing and industrial operations, enabling organizations to achieve unprecedented levels of efficiency, flexibility, and competitiveness [1, 2, 3, 4]. By integrating technologies such as the Internet of Things (IoT), artificial intelligence (AI), big data analytics, cloud computing, and cyber-physical systems, Industry 4.0 represents a paradigm shift in how industries operate and innovate [3]. However, the successful adoption of these technologies requires more than technical implementation; it demands a comprehensive understanding of organizational readiness across multiple dimensions, including technology, workforce, processes, and strategy.

Assessing readiness and maturity for Industry 4.0 is critical for organizations to identify their current capabilities, recognize gaps, and prioritize efforts to address them. Traditional readiness assessment approaches, such as the IMPULS and SIRI frameworks [4, 5], provide valuable starting points by identifying key dimensions and establishing readiness levels. However, these models often rely on static evaluations and qualitative surveys, limiting their ability to provide real-time insights or address the interconnected nature of Industry 4.0 dimensions. Their lack of reasoning capabilities results in assessments that may overlook nuanced interdependencies between readiness factors, leading to generalized or incomplete recommendations.

To address these limitations, this paper proposes an ontology-based framework for assessing Industry 4.0 readiness. Ontologies offer a structured and formal representation of knowledge, enabling the modeling of complex and dynamic domains. The proposed framework captures key readiness dimensions, including connectivity, digital infrastructure, cybersecurity, workforce capabilities, strategy, and data analytics. It also intends to integrate in next steps reasoning capabilities through the Semantic Web Rule Language. This innovative approach automates the inference of readiness levels based on organization-specific input data, ensuring consistency, transparency, and scalability.

The use of ontology and reasoning mechanisms provides several benefits. It not only standardizes the assessment process but also generates actionable insights by uncovering the interdependencies between readiness dimensions. For instance, the framework can identify how gaps in workforce skills might affect the effective deployment of digital infrastructure or how a lack of cybersecurity measures could hinder data analytics capabilities. By offering dynamic and tailored recommendations, the framework empowers organizations to make data-driven decisions, strategically allocate resources, and accelerate their digital transformation.

This article is structured as follows. First, a review of existing Industry 4.0 readiness models highlights their contributions and limitations, establishing the need for an ontology and rule based reasoning approach. Next, Section III methodology for developing the ontology, including the design of its classes, properties, and conceptual models. This is

followed by a detailed discussion of the theoretical underpinnings and practical applications of the model in fostering a smooth transition toward a more human-centric, sustainable, and technologically advanced industrial future in Section IV. Finally, the paper is concluded in Section V.

II. Related Work

This section reviews the fundamental concepts and characteristics of Industry 4.0, its requirements and enabling technologies, and the existing readiness models that evaluate the preparedness and implementation of Industry 4.0.

A. Industry 4.0 Overview and Characterization

The term Industry 4.0 originated in Germany, specifically in Hannover in 2011, marking the transformative potential of integrated technologies in reshaping global value chains. This initiative highlighted the possibilities of product customization and novel production methods [6], facilitated by the interaction of technologies across physical, digital, and biological domains. This fusion marks a distinct advancement in the fourth industrial revolution compared to its predecessors.

According to the Germany Trade and Invest Institute [7], Industry 4.0 represents a technological leap from embedded systems to cyber-physical systems (CPSs), leveraging the power of the Internet, data, and services. In this paradigm, industrial machinery evolves to not only process products but also enable communication between products and machinery, effectively directing the production process. Industry 4.0 represents a transformative leap forward in manufacturing, characterized by the integration of advanced technologies and their constant interaction across physical, digital, and biological domains.

It marks a new maturity stage for manufacturing companies, leveraging technologies such as the Internet of Things (IoT), Cloud Computing (CC), Big Data (BD), Artificial Intelligence (AI), and Cyber-Physical Systems (CPSs) to create interconnected, intelligent production environments. These systems enable real-time decision-making, remote monitoring, and flexible modular production processes [6-9]. To implement Industry 4.0, several main features are identified that support the evolution of intelligent production systems [10]:

- Interoperability, Integrity, and Awareness: The degree of system collaboration in utilizing capabilities, sharing information, and intelligent decision-making [11].
- Virtualization: Enabling remote traceability and monitoring of processes through sensors, creating smart factories.
- Service Orientation: Utilizing service-oriented software alongside IoT technologies.
- Real-Time Operation Capability: Facilitating instant data gathering, processing, and decision-making.
- Modularity: Flexible production processes involving the coupling and decoupling of production modules.
- Decentralization: Allowing CPSs to make independent decisions and produce locally, utilizing technologies like 3D printing.

The integration of these features is made possible by enabling technologies. For instance, IoT involves billions of interconnected devices like sensors and industrial equipment, facilitating real-time data collection and analysis [2,10].AI plays a central role in enabling Industry 4.0 by integrating intelligent functionalities across the value chain, from customer acquisition to operations management [3,4,9]. Moreover, big data and cloud computing address the exponential growth of data generated in manufacturing, offering scalable solutions for data storage, analysis, and processing.

The exponential growth of data in Industry 4.0 often exceeds human processing capabilities. Cloud Computing (CC) addresses this challenge by offering shared, on-demand resources via the Internet, delivering high-quality services at reduced costs [4, 9]. Simultaneously, Big Data (BD) enhances decision-making by analyzing vast amounts of information characterized by volume, velocity, variety, and veracity [4]. In addition to enabling technologies, Industry 4.0 fosters smarter work environments where technologies enhance human capabilities rather than replace them. Meindl et al. [12] emphasize that advanced systems support decision-making, creativity, and safety, leading to smarter workplaces. This human-centric approach has given rise to discussions around Industry 5.0 [13], which emphasizes workers' central role in digital transformation.

Industry 4.0 is more than the adoption of cutting-edge technologies; it’s about connecting these technologies to foster organizational growth and operational efficiency [14]. By leveraging advanced technologies like robotics, additive manufacturing, and analytics, companies can drive innovation, improve customer experiences, and enable predictive decision-making. For instance, smart products and connected systems enhance customer interactions through enriched post-sales support and tailored marketing strategies [15].

However, the implementation of Industry 4.0 varies widely based on organizational readiness, technological infrastructure, and economic development. Developing nations often face challenges in adopting these technologies while maintaining competitive advantage [16, 17]. Organizations must navigate talent development, process changes, and strategic human resource management to align with Industry 4.0 demands [18,19]. Readiness and Maturity models serve as valuable tools in assessing current adoption levels and guiding strategic implementation efforts.

B. Industry 4.0 Readiness and Maturity Models

The emergence of Industry 4.0 has prompted the development of numerous readiness and maturity models, each aiming to assess and guide organizations through their transformation journeys. While these models have made significant contributions to understanding readiness, they also exhibit notable limitations, particularly in terms of validation, granularity, and cross-industry applicability. Table I reviews prominent Industry 4.0 readiness models and identifies gaps that motivate the proposed ontology based readiness assessment smart system.

TABLE I. ANALYSIS OF EXISTING INDUSTRY 4.0 READINESS AND MATURITY MODELS

Model Name	Ref	Country	Contribution	Focus Areas
ACATECH I4.0 Maturity Index	[20]	Germany	Six-level progression emphasizing adaptability through technology and organization integration.	Technology & organization integration
IMPULS Industrie 4.0 Readiness	[21]	Germany	Practical tool tailored for German manufacturing industries.	Manufacturing industries
Singapore Smart Industry Readiness Index	[22]	Singapore	Comprehensive framework with 16 dimensions under three pillars: process, technology, and organization.	Holistic transformation
6Ps Maturity Model for SMEs	[23]	Italy	Tailored to SMEs, addressing unique small enterprise challenges with six stages.	SME-specific
Integrated IoT Capability Maturity Model	[24]	Netherlands	Combines capabilities from diverse frameworks to improve IoT management through five stages.	IoT management
SIMMI 4.0	[25]	Germany	Structured focus on digitization levels and integration across technologies and departments.	Digitization & IT integration
Industry 4.0 Readiness and Maturity Model	[26]	Austria	Comprehensive framework with nine dimensions, including strategy, leadership, governance, and innovation.	Technological & organizational aspects
Maturity Model for Smart Manufacturing	[27]	Turkey / Cyprus	Modular and incremental design adaptable to manufacturing contexts.	Modular for manufacturing
Categorical Framework of Manufacturing	[28]	United Kingdom	Multi-level approach integrating intelligence and automation across four dimensions.	Intelligence & automation

The ACATECH I4.0 Maturity Index [20] outlines a six-level progression from computerization to adaptability, emphasizing the integration of technological and organizational capabilities. This theoretical framework offers a structured approach to transformation but remains limited to conceptual discussions without cross-industry validation or practical implementation examples. Its lack of application-oriented guidance reduces its relevance for diverse industrial contexts. The IMPULS Industrie 4.0 Readiness model [21] assesses readiness across six dimensions: strategy, organization, IT infrastructure, smart products, smart services, and employees. It provides practical tools for manufacturing industries, particularly in Germany, making it highly relevant for this sector. However, its sector-specific focus restricts its versatility, limiting its applicability to other industries or global contexts.

The Singapore Smart Industry Readiness Index [22] provides a comprehensive framework encompassing 16 dimensions across three pillars: process, technology, and organization. Its holistic approach and intuitive tools are valuable for readiness assessment. Nevertheless, the model lacks extensive validation across sectors and does not provide detailed action plans for achieving specific maturity levels, reducing its utility for organizations seeking granular guidance. The 6Ps Maturity Model for SMEs [23] is tailored to address the unique challenges of small and medium enterprises (SMEs). With six stages: Plan, Prepare, Predict, Produce, Promote, and Proliferate. The model has been validated through case studies involving nine SMEs. However, its applicability to larger enterprises or different industrial contexts remains underexplored, highlighting a need for broader adaptability. The Integrated IoT Capability Maturity Model [24] combines capabilities from various frameworks to improve IoT management through five stages, ranging from primitive to maximizing, and three dimensions: technology, authority & culture, and knowledge management. While its operational focus on IoT is notable, the absence of practical validation and its narrow scope limit its broader relevance in the Industry 4.0 landscape. SIMMI 4.0 [25] emphasizes digitization and IT integration, guiding organizations through five maturity stages, from basic to optimized full digitization. Its focus on vertical and horizontal integration, cross-technology criteria, and digital product development provides a robust framework for IT landscapes. However, it neglects critical factors such as organizational culture and employee readiness, which are essential for successful Industry 4.0 adoption.

The Industry 4.0 Readiness and Maturity Model [26] offers a comprehensive framework with nine dimensions, including strategy, leadership, governance, and innovation. This balanced approach to technological and organizational readiness has been validated through a manufacturing case study. However, the complexity of its framework poses challenges for smaller enterprises, making adoption difficult without significant resources. The Maturity Model for Smart Manufacturing [27] adopts a modular and incremental design, making it adaptable to various manufacturing contexts. It evaluates five key dimensions: strategy, leadership, technology, culture, and operations. While validated through a case study, the model\u2019s scalability to larger enterprises and applicability in non-manufacturing sectors are not well explored. The Categorical Framework of Manufacturing [28] integrates intelligence and automation across four dimensions: factory, business, process, and customers. This multi-level framework offers a comprehensive perspective on readiness. However, its lack of practical examples and detailed guidance for progression limits its effectiveness and real-world applicability.

Despite the diverse contributions of these models, significant limitations persist. Many models focus heavily on technological aspects, such as digitization and IoT integration, while overlooking critical organizational factors like culture, leadership, and employee readiness. Additionally, most frameworks are constrained by sector-specific designs, limiting their adaptability to different industries. The lack of practical validation and real-world case studies further hampers their utility, as organizations struggle to translate theoretical guidance into actionable strategies. Moreover, the absence of reasoning-based approaches reduces their ability to dynamically adapt to evolving industrial contexts, creating a gap for more intelligent and flexible assessment tools.

Existing Industry 4.0 readiness models provide valuable insights but fail to address the need for holistic, adaptive, and validated frameworks. To overcome these challenges, this paper

proposes an ontology-driven framework incorporating reasoning mechanisms to enable dynamic and adaptive readiness assessment. This approach aims to offer a more comprehensive and actionable tool for Industry 4.0 transformation across diverse industries and organizational contexts.

III. Conceptual Model Development Methodology

This section outlines the methodology employed to construct a robust ontology specifically designed for Industry 4.0, along with its corresponding conceptual model. Recognizing the interdependence between these two components, the development process adopts a systematic and strategic approach to conceptualization. The conceptual model, which serves as the foundation for ontology development, provides a clear and visual representation of the domain, facilitating better understanding and usability. Moreover, the model is designed to be reusable, allowing adaptation across various ontology representation languages. The methodology applied here is informed by well-established practices in ontology development, tailored to address the unique characteristics and challenges of Industry 4.0.

To achieve this, a hybrid methodology was employed, integrating elements from different prominent approaches: the Uschold and King methodology [29], METHONTOLOGY [30] and Ontology Development [31]. Specifically, the structured framework proposed by Uschold and King was combined with the iterative processes of Ontology Development 101 and further enriched by the conceptualization techniques outlined in METHONTOLOGY.

This hybrid approach leverages the strengths of each methodology. The Uschold and King framework offers a systematic starting point for constructing the ontology's initial structure. Ontology Development 101 introduces iterative refinement, enabling detailed and comprehensive development. The inclusion of METHONTOLOGY ensures a deep understanding of the domain through its emphasis on conceptualization, ensuring that the model accurately captures the semantics of Industry 4.0.

The construction of the conceptual model followed a series of structured steps:

1) Define the domain, scope, and purpose: This step establishes clear boundaries and objectives for the ontology, ensuring it aligns with the specific requirements of Industry 4.0 readiness assessment.

2) Capture knowledge and develop conceptualization: Through iterative refinement, the following tasks were performed:

Identifying key terms and concepts relevant to Industry 4.0, such as smart factories, cyber-physical systems, IoT, and advanced analytics.

Defining classes and their hierarchical relationships to represent the domain's structure.

Establishing object and data properties to define relationships and attributes within the domain.

3) Create the conceptual model: The captured knowledge was transformed into a graphical representation, ensuring clarity and reusability. The model was designed to reflect core aspects of Industry 4.0, such as interoperability, automation, and digital transformation.

Fig. 1 illustrates the workflow adopted in this methodology, demonstrating the integration of the selected approaches and their application to Industry 4.0 ontology development.

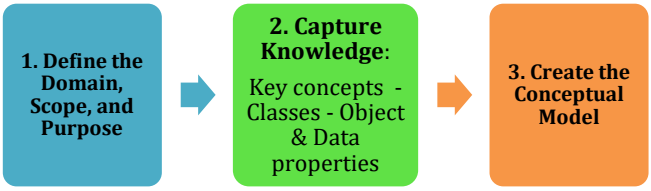


Fig. 1. The workflow of the ontology conceptual model development methodology.

B. Domain, Scope, and Purpose Definition

In developing our ontology for Industry 4.0 readiness assessment, the first step is to define the domain, scope, and purpose clearly. The domain centers on Industry 4.0, specifically focusing on the readiness and maturity of industrial companies in adopting advanced technologies and practices. The scope includes dimensions such as technological capabilities, organizational processes, workforce skills, and strategic alignment while excluding unrelated areas like consumer behaviors or non-industrial sectors.

The purpose of this ontology, as presented in Table II, is to provide a structured framework for evaluating and comparing the readiness of companies for Industry 4.0 transformation, facilitating informed decision-making and guiding improvement strategies. This step ensures that the ontology is both focused and relevant, addressing the key challenges and requirements of stakeholders in the Industry 4.0 landscape.

C. Capturing Knowledge and Conceptualization Definition

1) Identifying key terms and concepts: The process of capturing knowledge within the context of Industry 4.0 readiness involves identifying the core elements that influence an organization's journey toward digital transformation. This is achieved through the use of an ontology, which structures and organizes these elements into defined categories that enable clearer understanding, assessment, and decision-making. The conceptualization phase takes these elements and translates them into formal representations, allowing them to be analyzed and applied practically across industries. Our ontology is structured around four main dimensions: Strategic, Technological, Operational, and Cultural as presented in Fig. 2.

TABLE II. SCOPE, DOMAIN AND KNOWLEDGE SOURCE OF THE INDUSTRY 4.0 ASSESSMENT ONTOLOGY

Domain	The domain of interest of this work is industry 4.0 readiness and maturity
Date	2024-2025
Purpose	• Establish a standardized framework for organizing and categorizing Industry 4.0 readiness data, ensuring consistency and comparability across evaluations. • Enable reasoning to infer implicit knowledge, identify gaps in readiness, and derive automated recommendations for targeted improvements. • Facilitate evidence-based decision-making by linking readiness dimensions to actionable strategies, prioritizing critical areas, and enabling benchmarking. • Provide adaptability for evolving Industry 4.0 practices by integrating new concepts, technologies, and criteria as the domain progresses. • Support cross-functional and cross-organizational alignment by offering a shared vocabulary for clear communication and collaboration. • Serve as a foundational tool for advancing research, enabling data-driven insights, and fostering innovation in Industry 4.0 readiness assessments and practices.
Scope	The scope of the ontology is to provide a structured framework for assessing and auditing industrial companies' readiness for Industry 4.0 transformation across strategic, technological, operational, and cultural dimensions, enabling organizations to benchmark their progress, identify gaps, and guide their digital transformation journey.
Source of Knowledge	• ACATECH I4.0 Maturity Index [20] • IMPULS Industry 4.0 Readiness [21] • Singapore Smart Industry Readiness Index [22] • 6Ps Maturity Model for SMEs [23] • Integrated IoT Capability Maturity Model [24] • SIMMI 4.0 [25] • Industry 4.0 Readiness and Maturity Model [26] • Maturity Model for Smart Manufacturing [27] • Categorical Framework of Manufacturing [28] • Reference Architecture Model for Industry 4.0 • Surveys and Case Studies from Leading Manufacturers • Reports and Whitepapers from Industry Associations • Public Sector Digital Transformation Initiatives • Interviews and Insights from Technology Providers

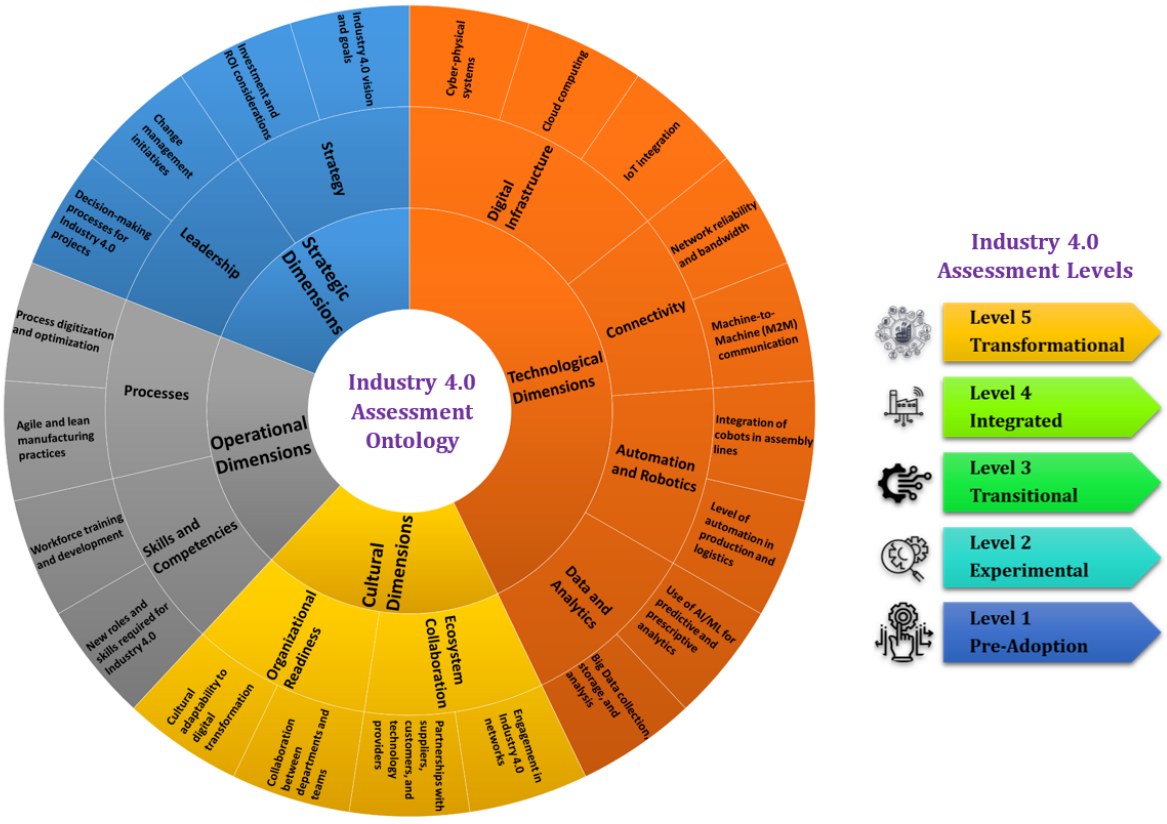


Fig. 2. Dimensions, associated fields and evaluation levels of the proposed industries 4.0 assesment ontology.

2) Capture knowledge and develop conceptualization: The assessment dimensions encompass the essential factors that determine an organization's preparedness for Industry 4.0 adoption. Within each dimension, specific classes are defined to capture the knowledge associated with different aspects of transformation. For instance, the Strategic Dimensions contain classes such as Strategy, which represents the company's approach to aligning its goals with Industry 4.0, ranging from early awareness to full integration. Similarly, Leadership in this dimension reflects varying stages of leadership involvement, from lack of awareness to driving an innovation culture.

In the Technological Dimensions, the ontology includes classes such as Digital Infrastructure, which captures the evolution of IT systems from basic setups to fully integrated, smart manufacturing solutions. Classes like Processes, Data and Analytics, and Automation and Robotics represent the

increasing sophistication of processes, data utilization, and automation in an Industry 4.0 environment. Each of these classes models the different stages of digital transformation in terms of technology adoption, from initial automation trials to the use of AI and machine learning in decision-making processes.

The Operational Dimensions focus on the organization's day-to-day operations, particularly the level of connectivity within the enterprise. The Connectivity class, for example, spans from isolated devices to seamless, real-time communication across the organization, while the Workforce Skills class tracks the evolution of employee competencies, capturing the transition from basic awareness to continuous adaptation of advanced digital skills. Finally, the Cultural Dimensions include classes like Organizational Readiness, which represents the organization's cultural alignment with Industry 4.0 principles. This class captures the evolution from resistance to change to a fully agile and innovation-driven environment. Additionally, the Ecosystem Collaboration class addresses the growing importance of strategic partnerships, from no collaboration to leadership in collaborative networks within the industry.

The conceptualization of the ontology involves translating these real-world concepts into formal classes and defining relationships between them. For example, Leadership is conceptualized as a class that influences both Strategy and Organizational Readiness, with relationships indicating how leadership's engagement drives the alignment of strategy and cultural transformation. Each class is further defined with specific properties and attributes, representing the maturity levels or stages of development. In the Strategy class, for instance, these properties might range from "Initial Exploration" to "Fully Aligned Corporate Strategy," reflecting different levels of maturity in aligning the company's strategy with Industry 4.0 goals. As part of the knowledge capture process, the ontology also formalizes the data collection methods that link these conceptualized classes to real-world assessments. This allows organizations to map their existing capabilities against the maturity levels defined in the ontology, offering a detailed picture of their current readiness. For example, Workforce Skills can be tied to specific skill data collected through employee assessments, training programs, and competency evaluations.

3) Create the conceptual model: The conceptualized ontology forms the basis of structured audits and assessments of Industry 4.0 readiness. By representing the different stages of maturity in each class, it allows organizations to pinpoint their strengths and weaknesses, identifying areas where they need to improve. This structured approach to capturing and organizing knowledge ensures that the assessment process is objective, repeatable, and aligned with the overall goal of driving digital transformation within industrial settings. The conceptualization phase, therefore, not only organizes knowledge but creates a dynamic framework for ongoing evaluation, facilitating continuous progress toward Industry 4.0. In fact in this stage, we defined not only the relations between the different classes but also the details on the minimum requirements for each maturity level in our model as represented in Fig. 3.

Industry 4.0 Assessment Dimensions		Level 1 Pre-Adoption	Level 2 Experimental	Level 3 Transitional	Level 4 Integrated	Level 5 Transformational
Strategic Dimensions	Strategy	No formal Industry 4.0 strategy or awareness.	Initial exploration of Industry 4.0 opportunities and risks.	Defined strategy with clear milestones.	Fully aligned corporate strategy with Industry 4.0 goals.	Continuous evolution of strategy based on innovation and market trends.
	Leadership	Leadership lacks awareness or engagement in Industry 4.0.	Leadership explores pilot initiatives and starts allocating budget.	Leaders actively drive initiatives and engage teams.	Leadership ensures organization-wide commitment and resource allocation.	Leadership promotes innovation culture and global thought leadership.
Technological Dimensions	Digital Infrastructure	Basic IT systems with minimal automation or connectivity.	Isolated systems and initial pilot setups for IoT or cloud computing.	Moderate IT integration; emerging IoT networks and cloud usage.	Fully integrated systems with real-time data access and decision-making.	Adaptive and scalable systems, leveraging edge computing and advanced networks.
	Processes	Highly manual processes with minimal standardization.	Early-stage digitalization of selected processes.	Processes are increasingly digitized and standardized across units.	Digital and optimized processes; cross-departmental collaboration.	Adaptive and self-optimizing processes with AI-driven automation.
	Data and Analytics	Minimal data collection; analysis is manual or nonexistent.	Basic data collection for pilot projects; spreadsheets or simple tools used.	Systematic data collection and some use of business intelligence tools.	Advanced analytics with AI/ML for predictive insights across operations.	Prescriptive analytics and autonomous decision-making are fully implemented.
	Automation and Robotics	Little or no automation; reliance on manual labor.	Robotics or automation applied in isolated tasks or areas.	Semi-automated processes with robotic integration in key operations.	High level of automation in core operations, including cobotic systems.	Fully autonomous production systems with cyber-physical integration.
Operational Dimensions	Connectivity	No machine-to-machine (M2M) or IoT connectivity.	Isolated devices connected in pilot projects.	Partial connectivity across production lines or departments.	Organization-wide IoT and M2M communication for seamless data flow.	Ubiquitous connectivity enabling real-time, cross-enterprise collaboration.
	Workforce Skills	Workforce lacks Industry 4.0 awareness and skills.	Limited upskilling initiatives in response to pilot projects.	Training programs are in place; workforce acquires basic digital competencies.	Comprehensive training and workforce adaptability to Industry 4.0 demands.	Workforce continuously evolves with advanced skills; innovation-driven roles.
Cultural Dimensions	Organizational Readiness	Resistance to change; no alignment with Industry 4.0 objectives.	Isolated initiatives with some departmental support.	Organization shows buy-in and cross-departmental collaboration improves.	Entire organization is aligned and agile, supporting Industry 4.0 objectives.	Agile, innovation-focused culture with strong collaboration and change management.
	Ecosystem Collaboration	No collaboration with external partners for Industry 4.0 initiatives.	Initial partnerships for pilot projects or technology trials.	Active collaboration with select suppliers, customers, and tech providers.	Strong partnerships and ecosystem participation for mutual growth.	Strategic leader in Industry 4.0 networks; driving standards and innovation.

Fig. 3. The proposed industry 4.0 assessment model for the four dimensions – minimum requirments.

IV. Industry 4.0 Assessment Ontology Conceptual Model: Unveiling the Interconnected Landscape of Industry 4.0 Assessment Models

The Industry 4.0 readiness assessment ontology is designed to evaluate and track the maturity of organizations in their adoption of Industry 4.0 technologies. The ontology is structured around several key dimensions, as presented in Tables III and IV that represent critical aspects of organizational readiness, with each dimension encompassing various sub-dimensions, attributes, and functions to holistically assess an organization's progress and capabilities in adopting Industry 4.0 principles. These dimensions are organized into four main categories: Strategic Dimensions, Technological Dimensions, Operational Dimensions, and Cultural Dimensions.

To assess the Strategic readiness, the Strategy sub-dimension captures the organization's preparedness and strategic direction towards Industry 4.0 adoption. Key attributes include strategyLevel, which categorizes the organization's strategy (e.g., Pre-Adoption, Experimental), and awareness, which assesses the level of understanding about Industry 4.0 technologies. The milestones attribute tracks the significant stages in the organization's adoption journey. The associated functions like defineStrategy() and evaluateProgress() allow for dynamic updates and evaluation of the organization's strategic alignment with Industry 4.0 goals. The Leadership sub-dimension focuses on the commitment and engagement of leadership in driving the Industry 4.0 transformation. The attributes engagementLevel, budgetAllocation, and innovationCulture reflect leadership's role in fostering technological adoption. Functions such as assessLeadershipCommitment() and promoteInnovation() evaluate leadership's contribution to innovation and resource allocation.

TABLE III. Description of the Industry 4.0 Assessment Ontology Main Classes

Class	Description
Industry4.0ReadinessAssessment	Represents the overall assessment process to evaluate a company's Industry 4.0 readiness.
Dimension	Represents a high-level category of readiness, such as Strategy, Technology, or Operations.
Strategic Dimensions	Represents the readiness related to strategy and leadership.
Criteria: Strategy	Evaluates the existence and maturity of Industry 4.0 strategy in the organization.
Criteria: Leadership	Assesses leadership involvement and vision in adopting Industry 4.0 technologies.
Technological Dimensions	Represents readiness related to infrastructure, processes, data, and automation.
Criteria: Digital Infrastructure	Assesses the maturity of IT systems, connectivity, and integration.
Criteria: Connectivity	Evaluates the integration of IoT and machine-to-machine communication.
Criteria: Data and Analytics	Measures the ability to collect, analyze, and use data for decision-making.
Criteria: Automation and Robotics	Assesses the level of automation, including robotics and cobotics integration.
Operational Dimensions	Represents the readiness of operations, workforce, and connectivity.
Criteria: Processes	Evaluates the level of digitization and optimization in organizational processes.
Criteria: Workforce Skills	Measures the readiness and adaptability of the workforce to Industry 4.0 changes.
Cultural Dimensions	Represents readiness in terms of organizational readiness and external collaboration.
Criteria: Organizational Readiness	Assesses the organization's openness and alignment with Industry 4.0 objectives.
Criteria: Ecosystem Collaboration	Measures the level of partnership and collaboration within Industry 4.0 ecosystems.
Indicator	Represents the quantitative or qualitative measures used to evaluate a Criterion.
AssessmentResult	Captures the outcome of the readiness assessment, including scores and detailed feedback.
Company	Represents the organization being assessed for Industry 4.0 readiness.

TABLE IV. Description of the Industry 4.0 Assessment Ontology main Properties

Property Name	Domain	Range	Cardinality	Inverse Property
aggregates	Industry4.0ReadinessAssessment	Dimension	Multiple : A single Industry4.0ReadinessAssessment can aggregate multiple Dimensions.	-
hasCriteria	Dimension	Criteria	Multiple : Each Dimension must have at least one associated Criteria, but it can have multiple criteria.	isCriteriaOf
isEvaluatedUsing	Criteria	Indicator	Single : A single Criteria is evaluated using one or more Indicators.	evaluates
generates	Industry4.0ReadinessAssessment	AssessmentResult	An Industry4.0ReadinessAssessment generates exactly one AssessmentResult. This ensures that each assessment leads to a unique, consolidated result.	isGeneratedBy
isAssociatedWith	Company	Industry4.0Readiness Assessment	Multiple : A single Company can be associated with one or more Industry4.0ReadinessAssessments. This allows a company to conduct multiple assessments over time or for different operational units.	isCompanyOf

In order to assess the technological readiness and maturity, the Digital Infrastructure sub-dimension assesses the foundational technological components necessary for Industry 4.0. It includes attributes like systemIntegrationLevel, which measures the extent to which systems are integrated within the organization, and realTimeAccess and scalability, which evaluate the system's ability to handle real-time data and scale accordingly. The functions integrateInfrastructure() and evaluateInfrastructureReadiness() provide methods for enhancing and assessing the state of technological infrastructure. The Processes sub-dimension examines the degree of process optimization, with attributes such as standardization, digitizationLevel, and crossDeptCollaboration. Functions like analyzeProcessMaturity() and optimizeProcesses () allow for the assessment and improvement of business processes through digital transformation.

The Data and Analytics sub-dimension addresses data management and analytics capabilities within the organization. Key attributes include dataCollection, analyticsCapability, and predictiveAnalyticsUsage. The functions collectData() and generateInsights() help organizations manage data collection and derive valuable insights for decision-making. The Automation and Robotics sub-dimension evaluates the automation and robotics capabilities within the organization. The attributes automationLevel, coboticsUsage, and integrationComplexity allow organizations to assess their automation maturity. The functions evaluateAutomation() and implementRobotics() guide the improvement and integration of robotic systems into operations.

For the Operational assessment, the Connectivity sub-dimension is central to evaluating the effectiveness of data exchange across systems. It includes attributes like connectivityLevel, iotNetworks, and dataFlowEfficiency. Functions such as ensureSeamlessConnectivity() and evaluateNetworkPerformance() ensure that connectivity is optimized and functioning at a level necessary for Industry 4.0 operations. The Workforce Skills sub-dimension evaluates the workforce's readiness for Industry 4.0. Attributes such as trainingPrograms, skillCompetency, and upskillingFrequency measure the organization’s commitment to continuous workforce development. Functions like analyzeSkillGap() and designTrainingProgram() ensure that the workforce remains competitive and capable of handling Industry 4.0 challenges.

For the cultural readiness and maturity, the Organizational Readiness sub-dimension examines cultural factors such as changeResistance, departmentalAlignment, and readinessLevel. These attributes help measure the internal alignment and the organization’s preparedness to embrace change. Functions such as fosterAlignment() and assessCulturalReadiness() aim to promote cultural alignment across departments. The Ecosystem Collaboration sub-dimension explores the organization’s external engagement and collaboration with partners. Attributes like partnerships, externalEngagement, and innovation Contribution gauge the organization’s collaborative efforts with external stakeholders. The functions buildPartnerships() and evaluateCollaborationImpact() help foster and assess the impact of external collaborations.

A. Theoretical Implications of the Conceptual Model on the Research Landscape of Industry 4.0 Readiness and Maturity Models

The conceptual model of the Industry 4.0 readiness ontology contributes significantly to the theoretical discourse on maturity and readiness models in Industry 4.0. By systematically integrating strategic, technological, operational, and cultural dimensions, the ontology offers a holistic framework that bridges previously siloed perspectives. Its structured approach to defining sub-dimensions, attributes, and their associated properties enhances the granularity and depth of readiness assessments, making it a pivotal reference point in the research landscape.

One critical implication is the interoperability this model introduces between disparate maturity and readiness frameworks. Existing models, such as IMPULS, SIRI, and Acatech, often emphasize specific aspects of readiness, such as technology deployment, organizational strategy, or workforce skills. The proposed ontology synthesizes these elements, offering a unified structure that incorporates strategic foresight, leadership commitment, digital infrastructure, process optimization, workforce competency, and cultural readiness. This integration ensures that no key dimension is overlooked, enabling researchers to analyze Industry 4.0 readiness through a comprehensive lens. Moreover, the ontology's inclusion of properties and cardinalities facilitates interoperability between different domains of analysis. For instance, relationships such as aggregates, hasCriteria, and isEvaluatedUsing allow researchers to map criteria and indicators across dimensions, enabling comparative studies between industries or geographic regions. This ability to establish linkages between criteria across strategic, technological, operational, and cultural domains helps advance the theoretical foundation for multi-dimensional readiness studies.

The ontology also advances the understanding of interdependencies between dimensions. For example, the cultural dimension’s attributes, such as readinessLevel and changeResistance, are intrinsically linked to strategic and operational dimensions, such as leadership engagement and process standardization. By explicitly modeling these interdependencies, the ontology highlights the cascading effects of progress or bottlenecks in one dimension on others, offering new insights into the dynamic nature of Industry 4.0 readiness. In addition, the ontology introduces a functional perspective, with defined methods (e.g., evaluateProgress(), analyzeSkillGap (), and generateInsights()) that provide a basis for operationalizing readiness assessment. This functional view enables researchers to explore not only static maturity levels but also dynamic transitions and improvement trajectories, which are often missing from traditional models. This theoretical shift aligns with the ongoing need in the research landscape to transition from static assessments to continuous, iterative improvement frameworks. The ontology emphasizes also the scalability and adaptability of readiness models. By accommodating both single-value attributes (e.g., readiness Level) and multi-value attributes (e.g., trainingPrograms or partnerships), it ensures flexibility for application across industries of varying sizes and complexities. This adaptability

addresses a significant gap in existing research, where models are often criticized for being too rigid or industry-specific.

B. *Practical Applications in Industry 4.0 Maturity Assessment: Bridging to Industry 5.0 Readiness*

The conceptual model of the Industry 4.0 readiness ontology, presented in Fig. 4, extends its theoretical robustness to practical applications in assessing Industry 4.0 maturity while paving the way for Industry 5.0 readiness. By capturing essential dimensions such as strategy, technology, operations, and culture, the model equips organizations with a structured and actionable framework to evaluate their current capabilities and chart a clear path toward technological and organizational evolution.

The ontology's multi-dimensional structure allows organizations to conduct a detailed maturity assessment, evaluating their performance across strategic, technological, operational, and cultural dimensions. For example, companies can assess their digital infrastructure and automation capabilities under the technological dimension while simultaneously evaluating leadership commitment and workforce skills under strategic and operational dimensions. This holistic approach ensures that organizations not only implement technology but also align it with strategy and culture, avoiding common pitfalls of fragmented adoption. By using properties such as analyzeSkillGap(), organizations can identify specific areas of improvement. The ontology's capability to aggregate multiple dimensions ensures that assessments are not isolated but interlinked, highlighting interdependencies that can accelerate or hinder progress. For instance, a lack of investment in leadership engagement may directly impact the success of cultural transformation initiatives. The scalability of the ontology makes it adaptable to various industries and business sizes. For large-scale manufacturers, the model can evaluate complex systems like real-time data access, IoT network integration, and robotic automation. Meanwhile, for small and medium-sized enterprises (SMEs), the focus can shift to incremental improvements, such as standardization of processes and upskilling the workforce. This flexibility ensures the model's relevance across the industrial spectrum, addressing both high-tech innovators and businesses in the early stages of transformation.

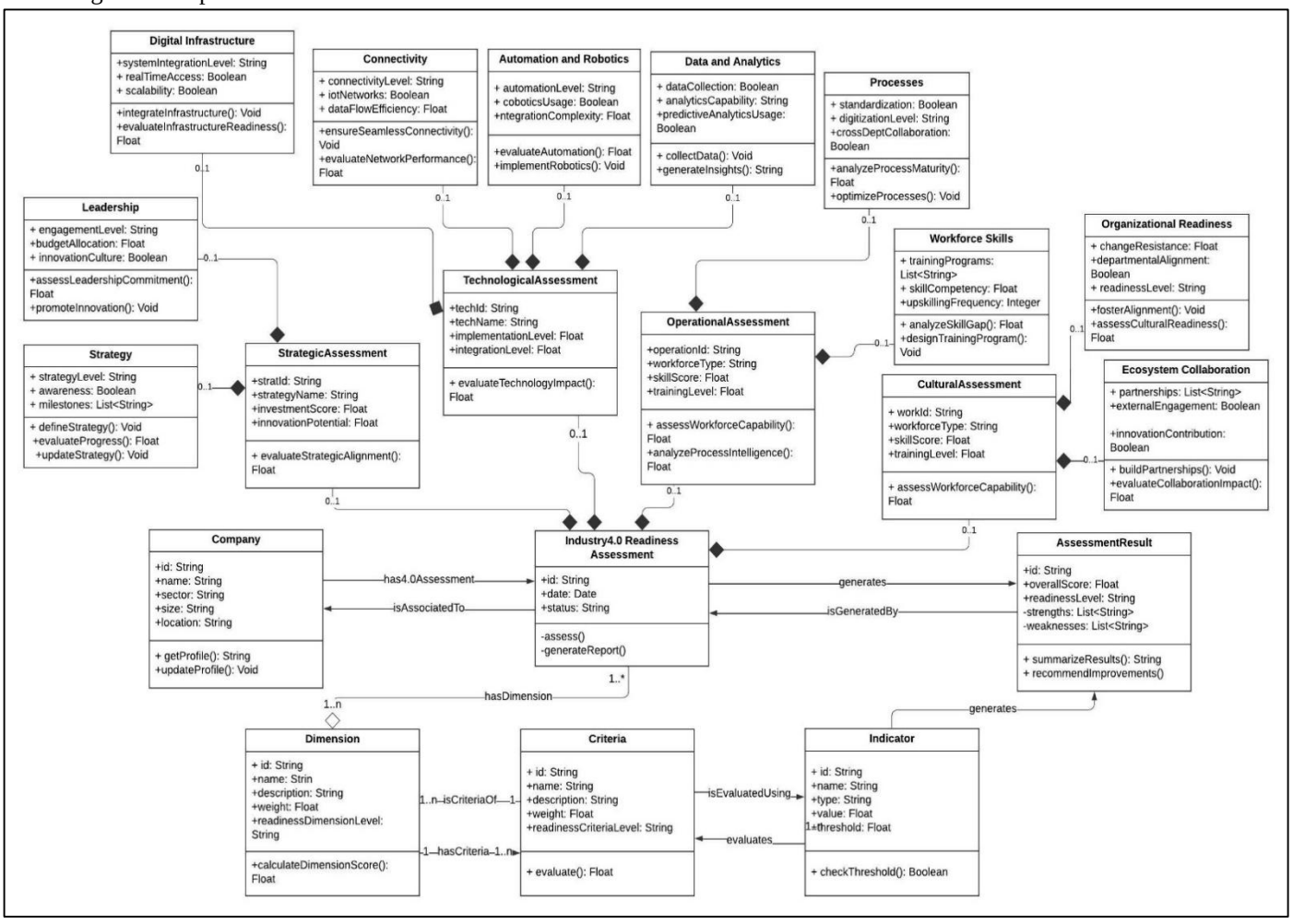


Fig. 4. Industry 4.0 Assessment conceptual model.

While rooted in Industry 4.0 principles, the ontology lays the groundwork for Industry 5.0 readiness by emphasizing human-centric innovation and sustainability. For instance, dimensions like workforce skills and cultural readiness align with Industry 5.0's focus on human-machine collaboration, where the role of cobotics and innovation culture becomes increasingly significant. The ontology's inclusion of attributes such as coboticsUsage, changeResistance, and trainingPrograms allows organizations to assess and enhance their preparedness for the collaborative, human-centered environments that define Industry 5.0. Furthermore, the ecosystem collaboration sub-dimension promotes partnerships and external engagements that are critical for sustainability and co-innovation in Industry 5.0. Functions like buildPartnerships() and evaluate Collaboration Impact() enable organizations to strengthen their position within an interconnected industrial ecosystem, fostering resilience and adaptability. The functional perspective embedded in the ontology supports real-time monitoring and continuous improvement. Methods such as evaluateProgress() and generateInsights() provide organizations with tools to regularly assess their Industry 4.0 maturity levels and dynamically adapt their strategies. This iterative approach ensures that organizations are not only maintaining their Industry 4.0 capabilities but are also actively transitioning toward Industry 5.0 readiness. The ontology facilitates benchmarking by enabling comparisons across industries, regions, and organizational sizes. Organizations can use indicators and criteria modeled in the ontology to measure their progress against industry standards or peers. This capability aids in identifying competitive gaps and aligning strategic decisions with industry trends.

For instance, companies can leverage insights derived from evaluateAutomation() and assessLeadershipCommitment() to prioritize investments and allocate resources more effectively. The ontology provides then a roadmap for organizations to future-proof their operations. By incorporating both current Industry 4.0 requirements and emerging Industry 5.0 principles, the model ensures that businesses are prepared for evolving technological landscapes. Attributes like scalability, predictiveAnalyticsUsage, and innovationCulture allow organizations to anticipate and adapt to future challenges, ensuring long-term competitiveness and sustainability.

The practical applications of the Industry 4.0 readiness ontology extend beyond assessing maturity. By offering a comprehensive, adaptable, and future-oriented framework, the model empowers organizations to not only excel in Industry 4.0 adoption but also position themselves as leaders in the transition to Industry 5.0. Its focus on interoperability, human-centric innovation, and continuous improvement makes it an essential tool for driving industrial transformation in an increasingly complex and interconnected world.

V. Conclusion

In conclusion, this research has developed a comprehensive conceptual model for assessing Industry 4.0 readiness and maturity, encapsulating the multifaceted dimensions and interrelationships critical to successful digital transformation. By integrating principles from prominent methodologies and drawing inspiration from established frameworks like IMPULS, SIRI, and Acatech, the proposed ontology provides a structured and holistic perspective on readiness assessment. The model delineates key connections between strategic, technological, operational, and cultural dimensions, offering insights into leadership engagement, digital infrastructure, workforce skills, and organizational alignment. Through its robust structure of classes, relationships, and functional properties, the ontology facilitates a nuanced understanding of interdependencies, enabling organizations to evaluate their maturity comprehensively and identify areas for improvement.

The conceptual model also establishes a foundation for bridging Industry 4.0 and Industry 5.0 readiness, emphasizing human-centric innovation, sustainability, and collaborative ecosystems. By accommodating diverse organizational contexts and fostering interoperability between dimensions, the model addresses critical gaps in existing frameworks, providing a flexible tool for continuous improvement and strategic decision-making. This work thus contributes to advancing the theoretical and practical landscape of Industry 4.0 readiness assessments, serving as a valuable resource for researchers, practitioners, and policymakers navigating the complexities of digital transformation.

The next critical step is to leverage this conceptual model to develop a fully operational ontology for Industry 4.0 readiness and maturity assessment. Such an ontology would formalize the domain knowledge, enable standardized representation, and enhance interoperability across systems, ultimately supporting tools for benchmarking, diagnostics, and strategic planning. By transitioning from conceptual modeling to ontology implementation, this work can catalyze meaningful progress in industrial transformation, positioning organizations to thrive in the evolving landscapes of Industry 4.0 and beyond.

References

[1] Elnadi, M., & Abdallah, Y. O. (2024). Industry 4.0: critical investigations and synthesis of key findings. Management Review Quarterly, 74(2), 711-744.

[2] Banitaan, S., Al-refai, G., Almatarneh, S., & Alquran, H. (2023). A review on artificial intelligence in the context of industry 4.0. International Journal of Advanced Computer Science and Applications, 14(2).

[3] Abadi, M., Abadi, C., Abadi, A., & Ben-Azza, H. (2022). A smart decision making system for the selection of production parameters using digital twin and ontologies. International Journal of Advanced Computer Science and Applications, 13(2).

[4] Ferreira, D. V., de Gusmão, A. P. H., & de Almeida, J. A. (2024). A multicriteria model for assessing maturity in industry 4.0 context. Journal of Industrial Information Integration, 38, 100579.

[5] Elhusseiny, H. M., & Crispim, J. (2024). A Review of Industry 4.0 Maturity Models: Theoretical Comparison in The Smart Manufacturing Sector. Procedia Computer Science, 232, 1869-1878.

[6] Schwab, K. (2024). 8. The Fourth Industrial Revolution-What It Means and How to Respond. Handbook of Research on Strategic Leadership in the Fourth Industrial Revolution, 29.

[7] GTAI (Germany Trade & Invest), Industrie 4.0-Smart Manufacturing for the Future, 2014, p. 21. Berlin.

[8] Abadi, C., Manssouri, I., & Abadi, A. (2020). An Artificial Intelligent based System to Automate Decision Making in Assembly Solution Design. International Journal of Advanced Computer Science and Applications, 11(7).

[9] Qi, Q., & Tao, F. (2019). A smart manufacturing service system based on edge computing, fog computing, and cloud computing. IEEE access, 7, 86769-86777.

[10] Monshizadeh, F., Moghadam, M. R. S., Mansouri, T., & Kumar, M. (2023). Developing an industry 4.0 readiness model using fuzzy cognitive maps approach. International Journal of Production Economics, 255, 108658.

[11] Ustundag, A., Cevikcan, E., Akdil, K. Y., Ustundag, A., & Cevikcan, E. (2018). Maturity and readiness model for industry 4.0 strategy. Industry 4.0: Managing the digital transformation, 61-94.

[12] Meindl, B., Ayala, N. F., Mendonça, J., & Frank, A. G. (2021). The four smarts of Industry 4.0: Evolution of ten years of research and future perspectives. Technological Forecasting and Social Change, 168, 120784.

[13] Romero, D., Stahre, J., & Taisch, M. (2020). The Operator 4.0: Towards socially sustainable factories of the future. Computers & Industrial Engineering, 139, 106128.

[14] Bai, C., Dallasega, P., Orzes, G., & Sarkis, J. (2020). Industry 4.0 technologies assessment: A sustainability perspective. International journal of production economics, 229, 107776.

[15] Cotteleer, M., & Sniderman, B. (2017). Forces of change: Industry 4.0. Deloitte Insights, 18(1), 1-16.

[16] Hopali, E., & Vayvay, Ö. (2018). Industry 4.0 as the last industrial revolution and its opportunities for developing countries. Analyzing the Impacts of Industry 4.0 in Modern Business Environments, 65–80.

[17] Cezarino, L. O., Liboni, L. B., Oliveira Stefanelli, N., Oliveira, B. G., & Stocco, L. C. (2021). Diving into emerging economies bottleneck: Industry 4.0 and implications for circular economy. Management Decision, 59(8), 1841-1862.

[18] Hecklau, F., Galeitzke, M., Flachs, S., & Kohl, H. (2016). Holistic approach for human resource management in Industry 4.0. Procedia cirp, 54, 1-6.

[19] Jan, Z., Ahamed, F., Mayer, W., Patel, N., Grossmann, G., Stumptner, M., & Kuusk, A. (2023). Artificial intelligence for industry 4.0: Systematic review of applications, challenges, and opportunities. Expert Systems with Applications, 216, 119456.

[20] Schuh, G., Anderl, R., Gausemeier, J., Ten Hompel, M., & Wahlster, W. (Eds.). (2017). Industrie 4.0 maturity index: die digitale transformation von unternehmen gestalten. Herbert Utz Verlag.

[21] K. Lichtblau, V. Stich, R. Bertenrath, M. Blum, M. Bleider, A. Millack, K. Schmitt, E. Schmitz and M. Schroter, "IMPULS-industrie 4.0-readiness, (2015).

[22] Utomo, S., & Setiastuti, N. (2019). Industri 4.0: Pengukuran Tingkat Kesiapan Industri Tekstil dengan Metode Singapore Smart Industry Readiness Index. Techno Nusa Mandiri, 16(1), 29-36.

[23] Spaltini, M., Acerbi, F., Pinzone, M., Gusmeroli, S., & Taisch, M. (2022). Defining the roadmap towards industry 4.0: the 6Ps maturity model for manufacturing SMEs. Procedia CIRP, 105, 631-636.

[24] V. Vachteryte, (2016) Towards an integrated IoT capability maturity model.

[25] Leyh, C., Bley, K., Schäffer, T., & Forstenhäusler, S. (2016, September). SIMMI 4.0-a maturity model for classifying the enterprise-wide it and software landscape focusing on Industry 4.0. In 2016 federated conference on computer science and information systems (fedcsis) (pp. 1297-1302). IEEE.

[26] Schumacher, A., Erol, S., & Sihn, W. (2016). A maturity model for assessing Industry 4.0 readiness and maturity of manufacturing enterprises. Procedia Cirp, 52, 161-166.

[27] Çınar, Z. M., Zeeshan, Q., & Korhan, O. (2021). A framework for industry 4.0 readiness and maturity of smart manufacturing enterprises: a case study. Sustainability, 13(12), 6659.

[28] Qin, J., Liu, Y., & Grosvenor, R. (2016). A categorical framework of manufacturing for industry 4.0 and beyond. Procedia cirp, 52, 173-178.

[29] H. Filho, (2010), Ontology Development 101: AGuide to Creating Your First Ontology.

[30] M. Uschold and M. King, (2011), “Towards a Methodology for Building Ontologies”.

[31] Law, N., Mahmoud, M. A., Tang, A. Y., Lim, F. C., Kasim, H., Othman, M., & Yong, C. (2019). A review of ontology development aspects. International Journal of Advanced Computer Science and Applications, 10(7), 290-298.

Eco-Efficiency Measurement and Regional Optimization Strategy of Green Buildings in China Based on Three-Stage Super-Efficiency SBM-DEA Model

Xianhong Qin*, Yaou Lv, Yunfang Wang, Jian Pi, Ze Xu
HCIG Xiong'an Construction Development Co., Ltd, Xiongan 070001, China

***Abstract*—With the increasing attention of society to sustainable development, green building as an important sustainable building form has attracted much attention. However, the comprehensive assessment of eco-efficiency of green buildings faces many challenges, including the insufficient comprehensive analysis of all stages of the building life cycle and the oversimplification of multidimensional input-output relationships. In addition, the existing methods have subjectivity and uncertainty in data processing and weight allocation, which reduces the reliability of evaluation. To overcome these difficulties, a measurement method based on the three-stage super-efficient data Enveloping analysis (SBM-DEA) model is introduced in this study. By constructing a three-stage super-efficiency SBM-DEA model, the eco-efficiency measurement model of green buildings is established, taking building resources and energy as input and economic and environmental value as output. The results show that after removing the interference of external environment variables and random errors, the measurement results of stage 3 are more reasonable. From 2011 to 2018, the eco-efficiency of green buildings in China showed obvious regional differences, showing a decreasing trend of "the highest in the east (0.884), followed by the central (0.704) and the lowest in the west (0.578)". The innovation of this study lies in the full consideration of timing and dynamics, which provides new theoretical and practical ideas for promoting sustainable development in the field of green building, and is expected to improve the assessment accuracy and reliability in the field of green building.**

Keywords*—Three stages; data envelopment analysis; super efficiency model; green buildings; ecological efficiency

I. Introduction

With the continuous warming of global environmental issues and the urgent need for Sustainable Development (SD), Green Buildings (GBS), as a sustainable building model, have gradually become the focus of attention. SD refers to a model of development that meets the needs of the present without compromising the ability of future generations to meet their own needs. At present, the measurement of Green Building Eco-efficiency (GBEE) is particularly important in evaluating its environmental friendliness and resource utilization, especially in the insufficient comprehensive analysis of various stages of the building lifecycle and the simple handling of 3D input-output relationships [1]. GBEE refers to minimizing environmental impact and improving resource efficiency throughout a building's life cycle through sustainable design, efficient use of resources and environmentally friendly technologies. However, the comprehensive evaluation of the comprehensive benefits of green buildings faces many difficulties and challenges. On the one hand, the existing measurement methods are insufficient in the comprehensive analysis of various stages of the building life cycle, and it is difficult to fully reflect the ecological benefits of green buildings at different stages. On the other hand, the existing methods are too simple when dealing with the multi-dimensional input-output relationship, and it is difficult to accurately evaluate the comprehensive ecological efficiency of green buildings. In addition, the existing methods have subjectivity and uncertainty in data processing and weight allocation, which reduces the reliability and credibility of the evaluation results. The Slacks-Based Measure in Data Envelopment Analysis (SBM-DEA) model based on three-stage super-efficiency combines Data Envelopment Analysis (DEA) and Super Efficiency Model (SEM) [2-3], i.e. 3SE-SBM-DEA model, which can more comprehensively and accurately measure the GBEE. The introduction of the 3SE-SBM-DEA model can better grasp the dynamic characteristics of the building lifecycle and more accurately evaluate the ecological benefits of GBS at different stages [4]. Based on this, this study proposes a GBEE measurement and optimization method based on the 3SE-SBM-DEA model. Firstly, by constructing a 3SE-SBM-DEA model with building resources and energy as inputs and economic and environmental values as outputs, a GBEE measurement model is established. Next, based on theoretical logic, the temporal differences in the GBEE from a temporal dimension are analyzed, and the dynamic evolution characteristics through kernel density estimation (KDE) are revealed. The aim of this study is to achieve comprehensive measurement of GBEE and propose corresponding optimization strategies through the method of this 3SE-SBM-DEA model. The innovation of this method lies in fully considering the temporal and dynamic aspects, providing new theoretical and practical ideas for promoting SD in GB. This paper aims to give more scientific and comprehensive basis for GB design and evaluation, and promote the GB field towards a more sustainable direction.

Section I introduces the research background, problems, and solutions of GBEE measurement. Section II provides a review of previous research on GBEE measurement, exploring

difficulties and shortcomings in methods. Section III is the method of using the 3SE-SBM-DEA model in GBEE measurement. Section IV designs simulation experiments to verify the effectiveness of the proposed method. Section V summarizes the research methods and analyzes the experimental results, pointing out the shortcomings of the methods and future research directions.

II. Related Works

The urgent need for global SD has gradually made GBS a mainstream form of construction that emphasizes resource conservation and environmental friendliness. However, to comprehensively evaluate the comprehensive performance of GBS, a single economic indicator is no longer sufficient. Therefore, researchers are gradually paying attention to GBEE measurement, aiming to comprehensively evaluate its sustainability from two aspects: Resource Utilization Efficiency (RUE) and environmental impact. Ishmael et al. addressed the contribution of buildings to climate change by using an intelligent energy building model based on the Internet of Things (IoT) to connect sensors of building equipment using M2M, IoT, and AEP technologies, achieving intelligent monitoring and improving energy efficiency. IoT intelligent building technology has been proven to be crucial in improving energy efficiency [5]. Tavana M et al. adopted a comprehensive DEA and lifecycle assessment approach to address the negative impact of the Construction Industry (CI) on the environment, particularly in material procurement and emissions, to measure the performance of environmentally friendly building materials in GB management. This method provided a scientific evaluation tool for the selection of GB materials [6]. Zhou Y et al. conducted long-term measurements and surveys of resident satisfaction, combined with environmental energy efficiency analysis, to assess the actual performance of the GB. The measured indoor thermal condition did not fully meet the design goals, especially with differences in relative humidity. However, residents had a higher level of satisfaction with IEQ [7]. Petre and other scholars have proposed a method to accurately determine the actual energy consumption of buildings by in-situ measuring the thermal resistance of building components in response to the high energy consumption problem of the CI in global warming. This study provided practical guidance for improving building energy performance and global warming prevention and control [8]. In response to the challenges encountered in implementing green practices in the chemical industry, Sinaga L et al. proposed an evaluation method that combines blockchain building information modeling (BIM) with structural equation model-Partial least squares (SEM-PLS). The research results show that green practices are becoming more and more common in the manufacturing industry and can reduce the adverse impact on the environment, but the adoption of green principles in industry is affected by a variety of factors [9]. Traditional building materials used in the construction industry significantly contribute to air pollution and greenhouse gas emissions, causing considerable environmental damage across Pakistan. Bashir et al., using closed questionnaires, interviews and observations to collect data using planning and random sampling techniques, focused on exploring the feasibility of adopting green building materials in Pakistan's building sector with the aim of mitigating environmental impacts. The results of the study show that factors such as high cost, low market demand and logistical challenges limit people's interest in environmentally friendly materials, with 73% of construction companies in Pakistan not using green building materials [10]. In view of the negative impact of the construction industry on the environment and the problems of resource depletion, emission and biodiversity loss, Kristinayanti W S et al. proposed a method combining local wisdom and green building practices, adopted the PRISMA framework method, and conducted a comprehensive systematic evaluation and qualitative analysis through NVivo software. The research results show that the construction industry needs a sustainable transformation, and combining local wisdom can provide innovative and adaptable solutions to help promote the transformation of construction practices to SD [11].

In addition, the 3SE-SBM-DEA combines the advantages of DEA and SEM to assess the relative efficiency of various units. In this model, the projection pursuit method is used to determine the unit's super efficiency boundary, so that an optimal super efficiency frontier can be found under certain constraints. Jiahui et al. utilized the 3SE-SBM-DEA to address the CO2 efficiency issues in the four Chinese major beef-cattle production areas in, and incorporated CO2 into the efficiency calculation framework. External random disturbances had greatly affected the efficiency measurement, and using the 3SE model made the results more in line with reality [12]. Junlong et al. constructed an indicator system to promote the high-quality development of China's shared manufacturing industry and used a 3-phase DEA-Malmquist model to dynamically measure 39 shared manufacturing enterprises from 2018 to 2020. The mean fluctuation of comprehensive/pure/scale efficiency had a significant impact on development efficiency [13]. Radimov N et al. proposed a novel control and optimization strategy for a bidirectional 3-level in vehicle battery charger (OBC) that achieves 80 PLUS titanium efficiency. OBC could change the direction of power flow within a few msec, providing reactive power support for the power grid, with 96.65% peak efficiency and 1% min-total harmonic distortion [14]. Qad et al. measured the relationship between technology industry agglomeration, green innovation, and development quality in the Yangtze River Delta urban agglomeration using superefficient SBM-DEA and improved TOPSIS method. There was a significant spatial connection between these factors, especially in the transformation stage where they mutually promote each other [15].

To sum up, the existing research has made remarkable progress in the field of eco-efficiency measurement of green buildings, but there are still some limitations. For example, although some studies have proposed comprehensive evaluation methods, they are still insufficient in dealing with the dynamic characteristics of the whole life cycle of buildings and the multi-dimensional input-output relationship. In addition, there are subjectivity and uncertainty in data processing and weight allocation in existing studies, which reduces the reliability and applicability of evaluation results. Although some studies try to solve these problems by

introducing advanced technical means or improving evaluation models, most studies fail to fully consider regional differences and dynamic evolution characteristics, resulting in the overall assessment of green building eco-efficiency is still insufficient. The DEA method is also adopted as the basic framework in this study. However, the innovation of this study lies in the introduction of the three-stage super-efficiency SBM-DEA model, which can not only evaluate the eco-efficiency of green buildings more comprehensively, but also better deal with data uncertainty and subjectivity by introducing the concepts of super-efficiency and stages. Improve the objectivity and robustness of the evaluation. In addition, this study also combined temporal dimension analysis and kernel density estimation to reveal the dynamic evolution characteristics of green building eco-efficiency, which was rarely involved in previous studies.

The contributions of this research are mainly reflected in the following aspects: First, by constructing a three-stage super-efficiency SBM-DEA model, this research provides a more scientific and comprehensive method for measuring the eco-efficiency of green buildings, which can effectively overcome the limitations of existing methods in data processing and weight allocation. Secondly, this study systematically analyzed the regional differences of eco-efficiency of green buildings in China for the first time, revealing the decreasing trend of "the highest in the east, the second in the central, and the lowest in the west" during 2011-2018, providing a scientific basis for formulating targeted optimization strategies. Finally, based on the measurement results, this study proposed specific optimization strategies to narrow the regional gap and improve the overall ecological efficiency of green buildings across the country.

III. Construction of 3SE-SBM-DEA Measurement Model

This study first constructs a 3SE-SBM-DEA model, using building resources and energy as inputs and economic and environmental values as outputs, to establish a GBEE measurement model. Next, based on theoretical logic, to analyze the temporal differences of GBEE from the temporal dimension, and to reveal the dynamic evolution characteristics through KDE.

A. Building the GBEE Measurement Model

Eco-efficiency is the efficiency of ecological resources to meet human needs, usually measured by the ratio of output to input [16]. Among them, "output" covers the value of the products and services produced by the enterprise, while "input" includes the resources and energy consumed by the enterprise, as well as the impact on the environment [17]. The mathematical formula related to ecological efficiency is Eq. (1).

$$\text{Eco-e} = \frac{\text{Value of p or s}}{\text{Ei}} \quad (1)$$

In Eq. (1), Eco-e represents ecological efficiency, and Value of p or s represents the value of the product or service. Ei represents environmental impact. GBEE refers to the use of sustainable design, efficient resource utilization, and environmental protection technologies to minimize the impact on the environment during the building lifecycle, increase RUE, and reduce the burden on natural ecosystems [18]. Fig. 1 shows the ecological efficiency evaluation model.

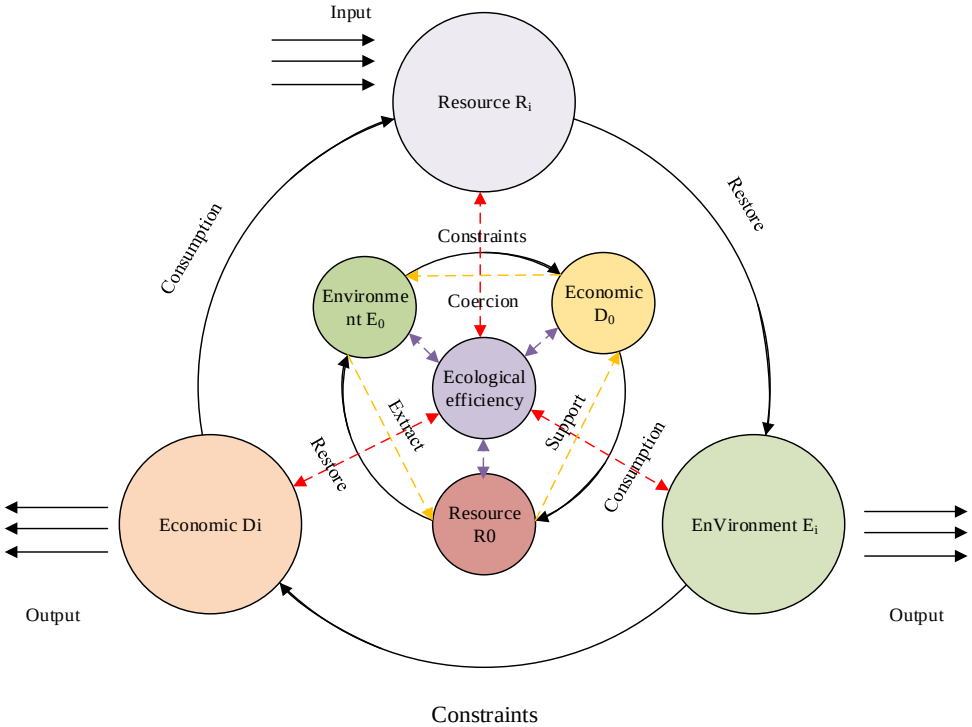


Fig. 1. Ecological efficiency evaluation model.

In Fig. 1, the model mainly revolves around economy, resources, and environment [19]. To evaluate ecological efficiency, it is necessary to first determine key indicators such as energy consumption, material utilization, and water resource utilization. Subsequently, by collecting relevant data from the system, including information on building energy usage, material sources and utilization, water resource management, etc., a comprehensive data foundation is established. Next, a mathematical model is used to quantify the system efficiency. Finally, based on the evaluation results, optimization suggestions are established to lift the Eco-efficiency. To derive the Eco-efficiency calculation formula for buildings based on Eq. (1), as shown in Eq. (2).

$$\text{B eco-e} = \frac{\text{Be}}{\text{El}} \tag{2}$$

In Eq. (2), B eco-e represents the ecological benefits of the building. B eco-e represents the value of the building. El represents environmental load. The GBEE measurement indicator system aims to comprehensively understand the comprehensive ecological benefits of buildings throughout their lifecycle by evaluating their performance in energy utilization, material selection, water resource utilization, environmental impact of design, and indoor environmental quality. Fig. 2 shows the GBEE measurement indicator system.

In Fig. 2, the measurement indicator system of GBEE is mainly divided into two parts: input indicators and output indicators. The investment indicators include capital investment, labor investment, energy investment, land investment, and technology investment. Output indicators include environmental output and economic output. This set of specific parameters is selected to ensure that the eco-efficiency measurement of green buildings based on the three-stage super-efficiency SBM-DEA model can fully and accurately reflect the actual ecological benefits of green buildings, while eliminating the interference of external environment and random errors, and improving the reliability and applicability of the model results. SBM is a distance function based method used to evaluate performance relative to other units, suitable for considering multiple input and output factors, and able to handle different weights and measurement standards. DEA is a non-parametric approach taken to evaluate relative efficiency. DEA does not require prior assumptions about weights or function forms, making it suitable for complex multi input multi output scenarios [20]. Fig. 3 shows the specific process of the 3-stage DEA.

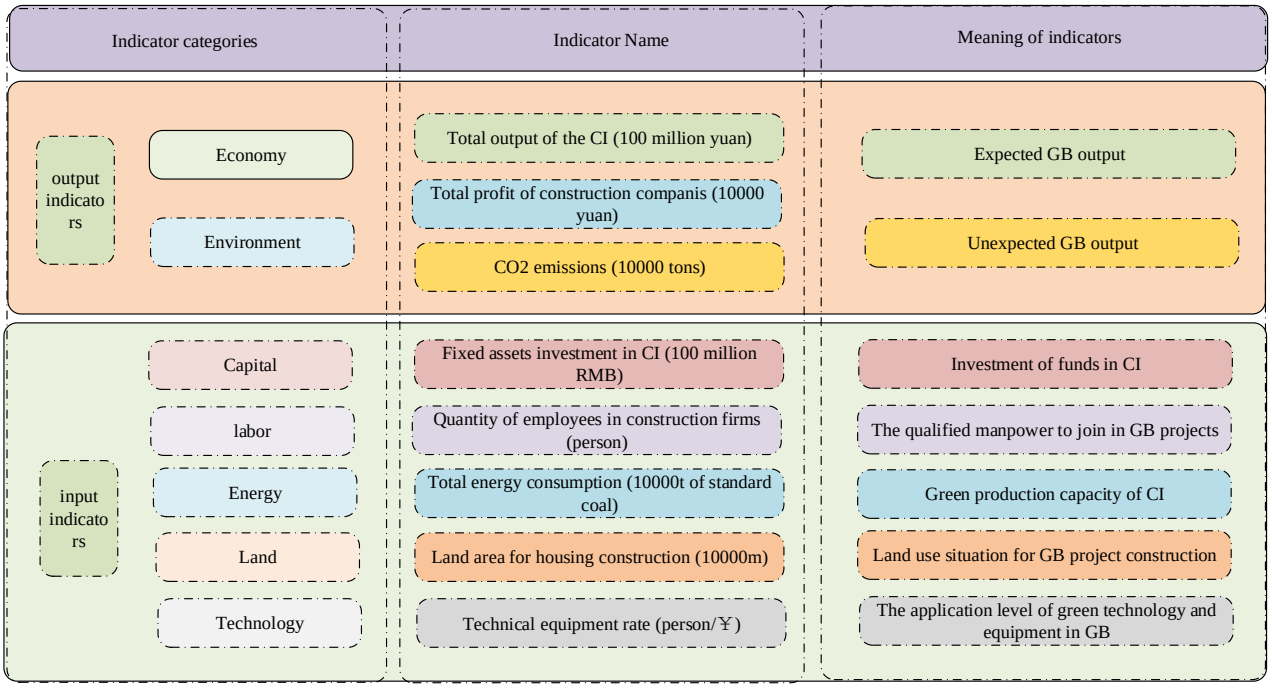


Fig. 2. Index system of GBEE measurement.

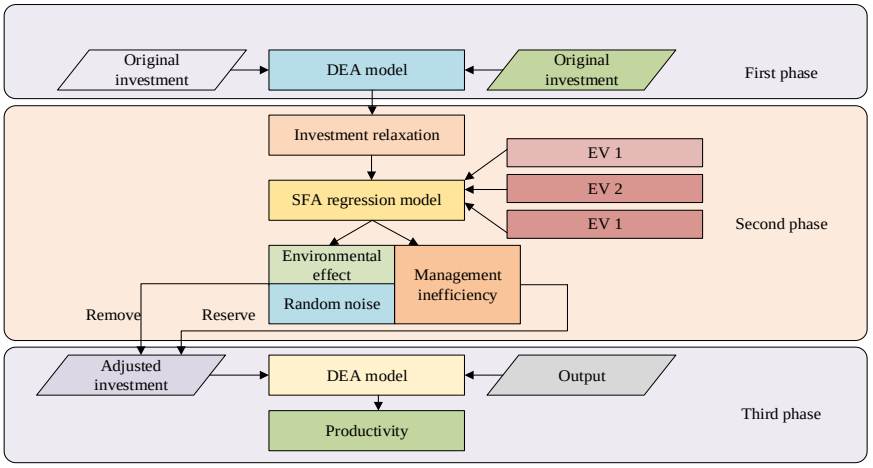


Fig. 3. Specific process of the three-stage DEA.

In Fig. 3, the first phase of constructing the three-stage DEA is the construction of the SBM model. The 2nd phase is to construct a stochastic frontier model. The third stage is based on the foundation of stages one and two, to obtain more accurate measurement values that reflect the GBEE levels of all DEAs [21]. Specifically, the first step is to construct an ultra efficient SBM model. This study adopted an input-oriented model to test the initial efficiency of GBEE. The improved super-efficient SBM model is obtained by optimizing the objective function based on relaxation variables. The process is shown in Eq. (3).

$$\rho = \min \frac{\frac{1}{m}\sum_{i=1}^{m}\frac{\bar{x}_i}{x_{i0}}}{\frac{1}{s_1+s_2}\left(\sum_{r=1}^{s_1}\frac{\bar{y}_r^a}{y_r^a}+\sum_{j=1}^{s_2}\frac{\bar{y}_j^b}{y_{j0}^b}\right)}$$

$$\text{s.t.}\begin{cases} x_0 = X\lambda + S^-, \quad y_0^a = Y^a\lambda - S^a, \quad y_0^b = Y^b\lambda - S^b \\ \bar{x} \geq \sum_{j=1,\neq 0}^{n}\lambda_j x_j, \quad \bar{y}^a \leq \sum_{j=1,\neq 0}^{n}\lambda_j y_j^a, \quad \bar{y}^b \leq \sum_{j=1,\neq 0}^{n}\lambda_j y_j^b \\ \bar{x} \geq x_0, \quad \bar{y}^a \leq y_0^a, \quad \bar{y}^b \geq y_0^b \\ \sum_{j=1,\neq 0}^{n}\lambda_j = 1, \quad S^- \geq 0, \quad S^a \geq 0, \quad S^b \geq 0, \quad \bar{y}^g \geq 0, \quad \lambda \geq 0 \end{cases} \tag{3}$$

In Eq. (3), m represents the number of input indicators. n represents Decision-making Unit (DMU). x represents the input item. λ and ρ represent weight vectors and objective function values, respectively. y_a and y_b represent expected and unexpected output items, respectively, with S_1 and S_2 as the number of indicators. S^- is the input, S^a is , expected output, and S^b is the unexpected output. After constructing the SBM, a random frontier model is then constructed [22]. By measuring GBEE, the original efficiency data and relaxation variable of each DMU can be obtained. The process function model is Eq. (4).

$$S_{ni} = f\left(Z_i;\beta^n\right) + v_{ni} + \mu_{ni} \quad ;i = 1,2,\ldots,I;n = 1,2,\ldots,N \tag{4}$$

In Eq. (4), i and n represent the number of DMUs and the input items, respectively. S_{ni} represents the input relaxation value (IRV) of n of the i -th DMU. Z_i represents P Environmental Variables (EVs). β^n represents the estimated parameter of i . $f\left(Z_i;\beta^n\right)$ represents the impact of external EV on the IRV. $v_{ni}+\mu_{ni}$ is the mixed-error, where v_{ni} is the Random Error (RE). To eliminate the influence of external EVs and REs, homogeneous adjustments are made to each input quantity. The specific adjustment method can refer to Eq. (5).

$$X_{ni}^A = X_{ni} + \left[\max f\left(Z_i;\hat{\beta}^n\right) - f\left(Z_i;\hat{\beta}^n\right)\right] + \left[\max\left(v_{ni}\right) - v_{ni}\right] \tag{5}$$

In Eq. (5), this study homogenized the input of each DMU and obtained the adjusted input value X_{ni}^A. By maximizing $\max f\left(Z_i;\hat{\beta}^n\right) - f\left(Z_i;\hat{\beta}^n\right)$, this study adjusts the external environment to the same state. Among them, $\max f\left(Z_i;\hat{\beta}^n\right)$ is the benchmark adjusted by other DMUs, indicating the worst environmental conditions. This adjustment takes into account both good and bad conditions, and increases or decreases investment. To adjust the RE to the same state through $\max\left(v_{ni}\right) - v_{ni}$, ensuring that all DMUs are in the similar external EV and RE status.

B. GBEE Based on Temporal Dimension

After establishing a scientific measurement index system, this study further analyzes the temporal changes of GBEE through the temporal dimension based on national, regional, and inter provincial differences, and uses KDE to reveal its dynamic evolution characteristics [23]. On the spatial dimension, the spatial distribution pattern of GBEE is visualized, and the spatial agglomeration and transition characteristics of GBEE in various provinces and cities through spatial auto-correlation (SAC) analysis are revealed. The specific technical road-map is Fig. 4.

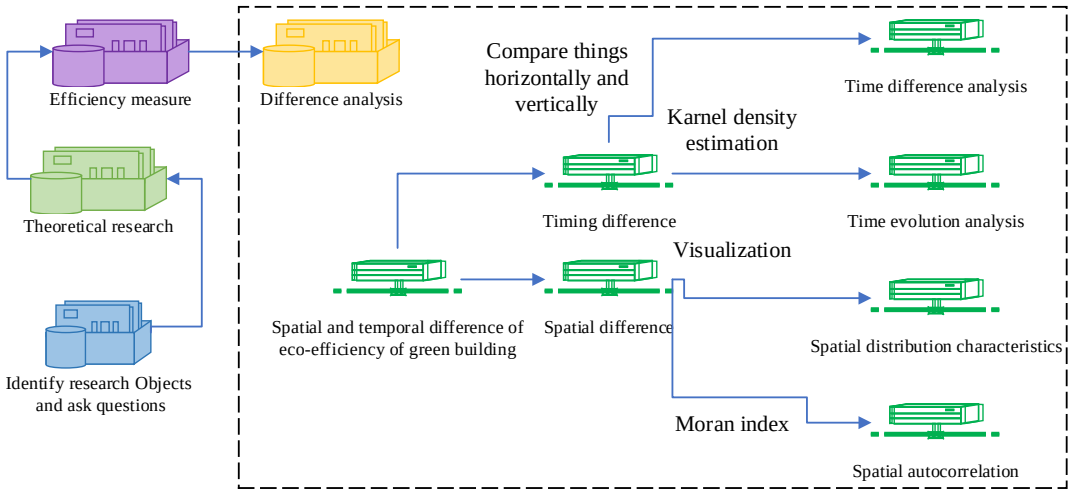


Fig. 4. Technology road-map.

In Fig. 4, the spatiotemporal differences of GBEE are divided into temporal differences and spatial differences. Kernel density estimation is a non-parametric statistical method utilized to estimate the shape of the probability density function. This method involves placing kernel functions (KerF) around each data-point and then overlaying these functions to form an estimate of the overall probability density. The specific expression is Eq. (6).

$$f(x)=\frac{1}{Nh}\sum_{i=1}^{N}K\left(\frac{X_i-\bar{x}}{h}\right) \tag{6}$$

In Eq. (6), bandwidth h is an important parameter. The set of observation samples is represented by N, and the values of independent and identically distributed observation samples are represented by X_i. x represents the input item. The K is a weighted or smooth transformation function. The mathematical formula that needs to meet the conditions is Eq. (7).

$$\begin{cases}\lim\limits_{x\to\infty}K(x)\cdot x=0\\ K(x)\geq 0 \quad \int_{-\infty}^{+\infty}K(x)dx=1\\ \sup K(x)<+\infty \quad \int_{-\infty}^{+\infty}K^2(x)dx<+\infty\end{cases} \tag{7}$$

KerFs generally include trigonometric KerFs, quadrilateral KerFs, and Gaussian KerFs. This study chose to use Gaussian density KerF, as shown in Eq. (8).

$$K(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \tag{8}$$

Building a spatial weight matrix is a commonly used task in spatial analysis, which is used to describe the degree of correlation between adjacent regions in geographic space. Construct a binary symmetric spatial matrix W_{n*n} of n*n to represent the spatial adjacency relationship between n positions. The matrix is specifically expressed as Eq. (9).

$$W_{n\times n}=\begin{bmatrix}w_{11} & w_{12} & \cdots & w_{1n}\\ w_{21} & w_{22} & \cdots & w_{2n}\\ \cdots & \cdots & \cdots & \cdots\\ w_{n1} & w_{n2} & \cdots & w_{nn}\end{bmatrix} \tag{9}$$

In Eq. (9), $W_{i\times j}$ represents the proximity relationship between region i and j. This study uses the spatial weight matrix under the geographical distance standard, as expressed in Eq. (10).

$$w_{ij}=\begin{cases}1, & \text{The distance between region i and j is<d}\\ 0, & \text{others}\end{cases} \tag{10}$$

Global SAC is a method taken to analyze the spatial correlation between geographic units in an entire region or system. It mainly focuses on the global distribution pattern of variables within the entire region to reveal spatial clustering or dispersion trends. This study uses a geographic distance spatial weight matrix and uses the global Moran's I index (MII) for measurement, as shown in Eq. (11).

$$I=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}\left(x_i-\bar{x}\right)\left(x_j-\bar{x}\right)}{S^2\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}} \tag{11}$$

In Eq. (11), n is the amount of provinces and cities. x_i and x_j represent the GBEE values of province and city i and j. $\bar{x}$ and S^2 represent the mean and variance of GBEE in each province and city. Further to adopt the Z-score normal distribution hypothesis to verify the accuracy of MII. Its expression is Eq. (12).

$$Z(I)=\frac{I-E(I)}{\sqrt{\mathrm{VAR}(I)}} \tag{12}$$

In Eq. (12), $E(I)$ is the expected value. $\mathrm{VAR}(I)$

represents the expected variance. The specific expression of $E(I)$ is Eq. (13).

$$E(I) = -\frac{1}{n-1} \tag{13}$$

Local SAC is usually measured by Local Moran's I. This paper uses local MII and Moran scatter plot to evaluate the local distribution characteristics of GBEE, as shown in Eq. (14).

$$I_i = \frac{(x_i - \bar{x})}{S^2} \sum_{j=1}^{n} W_{ij} (x_j - \bar{x}) \tag{14}$$

The spatial lag model is adopted to spatial data analysis. It considers spatial correlation, which is commonly used to describe the interactions and dependencies between spatial data, as shown in Eq. (15).

$$Y = \rho WY + X\beta + \mu \tag{15}$$

The relationship between GBEE Y and the influencing factor matrix X was studied in Eq. (15). Considering spatial interaction, spatial auto-regressive coefficient ρ and spatial weight matrix are introduced. The RE is represented by μ, while the spatial lag term WY reflects the influence of adjacent regions.

As a comprehensive efficiency evaluation method, the core assumptions of 3SE-SBM-DEA model mainly include the following points: First, the model assumes that there is a clear linear relationship between input and output, that is, under given technical conditions, the increase of input will lead to the increase of output in a certain proportional relationship. Secondly, the 3SE-SBM-DEA model assumes that the effects of external environment variables and random errors on efficiency are independent and can be separated and adjusted by appropriate statistical methods. In addition, the model also assumes that the data is accurate and complete, that is, the input and output data can truly reflect the actual operation of green buildings. Finally, the model assumes that each DMU is homogeneous in terms of technical conditions and production functions, that is, all evaluated green building projects are comparable at the technical level.

IV. Performance Analysis and Validation of the GBEE Measurement Model

This study first analyzes the statistical characteristics of the sample data and the assumption of homogeneity of measurement indicators, confirming the applicability of the measurement model. Subsequently, using indicator data, the GBEE in China is measured, and the efficiency levels and variation differences of Stage One and Stage Three at the national, regional, and inter provincial levels are compared and analyzed.

A. GBEE Analysis

This study uses panel data on input-output and EVs from 2011 to 2018, selecting 30 provinces in China as research subjects, with the aim of calculating the GBEE of each province and city. To better analyze its regional differences, the research area is segmented into eastern, central, and western regions. Table I presents the statistical results.

According to Table I, through statistical analysis of the collected data, it is found that the SD and range among individual indicators are huge, showing a significant difference, and the input-output situation also shows a significant difference. Before conducting the GBEE measurement, the measurement model requires that the input and output items meet the assumption of homogeneity, that is, the principle that "as the input increases, the output does not decrease.". To verify this hypothesis, this study conducts Pearson correlation tests between input indicators and output indicators using SPSS 20.0 software. Table II shows the test results.

In Table II, && is significantly correlated at the 1% level (bilateral). In Table II, the correlation coefficients are all positive and have all passed the bilateral test at the 1%, indicating that the input and output variables satisfy the principle of homogeneity assumption. This study further adopts the super efficiency SBM model and uses MAXDEA software to run GBEE input-output data, obtaining the GBEE levels of each province and city without considering the influence of external EVs and REs. The specific results are shown in Fig. 5.

TABLE I. Descriptive Statistical Results of Sample Data

Variable	Minimum	Mean	Maximum	SD	Sample
Housing construction land area	738	39362	249176	47499	240
Employees in construction company	54847	1604115	8110275	1785021	240
Total energy consumption	14	126	354	1785021	240
Rate of technical equipment	728	14354	91231	9978	240
Carbon emission	562	13178	52901	11055	240
Gross output of CI	52	1211	6717	1302	240
Investment in fixed assets in the CI	0.09	131	1136	201	240
Total profits of construction enterprises	64449	2093927	11617738	2070219	240

TABLE II. PEARSON CORRELATION TEST BETWEEN INPUT AND OUTPUT VARIABLES

Type	Index	Construction Enterprise Number of Employees	House Construction Land Area	Investment in Fixed Assets in CI	Energy Consumption Gross Amount	Technology Equipment Rate
Gross output value of CI	Significance (bilateral)	0.000	0.000	0.003	0.001	0.001
	Pearson correlation	0.948&&	0.974&&	0.626&&	0.566&&	0.603&&
Gross profit	Significance (bilateral)	0.000	0.000	0.003	0.001	0.002
	Pearson correlation	0.924&&	0.922&&	0.601&&	0.555&&	0.525&&
Carbon emission	Significance (bilateral)	0.000	0.000	0.002	0.004	0.001
	Pearson correlation	0.887&&	0.816&&	0.667&&	0.511&&	0.755&&

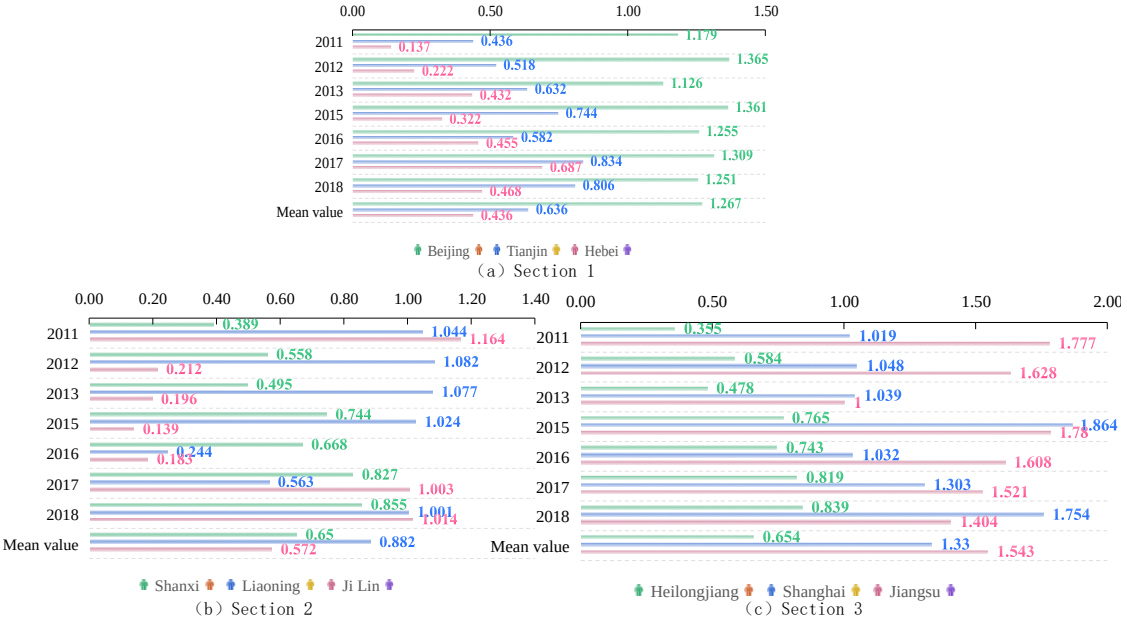


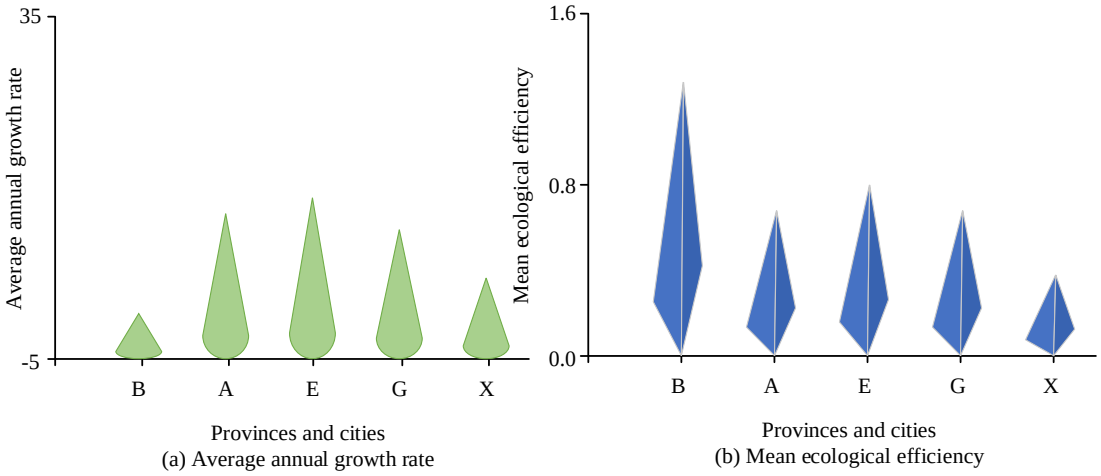
Fig. 5. Eco-efficiency level of green buildings.

Fig. 5 (a) shows the GBEE levels in provinces and cities B, T, and H. Annual Growth Rate (AGR) is used to describe the average annual growth of an indicator in a specific period of time. Fig. 5 (b) shows the GBEE levels in S, L, and J provinces. Fig. 5 (c) shows the GBEE levels of provinces and cities A, C, and D. Fig. 5 shows that from 2011 to 2018, China's GBEE showed a good development trend, with an overall average efficiency increasing trend, with an average AGR of 4.71%. After 2015, ecological efficiency achieved sustained and steady growth, reaching a peak of 0.916 in 2018, with an average AGR of 14.5%. The average and AGR of GBEE in some provinces and cities in China are shown in Fig. 6.

Fig. 6 (a) shows the AGR, and Fig. 6 (b) shows the Mean Ecological Efficiency (MEE). There are obvious discrepancies in GBEE among provinces. The MEE of province and city A is below 0.7, with an average AGR of 13.07%. Provinces and cities such as E show fluctuating fluctuations, with no significant increase or decrease trend. Fig. 7 shows the comparison of the MEE and average AGR of GBEE among different regions in China.

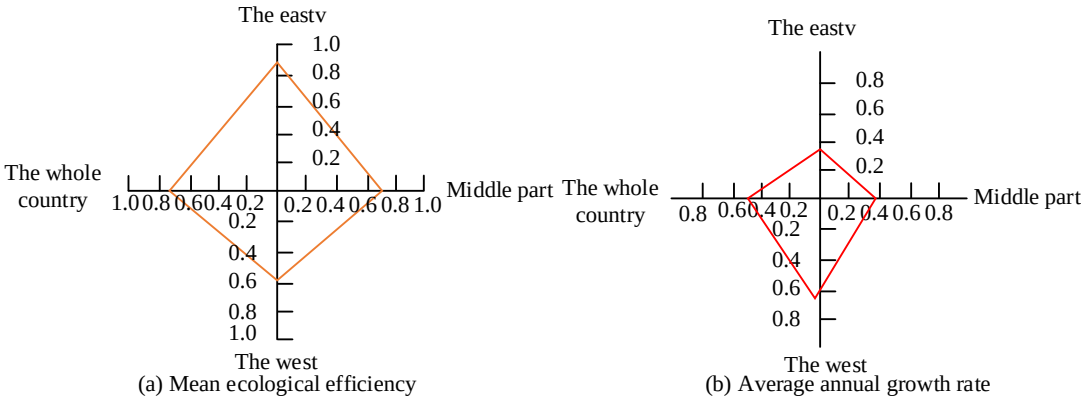
Fig. 7 (a) is the MEE, and Fig. 7 (b) shows the AGR. Between 2011 and 2018, China's GBEE showed significant regional differences, with the highest in the East (0.884), followed by the central (0.704), and the lowest in the West (0.578). Fig. 8 shows the trend of changes in the mean GBEE across different regions.

Fig. 8 (a) shows the eastern and central regions, while Fig. 8 (b) shows the national and western regions. Fig. 8 shows that the GBEE in the East showed fluctuating patterns from 2011 to 2018, consistently higher than the national average, with no significant increase or decrease trend. In contrast, the efficiency curve in the central region shows a similar development trend to the national average curve, with an overall low growth rate. Although the three major regions reached their peak efficiency in 2018, the average AGR in the West reached 6.99%, which is 48.84% and 41.33% higher than that in the East and Center, respectively.



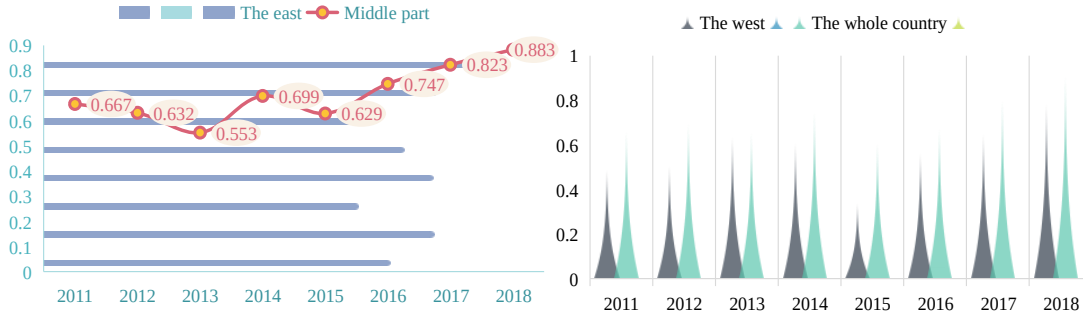
(a) Average annual growth rate (b) Mean ecological efficiency

Fig. 6. MEE and average AGR of GBEE in China.



(a) Mean ecological efficiency (b) Average annual growth rate

Fig. 7. Comparison of MEE and average AGR of GBEE in different Chinese regions.



(a) The east and Middle part (b) The west And The whole country

Fig. 8. Variation trend of mean value of GBEE in each region.

B. *Performance Verification Based on the 3SE-SBM-DEA Model*

After the SFA regression analysis in stage two, it is found that there are significant differences in the impact of external environmental differences and GB investment factors among provinces and cities, which in turn have a significant impact on GBEE. To eliminate these impacts, adjustments are made to the input variables to ensure that each province and city are compared under the equal external EV and RE. The final GBEE is obtained by using the SBM model and MAXDEA software for efficiency analysis. The GBEE mean measurement results for Stage 1 and Stage 3 are shown in Fig. 9.

Fig. 9 (a) and Fig. 9 (b) show the results of the first and third stages. By comparing Fig. 9 (a) and 9 (b), the MEE of the three stages is lower than that of the first stage, except for 2015 and 2016. Fig. 10 further illustrates the comparison of GBEE mean values.

Fig. 10 (a) shows the results of the first stage, and Fig. 10 (b) shows the results of the third stage. After excluding the external EVs and REs influences, GBEE still maintains a

pattern of "greater in the east than in the central and greater in the west", but the regional gap has widened. The efficiency in the East outperforms than the national average, while that in the Center and West has decreased. From 2011 to 2018, there is a slight decrease in the national average GBEE, indicating that there is still significant room for improvement in GBEE.

In order to verify the robustness of the eco-efficiency measurement results of green buildings based on the three-stage super-efficiency SBM-DEA model under different conditions, sensitivity analysis was conducted. By adjusting the key parameters of the model, the influence of these changes on the eco-efficiency of green buildings was analyzed, so as to evaluate the robustness of the model results. The details are shown in Table III.

In Table III, the eco-efficiency measurement results of green buildings show good robustness under different conditions. When the weight adjustment of input-output index is ±10%, the mean eco-efficiency of eastern, central and western regions decreases by 0.008, 0.006 and 0.008 respectively, and the mean eco-efficiency of the whole country decreases by 0.005. This indicates that weight adjustment has a certain impact on the measurement results of eco-efficiency, but the overall change range is small, indicating that the model is less sensitive to weight. When the adjustment amplitude of the influence of external environmental variables is ±15%, the mean eco-efficiency of eastern, central and western regions increases by 0.008, 0.008 and 0.008 respectively, and the national mean eco-efficiency increases by 0.008. This indicates that external environmental variables have a significant impact on the eco-efficiency measurement results, but the adjusted results still maintain the original regional difference pattern, that is, "East > central > west". When the random error adjustment amplitude is ±20%, the mean eco-efficiency in eastern, central and western regions decreases by 0.004, 0.004 and 0.004 respectively, and the national mean eco-efficiency decreases by 0.003. The adjustment of random error has relatively little effect on the eco-efficiency measurement results, which further verifies the robustness of the model results.

TABLE III. SENSITIVITY ANALYSIS RESULT

Parameter adjustment type	Adjustment range	Average ecological efficiency in eastern China	Average ecological efficiency in central region	Average ecological efficiency in western China	Average national ecological efficiency
Reference model	/	0.884	0.704	0.578	0.723
Input-output index weight adjustment	±10%	0.876	0.698	0.570	0.718
External environment variables affect adjustment	±15%	0.892	0.712	0.586	0.731
Random error adjustment	±20%	0.880	0.700	0.574	0.720

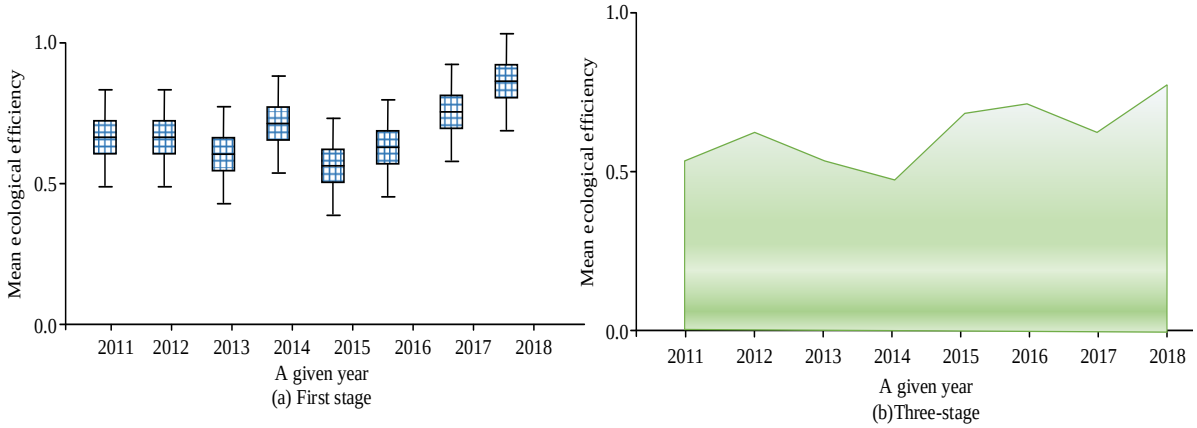


Fig. 9. Average measurement results of GBEE in stage 1 and stage 3.

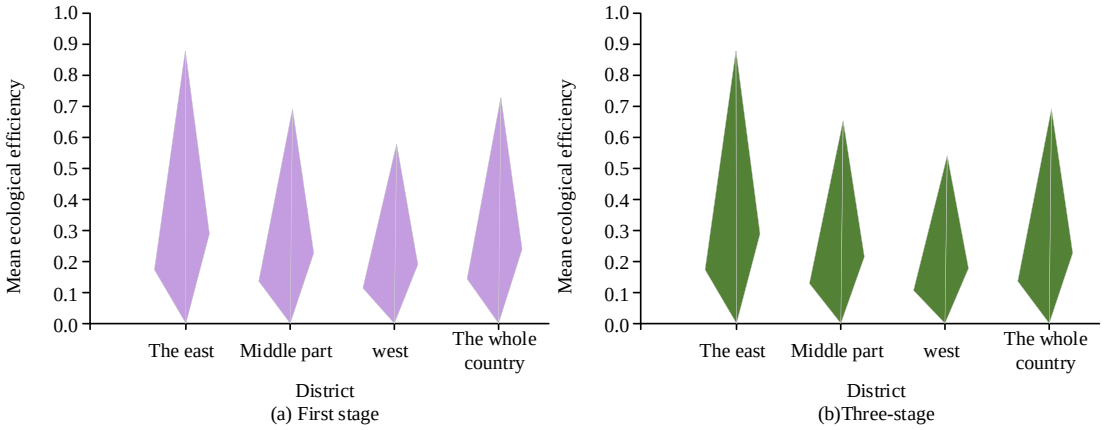


Fig. 10. Comparison of mean GBEE.

V. Optimization Strategy Discussion

The results showed that the eco-efficiency of green buildings in China showed obvious regional differences during 2011-2018, with the highest efficiency in the eastern region, followed by the central region and the lowest in the western region. The formation of regional differences is closely related to various factors such as economic development level, technological input, policy support and resource endowment. Therefore, according to the characteristics of different regions, the following optimization strategies are proposed in order to narrow the regional gap and improve the overall ecological efficiency of green buildings in the country. For the eastern region, although its green building ecological efficiency is at a high level, there is still room for further improvement. The eastern region has a relatively high level of economic development and relatively abundant technology and capital, so it should focus on strengthening the innovation and application of green building technology. The eco-efficiency of green buildings in the central region is at a medium level, and its optimization strategy should focus on the combination of technology introduction and talent training. The central region has certain advantages in terms of resources and market, but it is relatively short of technology and talents. Therefore, we should actively introduce advanced green building technology and management experience in the eastern region, and promote the popularization and application of green building technology in the central region through technical cooperation and project demonstration. The eco-efficiency of green buildings in western China is relatively low, and its optimization strategy should focus on solving the problems of weak infrastructure and shortage of funds. The economic development of the western region is relatively backward, and the infrastructure construction is insufficient. We should increase the investment in the construction of green building infrastructure and improve the basic conditions for the development of green building. The government should increase financial support for green building projects in the western region through financial transfer payments and special funds. At the same time, the western region should give full play to its own resource advantages, develop green buildings according to local conditions, increase the application proportion of renewable energy in buildings, and reduce the dependence on traditional energy.

VI. Conclusion

The continuous attention of society to SD has attracted attention to GBS as a key form of sustainable building. To comprehensively evaluate the comprehensive benefits of GBS, the focus is gradually shifting towards their ecological efficiency. In this context, this study constructed a measurement method based on the 3SE-SBM-DEA model, with building resources and energy as inputs and economic and environmental values as outputs, and established the GBEE measurement model. The results showed that from 2011 to 2018, China's GBEE saw overall growth with an average AGR of 4.71%. From 2011 to 2015, there was an M-shaped oscillation, reaching the lowest value of 0.611. After excluding the influence of external EVs and REs, the measuring data of stage three were more reasonable. From 2011 to 2018, China's GBEE showed a decreasing trend, with the highest in the East (0.884), the sequential Center (0.704), and the lowest in the West (0.578). The efficiency in the eastern area was higher than the national average, while the efficiency in the central and western areas has decreased. This study has made progress in GBEE measurement and optimization methods, but there are limitations to the data, uncertainty in model parameter selection, and important factors that have not been considered. Future research should focus on expanding the sample range, improving model parameter selection methods, improving data quality and reliability, and comprehensively considering more factors to promote the development of this field.

Funding and Acknowledgment

None.

Competing Interests

The author(s) declare none.

Data Availability Statement

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

REFERENCES

[1] Lin B, Chen H, Liu Y, He Q, Li Z. A preference-based multi-objective building performance optimization method for early design stage. Building Simulation, 2021, 14(3): 477-494.

[2] Badal P S, Sinha R. A multi-objective performance-based seismic design framework for building typologies. Earthquake engineering & structural dynamics, 2022, 51(6): 1343-1362.

[3] Pal S, Roy A, Shivakumara P, Pal U. Adapting a Swin Transformer for License Plate Number and Text Detection in Drone Images. Artificial Intelligence and Applications, 2023, 1(3), 145-154.

[4] Hebbi C, Mamatha H. Comprehensive Dataset Building and Recognition of Isolated Handwritten Kannada Characters Using Machine Learning Models. Artificial Intelligence and Applications, 2023, 1(3): 179-190.

[5] Ishmael A, Ogara S, Raburu G. Developing and Validating a Model for the Adoption of Internet of Things Based Smart Building. World Journal of Innovative Research, 2020, 9(1): 87-100.

[6] Tavana M, Izadikhah M, Saen R F. An integrated data envelopment analysis and life cycle assessment method for performance measurement in green construction management. Environmental Science and Pollution Research, 2021, 28(1): 664-682.

[7] Zhou Y, Cai J, Xu Y. Indoor environmental quality and energy use evaluation of a three-star green office building in China with field study. Journal of Building Physics, 2021, 45(2): 209-235.

[8] Petre S G, Isopescu D, Pruteanu M. Study on the Factors Affecting in Situ Measurement of the Thermal Resistance of Building Elements. Bulletin of the Polytechnic Institute of Iași. Construction. Architecture Section, 2021, 67(1): 87 - 94.

[9] Sinaga L, Husin A E, Kristiyanto K, Arif E J, Pinem M P. Blockchain-Building Information Modeling (BIM): Aplication in Cost Eficiency of Retrofitting Green Chemical Industrial Buildings. Computational Engineering and Physical Modeling, 2023, 6(2): 1-22.

[10] Bashir M T, Khan A B, Khan M M H, Rasheed K, Saad S, Farid F. Evaluating the implementation of green building materials in the construction sector of developing nations. Journal of Human, Earth, and Future, 2024, 5(3): 528-542.

[11] Kristinayanti W S, Zaika Y, Devia Y P, Wibowo M A. Green Construction and Local Wisdom Integration for Sustainability: A Systematic Literature Review. Civil Engineering Journal, 2024, 10(11): 3779-3802.

[12] Jiahui Yan, Yuejie Zhang. Spatio-temporal evolution characteristics and spatial distribution pattern of carbon emission efficiency in main beef cattle producing areas in China. Scientia Geographica Sinica, 2023, 43(5): 879-888.

[13] Junlong C, Qiu T. Research on the Efficiency Measurement of High-quality Development of Sharing Manufacturing in China Based on Three-stage DEA-Malmquist Method. Industrial Technology & Economy, 2022, 41(3): 106-115.

[14] Radimov N, Li G, Tang M. Three-stage SiC-based bi-directional on-board battery charger with titanium level efficiency. IET Power Electronics, 2020, 13(7): 1477-1480.

[15] Qad. Technological industry agglomeration, green innovation efficiency, and development quality of city cluster. Green Finance, 2022, 4(4): 411-435.

[16] Cheng H, Yu Y, Zhang S. Subsidies, green innovation, and the sustainable performance: evidence from heavy-polluting enterprises in China. Journal of Environmental Studies and Sciences, 2023, 14(1): 102-116.

[17] Leung K, Aréchiga N, Pavone M. Backpropagation through signal temporal logic specifications: Infusing logical structure into gradient-based methods. The International Journal of Robotics Research, 2023, 42(6): 356-370.

[18] Choudhuri S, Adeniye S, Sen A. Distribution Alignment Using Complement Entropy Objective and Adaptive Consensus-Based Label Refinement For Partial Domain Adaptation Artificial Intelligence and Applications. 2023, 1(1): 43-51.

[19] Long X M, Chen Y J, Zhou J. Development of AR Experiment on Electric-Thermal Effect by Open Framework with Simulation-Based Asset and User-Defined Input Artificial Intelligence and Applications. 2023, 1(1): 52-57.

[20] Mansir I B, Hani E H B, Ayed H, Diyoke C. Dynamic simulation of hydrogen-based zero energy buildings with hydrogen energy storage for various climate conditions. International journal of hydrogen energy, 2022, 47(62): 26501-26514.

[21] Qiao Z, Shan W, Jiang N, Heidari A A, Chen H, Teng Y. Gaussian bare-bones gradient-based optimization: Towards mitigating the performance concerns. International Journal of Intelligent Systems, 2022, 37(6): 3193-3254.

[22] Kim J, Kwon D, Woo S Y, Kang W M, Lee J H. Hardware-based Spiking Neural Network Architecture Using Simplified Backpropagation Algorithm and Homeostasis Functionality. Neurocomputing, 2020, 428(38). 153-165.

[23] Yang Z, Fang H. Research on Green Productivity of Chinese Real Estate Companies—Based on SBM-DEA and TOBIT Models. Sustainability, 2020, 12(8): 3122-3123.

Watermelon Rootstock Seedling Detection Based on Improved YOLOv8 Image Segmentation

Qingcang Yu*, Zihao Xu, Yi Zhu
School of Computer Science and Technology, Zhejiang Sci-tech University, Hangzhou 310018, China

***Abstract*—Automated grafting is an important means for modern agriculture to improve production efficiency and graft seedling quality, among which the use of visual systems to quickly segment target rootstock seedlings is the key technology to achieve automated grafting. This study aims to solve the problems of inaccurate image segmentation and slow detection speed in traditional rootstock seedling segmentation algorithms. To address these challenges, this study proposes a lightweight segmentation method based on an improved version of YOLOv8s-seg. The improved YOLOv8-seg introduces FasterNet as the backbone network and designs an RCAAM module to enhance feature extraction ability and lightweight model. The D-C2f module is improved to enhance feature fusion ability, achieving efficient and accurate segmentation of watermelon rootstock seedlings and improving grafting efficiency. This article designs a series of comparative experiments, comparing the improved version of YOLOv8-seg with classic models such as Unet, SOLO v2, Mask R-CNN, Deeplabv3+ on a test set containing watermelon rootstock seedlings, and evaluating the recognition performance and detection effect of the model. The experimental results show that the improved version of YOLOv8-seg outperforms other models in mAP coefficient index and can segment seedlings more accurately. This study provides reliable deep learning-based solution for the development of automatic grafting robots, which can effectively reduce labor costs and improve grafting efficiency, meeting the requirements of automated equipment for inference efficiency and hardware resources.**

Keywords*—Image segmentation; YOLOv8s-seg; lightweight; deep learning

I. Introduction

Grafting of watermelon rootstock seedlings is a key technology widely used in melon cultivation, which can improve the disease resistance, adaptability, and yield of watermelons through grafting, and is highly valued by agricultural producers [1]. In recent years, with the continuous growth of demand in the high-quality watermelon market, watermelon grafting technology has gradually become standardized and scaled up [2]. At present, most grafting operations use traditional manual methods, but manual grafting has problems such as low efficiency and high cost, which restrict the further promotion of grafting technology and the efficient development of the watermelon industry. In addition, due to the precise cutting and combination of the delicate characteristics of the seedlings during the grafting process, the skill requirements for operators are high, and even a slight carelessness may affect the survival rate of grafting. In order to achieve the large-scale and intelligent development of watermelon rootstock seedling grafting, the research and development of efficient and intelligent grafting equipment has become an important direction for promoting the modernization of the watermelon industry [3]. Automated grafting is an important means for modern agriculture to improve production efficiency and graft seedling quality, among which the use of visual systems to quickly segment target rootstock seedlings is one of the key technologies for achieving automated grafting. If the segmentation is not accurate, it will not only reduce the efficiency of grafting, but may also lead to failure, affecting the subsequent growth and survival rate of seedlings [4]. Therefore, utilizing advanced visual systems and image processing techniques for real-time segmentation and localization of target areas can effectively improve the accuracy and consistency of grafting operations, and reduce the cost and risk of manual intervention [5].

However, traditional methods for identifying rootstock leaves have significant limitations in practical applications, making it difficult to meet the precision and reliability requirements of automated grafting. Rootstock leaves usually have miniaturization characteristics. In addition, rootstock leaves are often mixed with surrounding branches, soil, or other plant leaves, and the background texture is complex and varied. In the process of leaf edge segmentation, traditional methods often have jagged and uneven segmentation boundaries, especially when facing diverse shapes or overlapping leaves, their performance is particularly inadequate. When the lighting conditions are uneven, the background is complex, or the leaf targets are small, traditional visual detection models are more prone to false positives and false negatives. These issues not only reduce the accuracy of identifying rootstock leaves, but may also lead to grafting failure or mechanical equipment misoperation, increasing the risks and costs of agricultural production.

Existing segmentation networks such as Unet, SOLO v2, Mask R-CNN, and Deeplab v3+ are simple and highly accurate in seedling segmentation tasks, they have several limitations that make them less suitable for the problem at hand. Specifically, these networks are difficult to perform shallow feature aggregation, resulting in poor segmentation performance on small and low contrast seedling leaves. In addition, their spatial perception ability is relatively weak, making it difficult to accurately depict the edges of overlapping or obscured seedlings. In addition, many of these models have high computational costs, which makes them perform poorly in real-time applications in actual agricultural production

environments. On the other hand, the YOLOv8 algorithm chosen in this article surpasses its predecessors in the YOLO series in terms of recognition accuracy, speed, and real-time performance. Its efficient feature extraction and multi-scale processing capabilities make it particularly suitable for seedling segmentation tasks. However, despite YOLOv8's strong performance on public datasets, there are still certain limitations in handling small object segmentation and complex background interference. To address these challenges, this paper proposes an improved YOLOv8 neural network architecture that enhances feature extraction, fusion mechanisms, and segmentation accuracy, making it more effective in rootstock seedling segmentation. By addressing the limitations of existing methods and leveraging the advantages of YOLOv8, the proposed method ensures high segmentation accuracy and real-time applicability, providing a powerful solution for automatic grafting. In the feature extraction stage, YOLOv8 utilizes its lightweight and powerful object detection capabilities as the backbone network to quickly segment the overall contour of the rootstock. In the feature extraction stage, FasterNet lightweight network is used to maintain high computational efficiency while achieving excellent performance in feature extraction. And integrate the RCAAM module into the backbone to alleviate the problem of high-frequency information loss in deep feature images. The D-C2f module was introduced in the feature fusion stage to enhance the learning ability of morphological features of different watermelon rootstock seedlings. By integrating these improvements, YOLOv8 demonstrates outstanding performance in tasks involving small and complex targets. It achieves higher segmentation accuracy, reduces false positives, and has stronger adaptability to different conditions, consolidating its position as the most advanced model in precision agriculture applications. The experimental results show that the neural network significantly improves the accuracy and real-time performance of rootstock seedling recognition tasks under complex backgrounds and varying lighting conditions. Compared with traditional models, this method exhibits stronger robustness and adaptability in detecting key parts of rootstock seedlings, providing reliable technical support for automated grafting equipment.

At the end of the introduction, the structure of this article is summarized as follows: Section III provides a detailed explanation of the improved YOLOv8 architecture and the modifications made to improve the accuracy of seedling recognition. The Section IV describes the experimental setup, including the dataset used and the evaluation metrics used to assess model performance. The Section V introduces the results and highlights the model proposed in this paper in comparison to other sub models. Finally, the Section VI summarizes the potential application exploration and future research directions of this model in real automatic grafting robots. This structure aims to guide readers through research and provide a clear understanding.

II. Related Work

Historically, grafting operations mainly relied on manual labor, which was not only time-consuming and labor-intensive, but also easily limited by workers' experience and technical level. In the process of leaf edge segmentation, traditional methods often have jagged and uneven segmentation boundaries, especially when facing diverse shapes or overlapping leaves, their performance is particularly inadequate [6]. When the lighting conditions are uneven, the background is complex, or the leaf targets are small, traditional visual detection models are more prone to false positives and false negatives [7]. Scholars have conducted in-depth research on the grafting process using advanced visual algorithms. In 2013, He et al. [8] proposed a method based on machine vision using ellipse fitting to restore seedling leaf surfaces and extract parameters for robot automatic grafting in order to improve the automation level of fruit and vegetable grafting robots. In 2015, Zhang et al. [9] proposed a comprehensive image processing algorithm to extract feature information of grafting seedlings for relevant vegetable grafting robots. The rapid target recognition technology achieved through visual systems can not only significantly improve grafting efficiency, but also reduce labor costs and intensity, while improving the quality and consistency of grafted seedlings. It is an important direction for promoting the intelligent and precise development of modern agriculture.

With the development of deep learning and computer vision technology, image-based seedling recognition methods have been widely studied and applied. Zuo et al. [10] proposed a crop seedling plant segmentation network model that integrates semantic and edge information of the target region in order to accurately segment crop seedlings in natural environments and achieve automatic measurement of seedling position and phenotype. The experimental results show that under the same network training parameters, the average cross merge rate and average recall rate obtained by testing the method proposed in this paper are 58.13% and 64.72%, respectively, which are better than the segmentation results corresponding to manually labeled samples; In addition, after adding 10% of outdoor seedling images to the training samples, the average pixel accuracy of this method on the outdoor test set can reach 90.54%, demonstrating good generalization ability. Image processing technology can be used for high-throughput collection and analysis of crop population phenotypes, which is of great significance for crop growth monitoring, seedling condition assessment, and cultivation management. However, existing methods rely on empirical segmentation thresholds, resulting in insufficient accuracy in extracting phenotypes. Li et al. [11] proposed a method for extracting phenotypes from aerial images of maize seedlings, using maize as an example. Explored an end-to-end segmentation network called PlantU-net, which uses a small amount of training data to achieve automatic segmentation of overhead images of maize seedling populations. Automatically extract morphological and color related phenotypes, including maize stem coverage, external radius, aspect ratio, and plant orientation plane angle. Ma et al. [12] proposed a method for processing greenhouse vegetable leaf disease symptom images in order to achieve robust segmentation, as uneven lighting and cluttered backgrounds are the most challenging problems in disease symptom image segmentation. The results show that the overall accuracy of the proposed method is 90.67%, indicating that the method can obtain robust segmentation of disease symptom images.

III. IMPROVED YOLOV8 MODEL

A. YOLO v8

The YOLO network has undergone various improvements to address the challenges brought by early versions, aiming to enhance its adaptability to specific tasks while maintaining a balance between detection speed and accuracy. The improved version YOLOv8-Seg adopts a modular design, including three main components: backbone, neck, and head. The backbone extracts basic features from the input image, the neck processes and integrates these features at multiple scales, and the head generates the final prediction, including object classification, bounding box coordinates, and segmentation masks.

The high accuracy of the YOLOv8 model is attributed to the replacement of the C3 module with the C2f module in the YOLOv5 backbone network and neck network. The C2f module first goes through a Conv, uses the chunk function to evenly split the out into two vectors, and then saves them to a list. The latter half is input into a Bottleneck Block, which contains n Bottlenecks. Each Bottleneck output is appended to the list. In the YOLOv8 model, the head part of the prediction head has undergone significant changes compared to the YOLOv5 model. It has been replaced from an anchor box based object detection algorithm to an anchor box free object detection algorithm, which has the advantages of fast convergence and improved regression performance. The decoupling head structure is adopted to separate the regression branch from the prediction branch, and the integral form representation method proposed in the Distribution Focal Loss strategy is used for the regression branch. The coordinates are transformed from a deterministic single value prediction to a distribution. Compared to using coupling heads, decoupling heads can effectively reduce the computational complexity of model segmentation of rootstock seedlings, which not only accelerates processing speed but also enhances the generalization ability and robustness of rootstock seedling recognition.

B. Improved YOLOv8 Model

In the process of rootstock seedling segmentation, the traditional method often appears jagged and unsmooth segmentation boundary, especially in the face of morphological diversification or leaf overlap, uneven lighting conditions, complex backgrounds, or small leaf targets, this paper proposes an improved YOLOv8 neural network architecture, which can be used to improve the performance of the neural network, for efficient and accurate segmentation of cotyledon parts of rootstock seedlings, the structure is shown in Fig. 1. In the feature extraction stage, YOLOv8 utilizes its lightweight and powerful target detection ability as the backbone network to quickly identify the overall contour of the rootstock. However, YOLOv8 has some limitations in the extraction of detailed features. To make up for this deficiency, the advanced feature extraction ability of FasterNet is integrated to capture finer features through deep convolution, especially in the localization of small target sites. At the same time, the RCAAM module is designed to non-uniformly weight the key features in the spatial and channel dimensions to highlight the useful information and suppress the interference of background noise. Combined with the task-aware classification and positioning module, the recognition accuracy and location accuracy of rootstock seedlings are further improved. In addition, the D-C2F module was improved to achieve effective fusion of multi-scale features through dynamic convolution, highlighting high-resolution features of key parts of rootstocks from coarse to fine, and improving the efficiency of rootstock management, it is used to improve the learning ability of different morphological rootstock seedling characteristics, and the deformable modeling is used to adapt to the target morphological change characteristics.

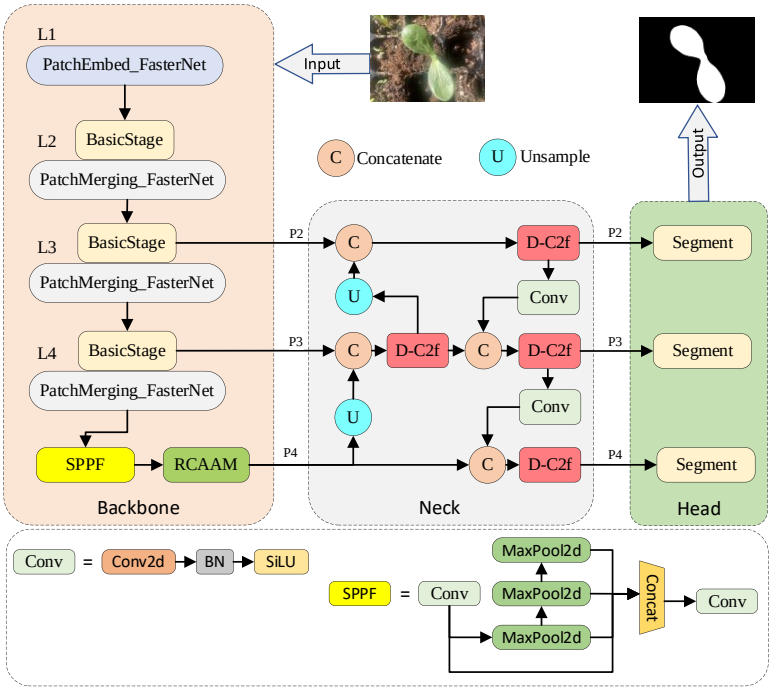


Fig. 1. Architecture of improved YOLOv8.

1) Based on FasterNet feature extraction: YOLOv8 uses DarkNet-53 as the backbone network, and its architecture consists of a series of convolutional layers and residual modules. Although DarkNet-53 performs well in general object detection tasks, it faces certain limitations in rootstock seedling segmentation, such as excessive redundant information, which limits training and inference speed, and insufficient segmentation performance for small target leaves and complex backgrounds. In addition, traditional backbone networks have certain bottlenecks in multi-scale information processing, making it difficult to effectively capture the global and local features of rootstock leaves. To address the aforementioned issues, CHEN et al. [13] proposed an efficient neural network called FasterNet, whose structure is shown in Fig. 2. Partial Convolution (PConv) is introduced in the article, which reduces redundant calculations and memory access by focusing on the features of specific regions, while significantly improving the efficiency of multi-scale feature extraction. This improvement enables FasterNet to achieve efficient model deployment on edge devices, while enhancing its ability to recognize small targets in complex scenes. In the task of rootstock seedling segmentation, combining FasterNet with YOLOv8 can effectively compensate for the shortcomings of DarkNet-53 in detail processing and feature extraction, providing new technical support for precise segmentation of rootstock leaves in complex agricultural scenes.

PConv only applies regular convolution to known regions in the input feature map, keeping other parts unchanged. The floating-point operands (FLOPs) of regular convolution and DWConv can be represented as:

$$F_{SC} = h \times w \times k^2 \times c^2 \tag{1}$$

$$F_{DWC} = h \times w \times k^2 \times c \tag{2}$$

In practical applications, PConv usually selects the first or last consecutive c_p channel, and requires that the input and output feature maps have the same channel. Therefore, the FLOPs of a PConv can be expressed as:

$$F_{PC} = h \times w \times c_p^2 \times k^2 \tag{3}$$

In the formula, cp=c/4, then the FLOPs of PConv are only 1/16 of those of regular convolution.

Its memory efficiency is significantly improved because PConv reduces memory access and data transmission by treating specific channels as representatives of the entire feature map, thereby accelerating computation speed without sacrificing accuracy. This paper replaces the backbone network of YOLOv8 with FasterNet to reduce redundancy and improve the overall computing speed, thus promoting efficient edge computing applications.

2) Residual channel adaptive attention module (RCAAM): To alleviate the problem of high-frequency information loss caused by the decrease in the number of convolutional channels in each layer, Wang et al. [14] proposed an adaptive attention module (AAM) to improve object detection in multi-scale scenes. Although this module improves the ability of feature extraction by adjusting the weight allocation of multi-scale features through adaptive average pooling, its operations mainly focus on low dimensional operations, resulting in an increase in shallow information redundancy and interference phenomena in feature images.

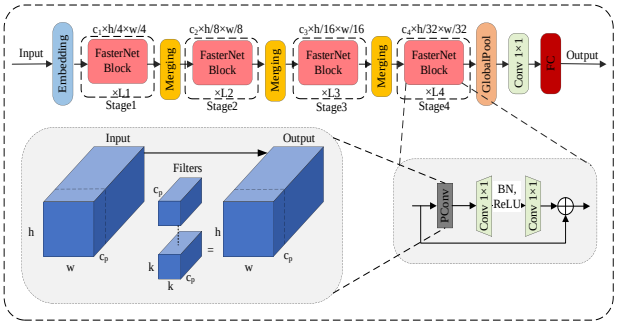


Fig. 2. Architecture of FasterNet.

To this end, this article has redesigned the AAM module and combined it with the residual channel attention network (RCAN) proposed by Zhang et al. [15], introducing super-resolution technology to enhance the detail representation and clarity of feature images. The core module in RCAN, residual group (RG), enriches feature encoding information through pixel addition strategy and enhances channel perception weights to recover more high-frequency information from the bottom layer, as shown in Fig. 3.

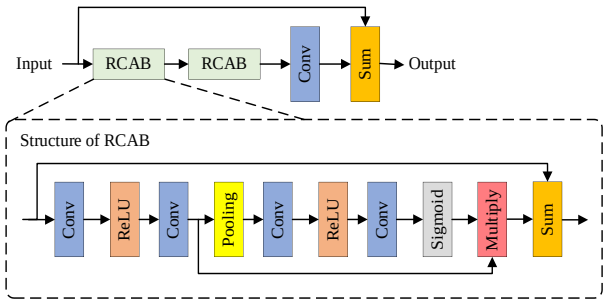


Fig. 3. Architecture of residual group.

This article proposes a new residual channel adaptive attention module (RCAAM) based on residual groups. RCAAM generates multiple sets of high-resolution feature maps in low resolution feature map reconstruction, which not only reduces redundant computation but also significantly restores shallow feature information of occluded targets.

In addition, this article introduces ECA module in RCAAM, which enhances contextual information correlation by weighted fusion of feature maps, thus more accurately segmenting rootstock seedlings in complex backgrounds. This article also compared multiple attention mechanisms, and the Seg loss function of the YOLOv8 model is shown in Fig. 4, with the model incorporating the ECA module performing the best in terms of loss function. The structure of the RCAAM module is shown in Fig. 5.

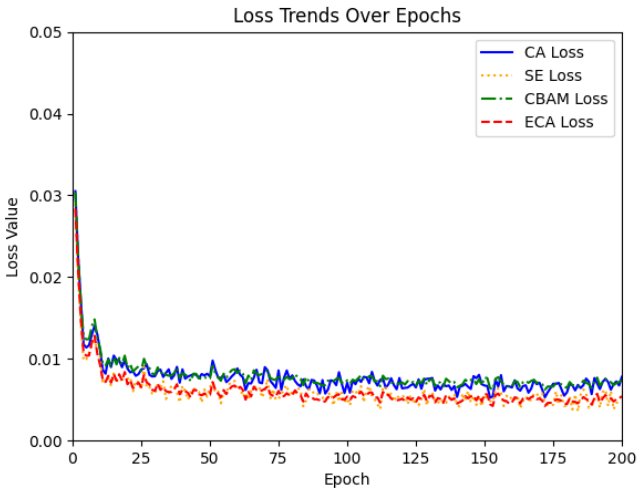


Fig. 4. Seg loss function curves for different attention mechanisms.

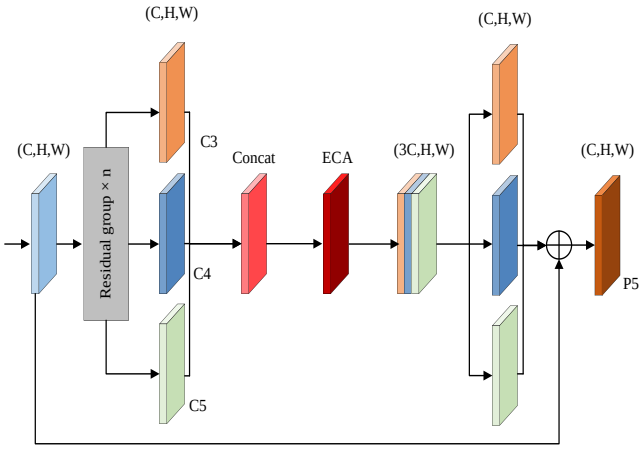


Fig. 5. The structure of the RCAAM module.

3) D-C2f Module: During the growth process of rootstock seedlings, their leaf shapes exhibit a high degree of diversity and are often accompanied by overlapping phenomena. In addition, due to the fact that rootstock seedlings usually grow in a plug environment, there are a large number of ridges and soil in the background, which further increases the difficulty of segmentation and recognition. Traditional standard convolutional neural networks (CNNs) use fixed convolution kernels for feature learning, which makes it difficult to fully capture the diverse morphological features of rootstock leaves, especially in cases of leaf overlap and blurred boundaries, leading to inaccurate feature extraction, false positives, and missed detections.

To address the above issues, this paper introduces deformable convolution (DConv) to enhance the model's ability to learn diverse rootstock leaf features. DConv adaptively adjusts the size and sampling position of the convolution kernel through deformable modeling, in order to better adapt to the changing characteristics of the target shape. Its structure is shown in Fig. 6.

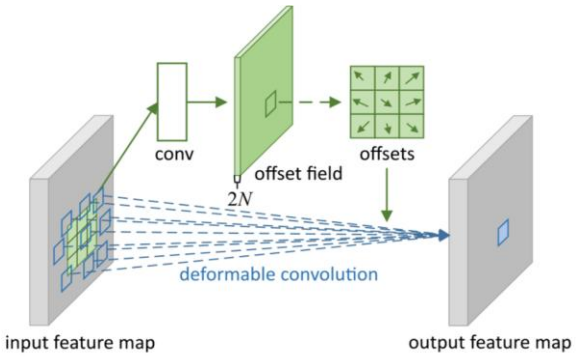


Fig. 6. The structure of the DConv module.

Specifically, DConv introduces offset in the convolutional receptive field and dynamically adjusts the position of each convolution kernel. For example, for a 3x3 convolution kernel, the regular sampling grid can be expanded by the offset, as shown:

$$R = \{\Delta p_n \mid n = 1,2,\cdots,n\} \tag{4}$$

By combining modulation variables, the model can automatically learn the offset and weight of each sampling point, enabling the sampling points to fit the target shape more accurately. For any position p, the output mapping Y is expressed as:

$$Y_{(p)} = \sum_{k=1}^{K} \omega_k \cdot x(p_k + \Delta p_k) \cdot \Delta m_k \tag{5}$$

Among them, ω_k and Δm_k respectively represent the weight and modulation variable of the kth position. Due to the fact that the offset Δp_k is usually in fractional form, bilinear interpolation is used for calculation, as shown in Eq. (6) to Eq. (8).

$$g(q_y, p_y) = \max(0, 1 - |a - b|) \tag{6}$$

$$G(q, p) = g(q_x, p_x) \cdot g(q_y, p_y) \tag{7}$$

$$X(p) = \sum_{q} G(q, p) \cdot g(q_y, p_y) \tag{8}$$

The application of DConv can dynamically adjust the sampling position and convolution kernel size according to the specific morphology of rootstock leaves. Firstly, by preprocessing the input feature map, offset and modulation variables are generated to calculate the offset direction of pixel points and obtain irregularly distributed sampling points. Subsequently, these sampling points are used to resample the feature map and combined with the convolution kernel to calculate the final convolution result.

To further improve the performance of the model, this paper redesigns the D-C2f module based on DConv convolution, as shown in Fig. 7. The D-C2f module can adaptively adjust the size and sampling position of the

convolution kernel based on the local features of the rootstock leaves by introducing learnable deformation parameters and offset weights. Compared with traditional methods, this module provides a larger receptive field range in the output features, greatly improving the model's ability to extract rootstock leaf features and model overlapping leaf spatial distribution in plug scenes, thereby effectively improving the segmentation and recognition performance of rootstock seedlings.

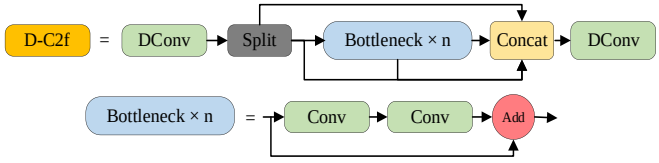


Fig. 7. The structure of the D-C2f module.

IV. Experiments And Analysis

A. Model Training and Testing Trials

1) Experimental data: The rootstock seedling task dataset of this study consists of images of watermelon seedlings in the cotyledon stage, constructed in two stages to ensure data diversity and representativeness. The first stage data was obtained from actual shooting at the seedling center, which captured 400 high-resolution images of watermelon seedlings. In the second stage, in order to unify and adapt the model training requirements, all images were preprocessed, cropped, and adjusted to a resolution of 640 × 640, while maintaining a 24 bit RGB format. After data augmentation and filtering, a total of 1600 images were generated, of which 1350 were divided into training and validation sets and randomly segmented in a 9:1 ratio. The remaining 250 images were used as the test set for independent evaluation of model performance, and the test set did not participate in model training. The dataset was annotated using the LabelMe tool and further converted to VOC dataset format. This dataset not only reflects the diversity of rootstock seedling characteristics, but also takes into account the plug scenes in real grafting environments, providing a solid foundation for training and evaluating segmentation and recognition models.

2) Comparison of segmentation network models: This study compared the performance of YOLOv8, Unet, SOLO v2, Mask R-CNN, and Deeplab V3+. Among them, SOLO v2 and Mask R-CNN belong to instance segmentation algorithms, while improved YOLOv8, Unet, and Deeplab v3+ belong to semantic segmentation models.

a) Unet: The Unet network structure is characterized by its clear U-shaped architecture, with symmetric encoders on the left and decoders on the right, which can effectively enhance the ability to extract feature map information. Unet has a low dependence on the number of images and only requires a small number of images to complete end-to-end training, making it very suitable for medical image segmentation. However, due to its relatively simple design, it is prone to inaccurate segmentation when dealing with complex backgrounds and small target tasks.

b) Mask R-CNN: Mask R-CNN is based on Faster R-CNN and is used to predict instance segmentation masks by adding a mask branch that runs in parallel with the classification and bounding box regression branches [16]. This method adopts a top-down detection approach, first detecting the regions of each instance, and then segmenting the instance masks within these regions. Detection based methods typically have high accuracy, but rely on precise bounding box detection, which places high demands on computational resources.

c) SOLO v2: Unlike Mask R-CNN, SOLO v2 transforms segmentation tasks into pixel classification problems, thereby eliminating the step of proposal generation [17]. The network consists of two branches: a category prediction branch for predicting the semantic category of the target, and a masking branch for predicting the instance mask of the target. This method reduces computational complexity and can improve the efficiency of instance segmentation to some extent, but its performance may be affected for scenes with complex backgrounds or overlapping targets.

d) Deeplab v3+: As the latest generation model in the Deeplab series, Deeplab v3+ adopts Deeplab v3 in its encoding structure and introduces a decoder to solve the problem of losing detailed information caused by directly upsampling feature maps in Deeplab v3, thus achieving higher performance in semantic segmentation tasks [18]. Deeplab v3+ has strong ability to recover detailed information, but it may still face certain challenges when dealing with small target tasks with complex backgrounds.

3) Testing trial setup: The hardware configuration is Intel i5-12490K CPU and Nvidia GeForce RTX 4060ti GPU. The experimental method of this paper is developed and implemented in Python 3.8 environment based on the deep learning framework PyTorch 2.2.2, using CUDA 12.4 for GPU acceleration. In the model training phase, the SGD optimizer is used to optimize the performance of the model, and the transfer learning strategy is used to initialize the model by loading the pre-trained model weight"yolov8s.pt" to accelerate the convergence speed of the model. The model training parameters in this paper are shown in Table I.

TABLE I. Experimental Model Parameters

Parameter Name	Value
Batch size	16
Epoch	200
Learning rate	0.01
Momentum factor	0.937
Image size	640 dpi×640 dpi

4) Evaluation metrics: In order to evaluate the segmentation performance of the improved YOLOv8 and the contrast model, this study uses the mean average Precision (mAP) as the main evaluation index, and combines Precision

and Recall to comprehensively analyze the performance of the model. Map is a commonly used metric in detection and segmentation tasks. Its calculation is based on the matching of the predicted results and the real labels, which can comprehensively measure the performance of the model in detection accuracy and coverage. See Eq. (9) and Eq. (10) for calculations of Precision and Recall.

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

In a single image, the calculation of mAP involves multiple target categories or instances. Firstly, calculate the accuracy and recall curves for each category: for a certain category, sort the model's predictions for that category in descending order of confidence, and gradually calculate the accuracy and recall at different thresholds. Secondly, calculate the average precision (AP): By integrating the precision and recall curves, obtain the AP value for that category. The higher the AP value, the stronger the detection ability of the model in that category. The AP formula is as follows:

$$AP = \int_0^1 P(R)dR \tag{11}$$

Finally, calculate the average precision mean (mAP) for all categories: take the average of the AP values for all categories, which is mAP, using the following formula:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{12}$$

where, N is the number of categories, and AP_i is the average accuracy of the i-th category.

In the segmentation task, the mAP calculation is usually based on the Intersection over Union (IoU) threshold setting. For example, when the threshold of IoU is set to 0.5, the ratio of the overlapping area representing the predicted result and the real label region to the joint area needs to be greater than 50% to be considered a correct detection. mAP@0.5 is a commonly used evaluation metric, in addition, the mAP can also be calculated at a higher IoU threshold to examine the sensitivity of the model to the target boundary.

5) Analysis of model training: In the training process of rootstock seedling segmentation algorithm, loss function plays an important role in evaluating the performance of the model. The smaller the loss value, the better the performance of the model and the more significant the optimization effect. As shown in Fig. 8, the training process of the improved algorithm generates a loss function curve, which intuitively reflects the obvious downward trend of the loss value with the number of iterations, indicating that the model performance is continuously optimized.

Among them, the decrease of box loss is the most significant, which reflects the improvement of target positioning accuracy. After approximately 75 training rounds, the Obj loss and seg loss tended to stabilize.

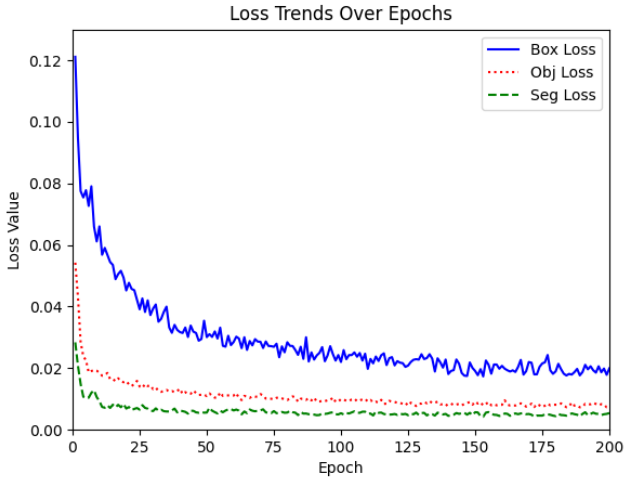


Fig. 8. Curve of model loss function.

The training process of YOLOv8 network before and after improvement is shown in Fig. 9. During the training process, the mAP of the model went through three stages: within the initial 25 training rounds, the mAP value rose rapidly, indicating that the model was in a rapid fitting stage; subsequently, between the 25th and 75th training rounds, the mAP value increased rapidly, indicating that the model was in a rapid fitting stage, the mAP value fluctuated greatly; from the 75 th to 200 th training round, the mAP curve tended to be stable and convergent with the change of learning rate gradually decreasing. With the increase of training times, the mAP value gradually converges, indicating that the model tends to be stable after further training.

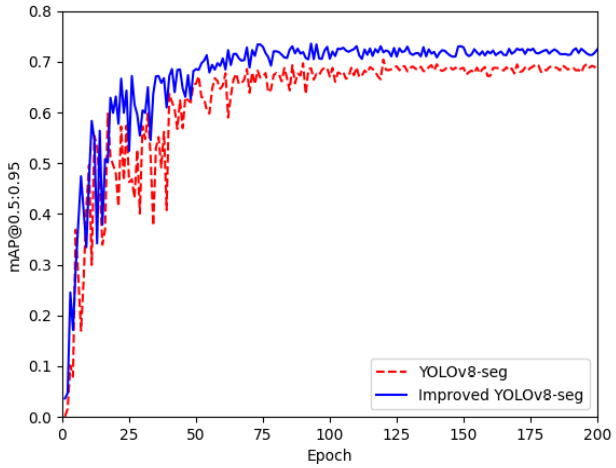


Fig. 9. Comparison of mAP curve changes with epoch variation.

B. Experimental Results

1) Ablation experiment: The purpose of ablation experiments is to explore the specific impact on model performance when certain parts of the network are removed or

modified. This article takes YOLOv8 as the baseline model, improves it in three aspects, designs an improved YOLOv8 model, and conducts five ablation experiments on the improved model. Among them, Improved Model 1 represents replacing the original backbone network with a FasterNet network, Improved Model 2 represents replacing the C2f module in the neck with a D-C2f module, and Improved Model 3 represents introducing the RCAAM module into the backbone network. The detailed experimental results are shown in TABLE II. I.

TABLE II. PERFORMANCE COMPARISON OF EACH IMPROVED MODEL

Model	FasterNet	D-C2f	RCAAM	P/%	R/%	mAP@0.5/%	Parameters/10^6	FLOPs/10^9
YOLOv8s	-	-	-	91.2	90.3	92.7	11.1	28.8
Improved model 1	√	-	-	91.4	93.6	95.3	6.0	11.8
Improved model 2	-	√	-	86.4	92.0	95.7	10.3	25.6
Improved model 3	-	-	√	95.6	91.2	94.4	7.8	21.8
Improved YOLOv8	√	√	√	95.4	95.0	96.6	3.2	10.6

According to the results in Table II, Improved Model 1 replaces the backbone network with a lightweight Faster Net network, mAP@0.5 Value increased by 2.6%, parameter and computation reduced by 5.1M and 17G. Indicating that lightweight FasterNet networks can reduce redundant information, optimize model parameter count and computational complexity, and improve detection speed. The FasterNet network adopts PConv to better extract multi-scale information of rootstock seedlings and cotyledons, improve detection ability and compress model volume, enhance the network's feature extraction ability, and reduce parameter counting.

Improved model 2 replaces the C2f module on the neck with a lighter D-C2f module, reducing the parameter and computational complexity by 0.8M and 3.2G compared to the baseline model, mAP@0.5. The value has increased by 3%. The D-C2f module can adaptively adjust the size and sampling position of the convolution kernel based on the local features of the rootstock leaves by introducing learnable deformation parameters and offset weights. This provides a larger receptive field range in the output features, effectively improving the segmentation and recognition performance of rootstock seedlings. Better save computational costs and improve training speed.

Improved model 3 then introduced the RCAAM module into the original backbone network, mAP@0.5 Compared to the baseline model, the value increased by 1.7%, while the parameter and computational complexity decreased by 3.3M and 7G, respectively. Compared with traditional pooling methods, RCAAM not only improves feature extraction efficiency, enhances contextual information correlation, but also suppresses useless information interference, thus more accurately segmenting rootstock leaves.

Finally, three modules were added simultaneously, namely the improved YOLOv8 algorithm proposed in this article, which increased mAP values by 3.9%, reduced parameter and computational complexity to only 28.8% and 36.8% of the baseline model, and achieved a recall rate of 95.0%. The ablation experiment results have verified the rationality and superiority of the algorithm proposed in this paper in terms of detection accuracy, speed, and lightweight.

2) Comparison of evaluation metrics between improved YOLOv8 and other models: In order to evaluate the performance of the improved YOLOv8 model in rootstock seedling segmentation, this study compared and analyzed the segmentation abilities of different models in the test set, focusing on their performance in handling significantly different backgrounds and complex low contrast scenes. To evaluate in detail the segmentation performance of the improved rootstock seedling recognition model, this study used mAP@0.5 Compare its performance with Unet, SOLO v2, Mask R-CNN, and Deeplab v3+ on the test set. The test set consists of 250 images, including two distinct types of seedlings, providing a diverse basis for performance evaluation. The results of evaluation metrics comparison are shown in TABLE III. .

TABLE III. COMPARISON BETWEEN MAINSTREAM SEGMENT ALGORITHMS AND THE PROPOSED METHOD

Model	P/%	R/%	mAP@0.5/%	Parameters/10^6	FLOPs/10^9
Mask-RCNN	87.3	88.4	92.3	44	44.7
SOLO v2	89.5	92.7	93.9	37.2	41
Deeplab v3+	84.8	85.0	88.9	41.2	12.6
Unet	91.3	93.3	94.7	31	8.3
Improved YOLOv8	95.4	95.0	96.6	3.2	10.6

The research results indicate that Mask R-CNN and Deeplab v3+ mAP@0.5 The indicators are significantly lower than the improved YOLOv8 model and Unet. This low performance reflects that these two models have difficulty accurately segmenting small seedlings under testing conditions. In contrast, improving the recognition model of rootstock seedlings mAP@0.5 The value is slightly higher than Unet, indicating its optimal performance among the four models. Specifically, the improved YOLOv8 model mAP@0.5 It reached 96.6%, indicating its significant advantage in the accuracy of rootstock seedling image segmentation and detection.

These results highlight the effectiveness of the improved model in segmenting rootstock seedlings and cotyledons under complex environmental conditions, and its stability and

robustness make it have important application potential in grafting management and production processes.

When segmenting rootstock seedling images with significant contrast between background and rootstock seedling cotyledons, the improved YOLOv8 model was compared with four other models (Unet, SOLO v2, Mask R-CNN, Deeplab v3+), and the results are shown in Fig. 10. The results indicate that Deeplab v3+ is relatively less effective than other models in segmenting small or edge blurred leaves. The segmentation performance of SOLO v2 and Mask R-CNN is superior to Deeplab v3+, but the computational cost is high and the performance is still insufficient when detecting small leaves or low contrast targets, which can easily miss some key leaf regions. The Unet model may encounter problems of over segmentation or under segmentation when segmenting overlapping or blurred boundary rootstock leaves, especially for leaves with complex shapes. The improved YOLOv8 model performs well in rootstock seedling segmentation tasks. Compared with other models, the improved YOLOv8 effectively enhances segmentation performance through stronger feature extraction ability and optimized feature fusion mechanism, and has excellent robustness and adaptability. The comparative experimental results show that the improved YOLOv8 model is suitable for precise segmentation of rootstock seedlings in complex environments, providing efficient and reliable technical support for automatic grafting in smart agriculture.

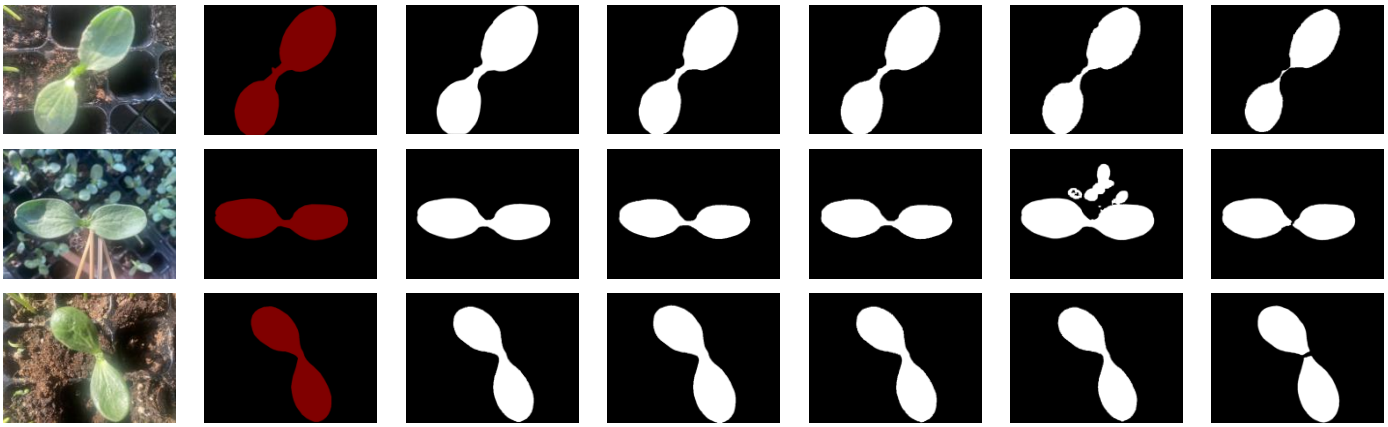


Fig. 10. Comparison of mAP curve changes with epoch variation.

V. Discussion

The improved YOLOv8 rootstock seedling recognition model outperforms the original YOLOv8 model in terms of recognition performance. In practical complex environments, the original model performs poorly in situations where lighting is uneven or the background is similar to the characteristics of rootstock seedlings, leading to inaccurate segmentation. The improved model can accurately segment rootstock seedlings with higher recognition rate, especially in complex environments, demonstrating stronger adaptability. At the same time, it effectively solves the problem of low accuracy in detecting small or distant rootstock seedlings in the original model.

This performance improvement does not come from the improvement of a single method, but from the comprehensive enhancement of feature extraction and feature fusion capabilities. By improving the network structure, optimizing the feature processing flow, and introducing more advanced feature fusion strategies, the robustness and segmentation accuracy of the model in complex environments have been significantly improved, better adapting to diverse hardware deployment requirements, providing more efficient and intelligent support for modern agricultural production, and promoting the implementation of precision agriculture. In order to ensure the efficient application of the rootstock seedling recognition system in practical agricultural production, it is possible to consider removing unimportant connections or neurons from the model in the future. Pruning technology can significantly reduce the size and inference time of the model while maintaining its recognition accuracy. This method is particularly suitable for devices with limited computing resources, making the rootstock seedling recognition system more efficient and meeting the real-time detection needs of resource constrained platforms. In large-scale plug seedling grafting, the segmentation model of the pruned rootstock seedlings can achieve real-time detection on unmanned aerial vehicles or vehicle platforms, improving the efficiency and accuracy of grafting.

VI. Conclusion

In order to solve the problems of inaccurate edge segmentation and low detection efficiency of traditional detection algorithms, this paper proposes a rootstock seedling detection method based on improved YOLOv8. By suppressing invalid features in high-order and low-order feature fusion and enhancing the ability of the model to extract rootstock seedling features, this method can effectively realize the recognition of narrow and small seedlings and large seedlings.

In this paper, mAP@0.5 is used as the evaluation index to compare the performance of the improved YOLOv8 and the basic UNET model in the segmentation of watermelon rootstock seedlings. The results show that the improved method is superior to the basic UNET algorithm in both quantitative and qualitative evaluation. In the aspect of model performance, the model parameters and FLOPs were used as evaluation criteria. Although the improved YOLOv8 outperforms Unet in recognition performance, its FLOPs are

2.3 g higher than Unet, showing a good balance between performance and recognition accuracy. Compared with the other three classical segmentation networks, the results show that the MAP@0.5 score of the improved YOLOv8 is 0.8403, which is better than the classical models such as SOLO v2, Mask R-CNN and Deeplab v3+. Compared with other models, the improved YOLOv8 model has the highest performance in identifying rootstock seedlings, and can accurately extract the characteristic information of seedlings.

The improved model has important application potential in watermelon rootstock grafting. Its ability to accurately segment the characteristics of small and irregular watermelon rootstock seedlings provides an important guarantee for improving the efficiency and accuracy of agricultural production. By integrating the improved model into the automatic grafting robot, the time and labor cost required for traditional manual grafting can be significantly reduced.

Although the improved YOLOv8 model performs well, there are still areas that need improvement in the future. Firstly, reducing computational complexity while maintaining high detection accuracy remains the main research direction. In future research, optimizing the model structure or using a more lightweight network architecture can be considered. Secondly, the actual agricultural production environment is more complex, with different lighting conditions, different seedling growth conditions, and different seedling varieties. Future work should improve the robustness of models in different environments. In addition, combining this model with edge computing equipment can enhance its practical applicability in precision agriculture.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (No.51375460).

REFERENCES

[1] Kyriacou M C, Rouphael Y, Colla G, et al. Vegetable grafting: The implications of a growing agronomic imperative for vegetable fruit quality and nutritive value[J]. Frontiers in plant science, 2017, 8: 741.

[2] Kumar P, Rouphael Y, Cardarelli M, et al. Vegetable grafting as a tool to improve drought resistance and water use efficiency[J]. Frontiers in plant science, 2017, 8: 1130.

[3] Lee J M, Kubota C, Tsao S J, et al. Current status of vegetable grafting: Diffusion, grafting techniques, automation[J]. Scientia Horticulturae, 2010, 127(2): 93-105.

[4] Maurya D, Pandey A K, Kumar V, et al. Grafting techniques in vegetable crops: A review[J]. International Journal of Chemical Studies, 2019, 7(2): 1664-1672.

[5] Gaion L A, Braz L T, Carvalho R F. Grafting in vegetable crops: A great technique for agriculture[J]. International Journal of Vegetable Science, 2018, 24(1): 85-102.

[6] Hétroy-Wheeler F, Casella E, Boltcheva D. Segmentation of tree seedling point clouds into elementary units[J]. International Journal of Remote Sensing, 2016, 37(13): 2881-2907.

[7] Scharr H, Minervini M, French A P, et al. Leaf segmentation in plant phenotyping: a collation study[J]. Machine vision and applications, 2016, 27: 585-606.

[8] He L, Cai L, Wu C. Vision-based parameters extraction of seedlings for grafting robot[J]. Transactions of the Chinese Society of Agricultural Engineering, 2013, 29(24): 190-195.

[9] Zhang L, He H, Wu C. Vision method for measuring grafted seedling properties of vegetable grafted robot[J]. Transactions of the Chinese Society of Agricultural Engineering, 2015, 31(9): 32-38.

[10] Zuo X, Lin H, Wang D, et al. A method of crop seedling plant segmentation on edge information fusion model[J]. IEEE Access, 2022, 10: 95281-95293.

[11] Li Y, Wen W, Guo X, et al. High-throughput phenotyping analysis of maize at the seedling stage using end-to-end segmentation network[J]. PLoS One, 2021, 16(1): e0241528.

[12] Ma J, Du K, Zhang L, et al. A segmentation method for greenhouse vegetable foliar disease spots images using color information and region growing[J]. Computers and Electronics in Agriculture, 2017, 142: 110-117.

[13] Chen J, Kao S, He H, et al. Run, don't walk: chasing higher FLOPS for faster neural networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 12021-12031.

[14] Wang J, Chen Y, Dong Z, et al. Improved YOLOv5 network for real-time multi-scale traffic sign detection[J]. Neural Computing and Applications, 2023, 35(10): 7853-7865.

[15] Zhang Y, Li K, Li K, et al. Image super-resolution using very deep residual channel attention networks[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 286-301.

[16] Shen L, Su J, Huang R, et al. Fusing attention mechanism with Mask R-CNN for instance segmentation of grape cluster in the field[J]. Frontiers in plant science, 2022, 13: 934450.

[17] Zhou R J, Zheng L M, Ren C L, et al. Image Segmentation Algorithm in Complex Environment Based on Improved SOLOV2[C]//2023 12th International Conference of Information and Communication Technology (ICTech). IEEE, 2023: 581-585.

[18] Yang T, Zhou S, Xu A, et al. An approach for plant leaf image segmentation based on YOLOV8 and the improved DEEPLABV3+[J]. Plants, 2023, 12(19): 3438.

Object Recognition IoT-Based for People with Disabilities: A Review

Andriana[1], Elli Ruslina[2], Zulkarnain[3], Fajar Arrazaq[4], Sutisna Abdul Rahman[5],
Tjahjo Adiprabowo[6], Puput Dani Prasetyo Adi[7]*, Yudi Yuliyus Maulana[8]
Department.of Electrical Engineering, Universitas Langlangbuana, Bandung, Indonesia[1, 3, 4, 5, 6]
Faculty of Law, Universitas Pasundan, Bandung, Indonesia[2]
Research Center for Telecommunication, BRIN, Bandung, Indonesia[7, 8]

***Abstract*—This research focuses on a literature study on developing a Mini Smart Camera (MSC) system that utilizes Internet of Things (IoT) technology to help people with disabilities interact with their environment. The MSC serves as an assistive device, which integrates object recognition and speech recognition technologies along with an internet-based two-way communication system. Utilizing state-of-the-art hardware and software, the system captures images, processes audio, and transmits data via Real Time Streaming Protocol (RTSP) and Message Queuing Telemetry Transport (MQTT). These protocols serve different purposes: managing data transmission and enabling communication between machines. The MSC is equipped with a 5 MP camera, 2.5 GHz Quad-Core processor, and 4G connectivity, and is connected to a high-performance Ubuntu 22.04 Linux cloud server. The use of OpenCV libraries and machine learning algorithms ensures fast and precise image analysis. By integrating machine learning and natural language processing (NLP), MSC efficiently handles both visual and audio inputs. Key features, including text-to-speech (TTS) and speech-to-text (STT), provide an interactive and adaptive communication interface. The system is designed to improve accessibility and encourage greater independence for people with disabilities in daily activities. The development of multispectral cameras for disabilities will provide a more detailed analysis for the detection of surrounding objects.**

***Keywords*—Internet of Things; mini smart camera; object recognition; speech recognition; assistive technology**

I. Introduction

The global population is witnessing an increasing focus on improving accessibility for individuals with disabilities, driven by both technological advancements and a growing awareness of inclusivity. According to the World Health Organization (WHO), over one billion people globally live with some form of disability, many of whom face challenges in accessing information and engaging with their surroundings. These challenges are particularly significant for individuals with sensory impairments, such as visual, auditory, or speech disabilities [1]. Mini Smart Camera (MSC) system addresses these challenges by integrating Internet of Things (IoT) technologies with machine learning algorithms to provide real-time information and communication tools. The proposed system uses object recognition, speech recognition, and conversion technologies like text-to-speech (TTS) and speech-to-text (STT), which enable individuals to interact with their environment more efficiently. Prior studies have shown that IoT devices can significantly enhance the quality of life for individuals with disabilities by improving environmental interaction [2], [3]. IoT technology allows the MSC system to perform complex tasks such as real-time object recognition and audio processing, which are essential for individuals with disabilities to navigate their environments independently. This paper aims to present the design, implementation, and evaluation of the MSC system, focusing on its potential to empower individuals with disabilities by improving accessibility and independence. In the system that has been made to help people with disabilities, there are still many shortcomings, from models, prototypes, and improper placement of prototypes. In this research, it is discussed comprehensively starting from the placement of the prototype, the size of the prototype, namely the Mini Smart Camera (MSC) used for disabilities so that there is comfort in using it, supported by IoT and Artificial Intelligence technology that can perfect this system and provide novelty.

Moreover, disability monitoring needs to be improved with a complete range of components, including a Global Positioning System (GPS) to monitor the patient's position in addition to detecting other vital parts such as heart rate. This research provides a detailed overview of the IoT system to be built. IoT and AI are key components in building this system, but not all AI components are discussed in this research, but the main contribution lies in IoT design and this design can provide specific results for people with disabilities.

For this reason, this research is presented starting from how this system should be built for people with disabilities, namely building an IoT architecture starting from a literature study on IoT and IoT architecture specifically for disabilities and improvements, for example on Hyperspectral and Multispectral Cameras in the future, in this research seen from the review mode how the use of thermal cameras to help people with disabilities which can then be improved in more detail and comprehensively.

II. Literature Review

A. Internet of Things in Assistive Technology

The integration of IoT in assistive technology has seen significant growth in recent years, supported by numerous studies that highlight its potential to assist individuals with disabilities. IoT-enabled devices offer continuous monitoring, real-time data exchange [21, 22, 23], and remote management,

making them highly adaptable to various user needs. Andriana, et al. [4] emphasized that IoT can transform assistive technologies by enabling real-time data processing, allowing users to receive context-specific support when needed.

The ability of IoT to connect multiple devices and sensors into a cohesive network has positioned it as an ideal choice for assistive systems [5]. Wearable IoT devices, for instance, can monitor physiological data for individuals with mobility limitations or provide real-time navigation assistance for those with visual impairments [25]. Research by Semmary et al. [6] underlined the value of real-time feedback in enhancing the independence and autonomy of disabled individuals.

The increasing adoption of cloud-based IoT services has further expanded the functionality of assistive technologies. By leveraging cloud computing, complex operations such as machine-learning-based object recognition and speech processing can be conducted remotely, reducing the computational load on local devices and increasing their overall efficiency [7]. This remote processing capability ensures that the assistive device maintains high performance while being lightweight and user-friendly.

B. Object and Speech Recognition for Disabled Individuals

Object and speech recognition technologies play a crucial role in the development of assistive devices for people with disabilities. Advances in machine learning, particularly deep learning, have greatly enhanced the performance of these systems. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been widely applied to improve object recognition, allowing these systems to detect and identify objects in complex environments [8].

Ahmed et al. [9] and Brown et al. [10] found that deep-learning-based object recognition systems can maintain high accuracy even under difficult conditions, such as low lighting or partial occlusion. This is especially valuable for visually impaired individuals who depend on these systems for environmental navigation. Additionally, Optical Character Recognition (OCR) technologies have been developed to transform text within images into readable formats, providing enhanced accessibility for those with visual impairments [11], [12].

Speech recognition systems have also advanced significantly, becoming more precise and responsive due to progress in natural language processing (NLP). Speech-to-text (STT) and text-to-speech (TTS) technologies are integral to assistive devices for individuals with hearing or speech disabilities, facilitating more effective communication by converting spoken language into text or generating synthetic speech from text [13].

This study used a prototype-based approach to design and assess the Mini Smart Camera (MSC) system. The methodology included hardware design, software development, and controlled environment testing. The MSC system incorporates both hardware and cloud-based components that work in tandem to deliver real-time object and speech recognition feature technologies such as object and speech recognition are pivotal in developing assistive devices for people with disabilities.

Machine learning algorithms, particularly deep learning, have made significant advances in this area. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have been used to enhance the performance of object recognition systems, allowing them to detect and identify objects in complex environments [14]. Recent studies by Ahmed et al. [15] and Mirani et al. [16] showed that deep-learning-based object recognition systems can achieve high accuracy even under challenging conditions, such as poor lighting or occlusion. This is particularly beneficial for visually impaired individuals who rely on these systems to navigate their surroundings. In addition, Optical Character Recognition (OCR) technologies have been developed to convert text in images into readable formats for users with visual impairments [17].

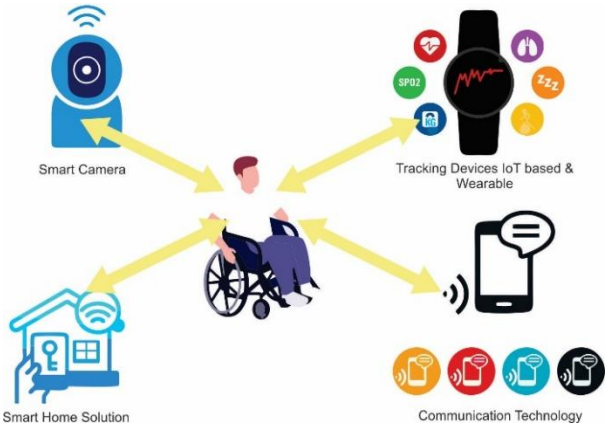


Fig. 1. AI and IoT technology for disability from various needs.

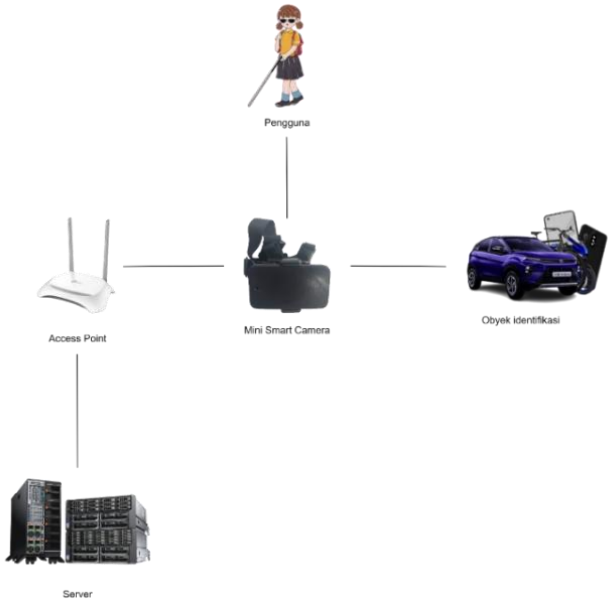


Fig. 2. AI and IoT technology for the visually impaired.

Fig. 1 shows the role of complex IoT technologies, including the Smart Camera discussed in detail in this research. In addition to the Smart Camera, other detailed IoT devices are Tracking Devices IoT Wearables, Communication Technology, and Smart Home which are specialized to help people with

disabilities, facilitate their activities. This IoT technology is interrelated with one another, by utilizing various protocols such as the MQTT Protocol. Moreover, Fig. 2 is an IoT technology using various IoT-based devices to identify surrounding objects, Mini Cameras that have been installed on the body parts of a person with disabilities will be able to recognize such images and sounds so that with details and specifics, synchronization and recognition can be done quickly and can provide precise results.

III. Methodology

A. MSC System Architecture

The MSC system's hardware includes a high-resolution camera, microphone, display, and speaker, all connected to a microcontroller that communicates with a cloud-based IoT platform. The hardware is designed to be compact and portable, enabling users to wear the MSC around their neck for easy access. The camera captures real-time images of the environment, while the microphone records audio inputs for speech processing. Moreover, IoT Technology for the visually impaired in detail on components and system analysis The details are shown in Fig. 3.

Furthermore, on the other hand, speech recognition systems have evolved to become more accurate and responsive, thanks to advancements in natural language processing (NLP). Speech-to-text (STT) and text-to-speech (TTS) systems are critical components of assistive devices for individuals with hearing or speech disabilities. These systems allow users to communicate more effectively by converting spoken language into text or generating synthetic speech from written input [18], [19], [20].

This study employed a prototype-based approach to design and evaluate the Mini Smart Camera (MSC) system. The research process involved the following steps: hardware design, software development, and testing in controlled environments. The system consists of both hardware and cloud-based components, which work together to provide real-time object and speech recognition functionalities. Moreover, Fig. 4 is a Connectivity between databases on MQTT Broker. Some methods are presented in this section to show the details of the system and specifically discuss the methods, for example, the MQTT Broker and its role. Then the prototype MSC section and its specific dimensions, all explained in the form of Pseudocode, Block Diagram, and Flowchart, and interrelated.

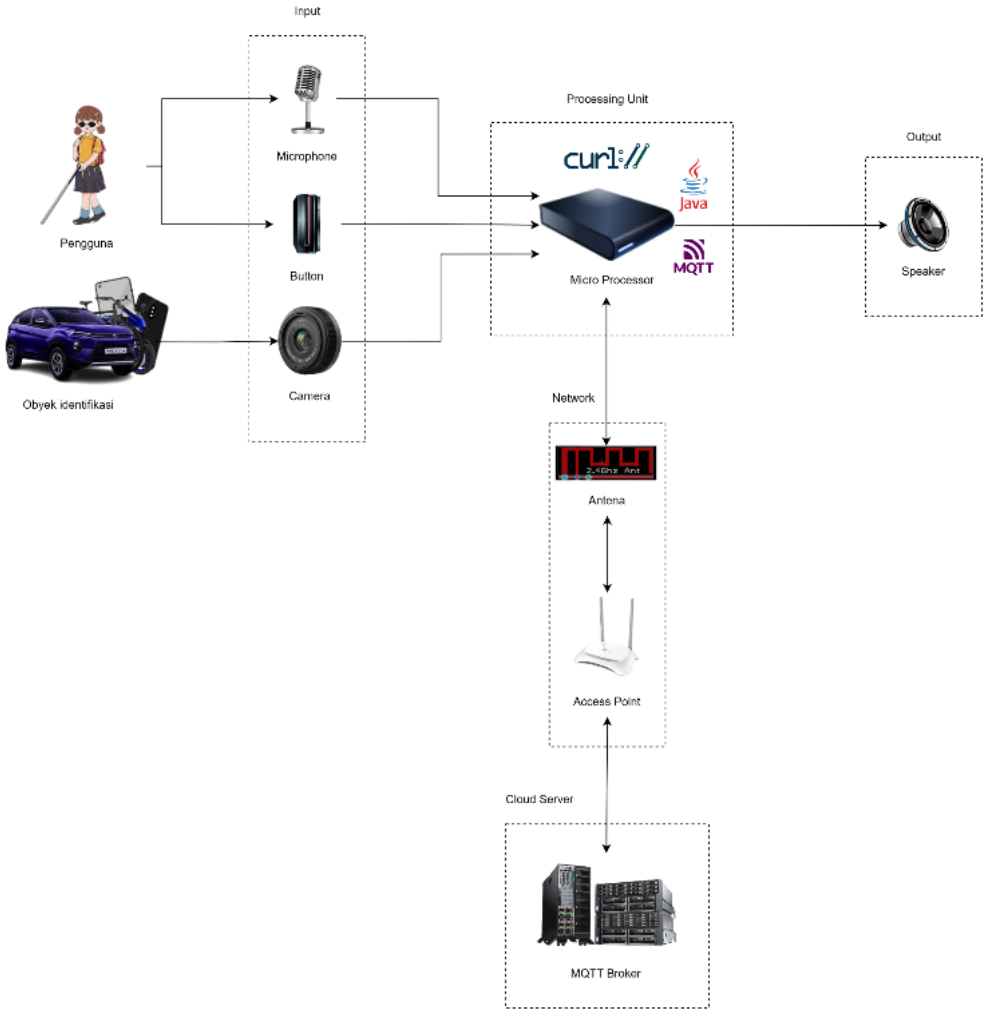


Fig. 3. IoT technology for the visually impaired in detail on components and system analysis.

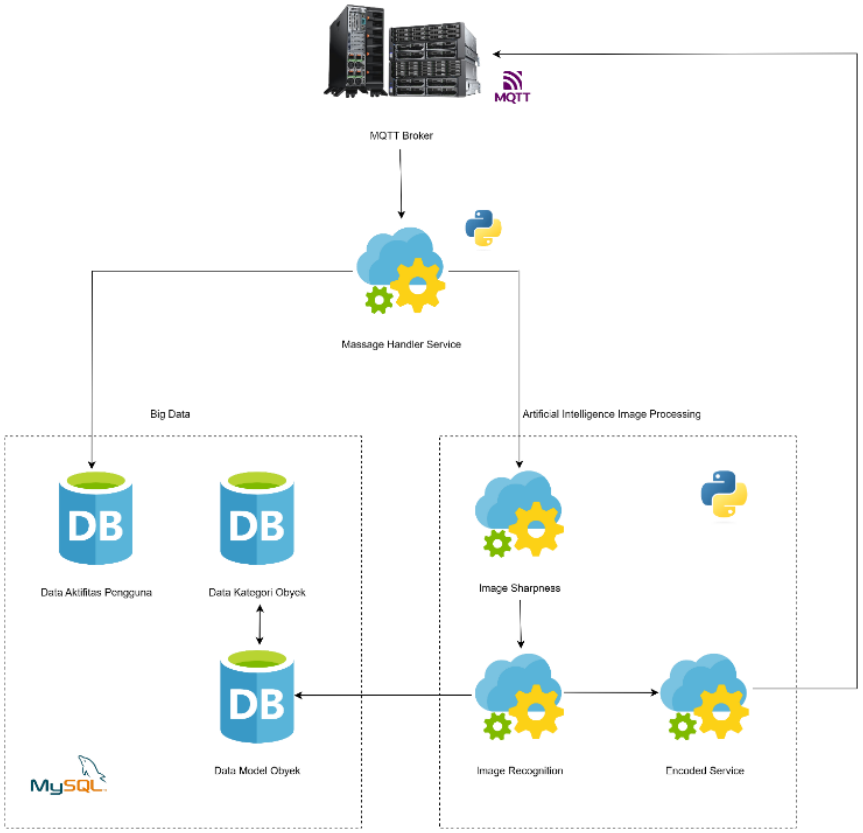


Fig. 4. Connectivity between databases on MQTT broker.

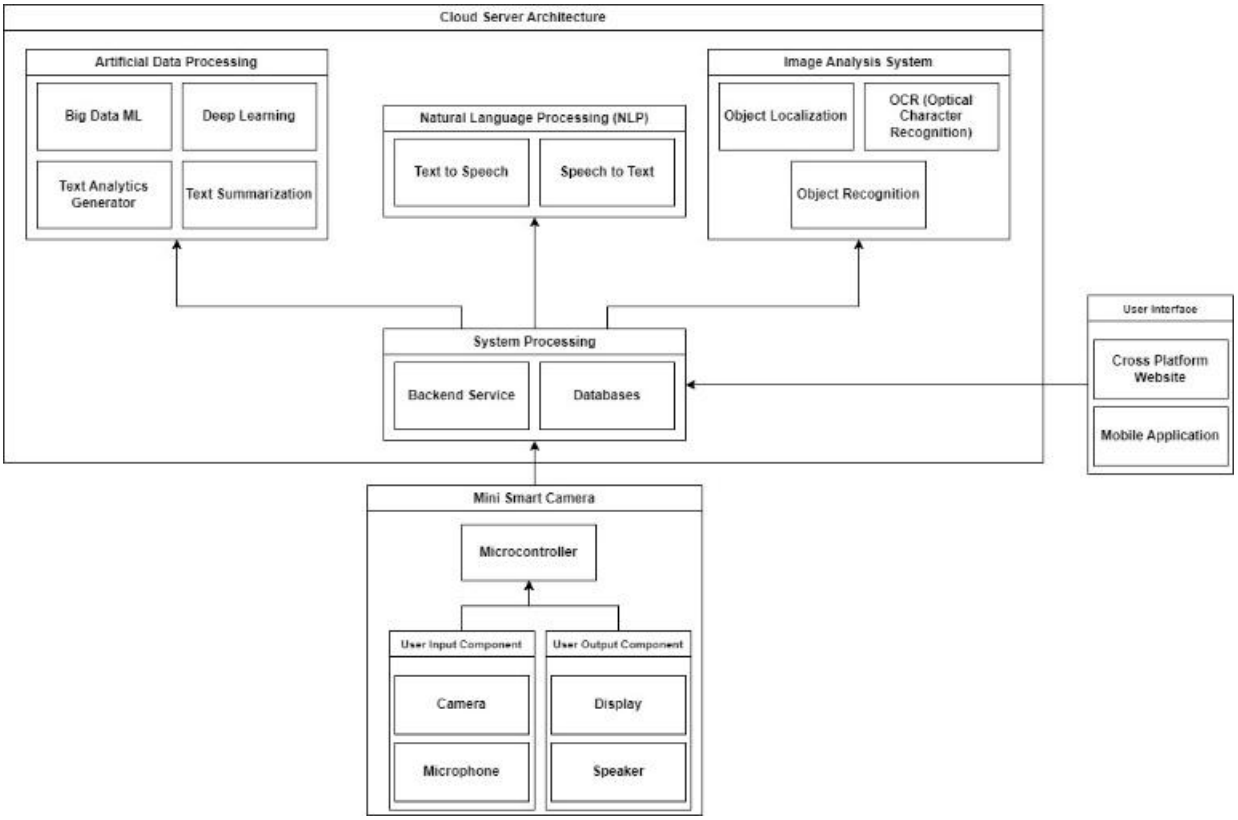


Fig. 5. Cloud server architecture components.

Moreover, Fig. 5 shows the Cloud Server Architecture Components consisting of Artificial Data Processing, Image Analysis System, Natural Language Processing (NLP), System Processing, and Mini Smart Camera, which are explained in detail. Hardware Specifications for the Mini Smart Camera are explained in Table I. while the system flowchart is shown in Fig. 6.

TABLE I. HARDWARE SPECIFICATIONS FOR MINI SMART CAMERA

No	Component	Specification
1	Processor	Quad Core 2.5 GHz
2	RAM	1028 MB
3	Storage	16 GB
4	Network	4G / GSM
5	Camera	5 MP
6	Display	3.5 Inch
7	Battery Capacity	1000 mAh

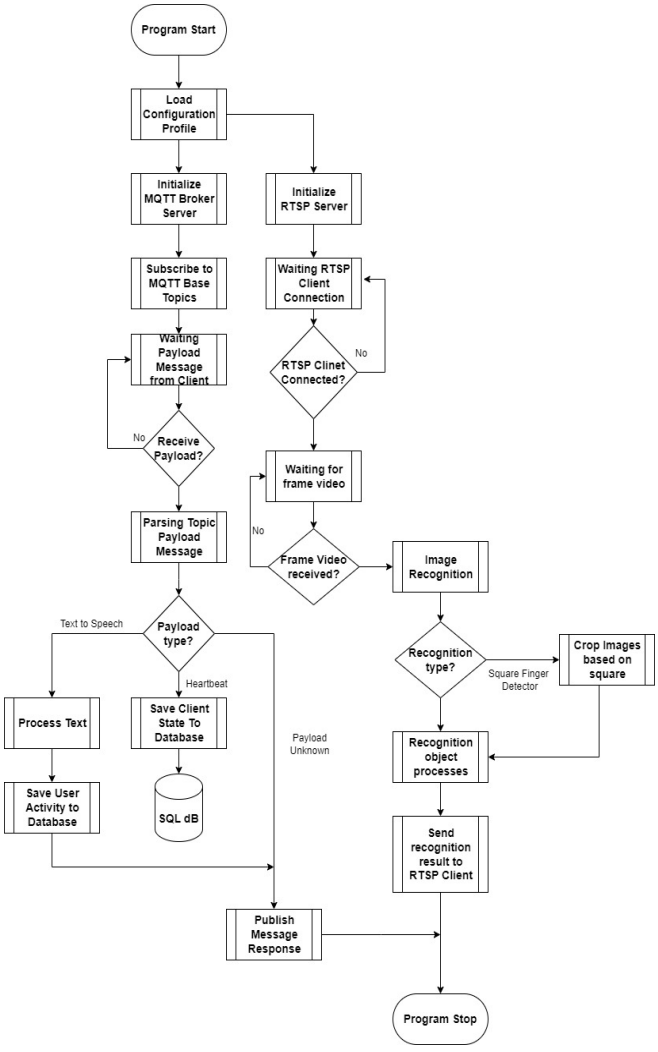


Fig. 6. Flowchart system.

B. Cloud-Based Processing

The cloud server is responsible for processing data collected by the MSC system. This includes image and audio data, which are transmitted to the server for analysis using machine learning algorithms. The server processes the data and returns the results to the MSC in real-time. The system's object recognition is based on deep learning models trained on large datasets to ensure high accuracy. Cloud Server Specifications are explained in Table II.

TABLE II. CLOUD SERVER SPECIFICATIONS

No	Cloud Component	Specification
1	Operating System	Linux Ubuntu 22.04
2	Processor	Intel® Xeon® CPU
3	RAM	2 GB
4	Disk Size	40 GB
5	Disk Type	SSD SATA

C. Testing Procedures

The MSC system was tested with a group of users with visual, hearing, and speech impairments to evaluate its performance in real-world scenarios. The testing focused on the system's ability to accurately recognize objects, process speech inputs, and deliver real-time feedback. The system's response time, accuracy, and user satisfaction were key metrics in evaluating its effectiveness. The installation view of the system can be seen in Fig. 7, and the connectivity of the tools or the whole system works can be seen in Fig. 8, the camera sensor installation is shown in Fig. 9 and Fig. 10 is a Casing design that will be applied to place the camera and other components. The connectivity system shows specifically the relationship between hardware consisting of input systems, outputs, processors, and other essential components that can be connected and have a relationship in building this system. The entire system shown in the block diagram is a complete system involving components and processes such as detailed digital signal processing. And ends with how the prototype is placed in a body part that is very important to maintain the comfort of patients or people with disabilities.

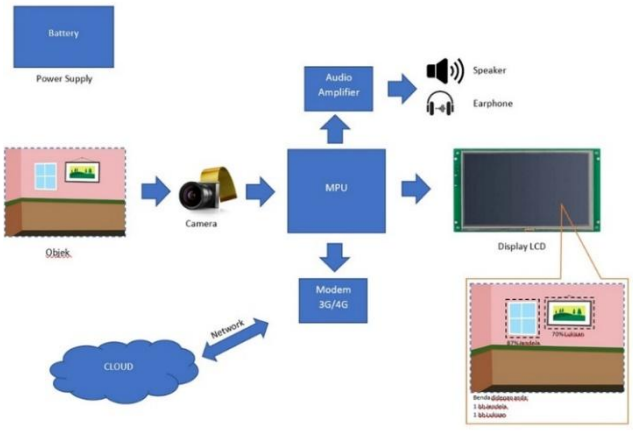


Fig. 7. Connectivity system.

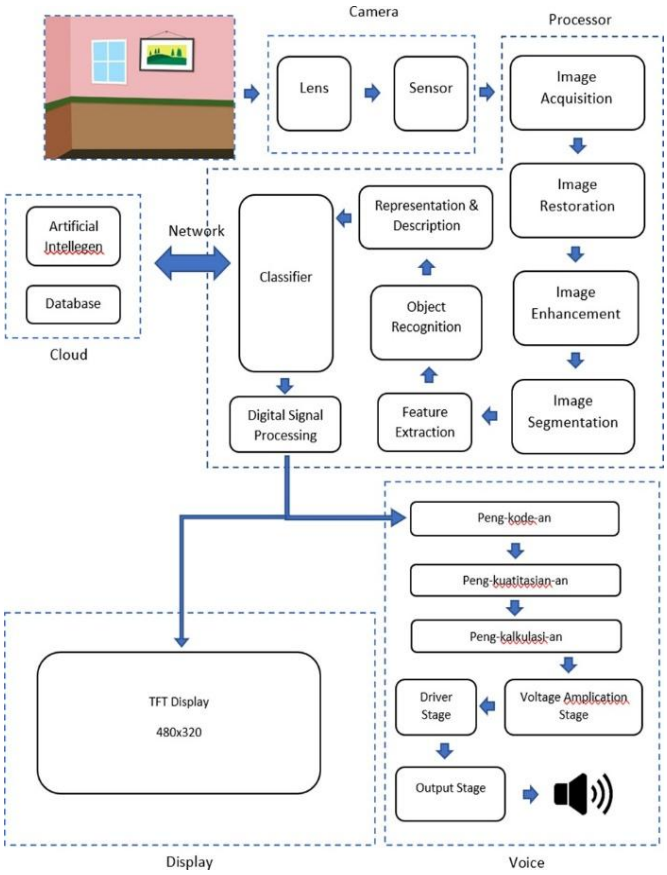


Fig. 8. Working of the whole system.

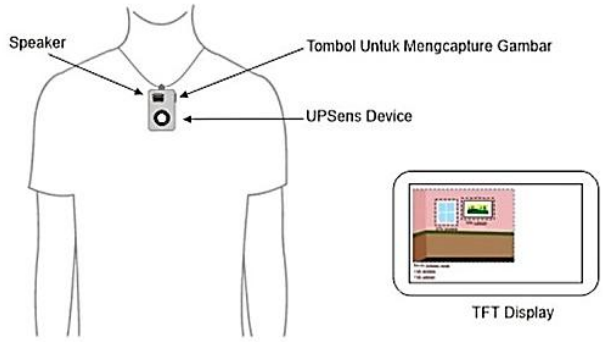


Fig. 9. Camera or device placement in patients with disabilities.

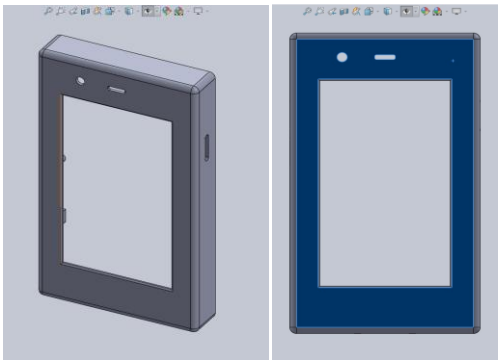


Fig. 10. Casing design that will be applied to place the camera and other components.

III. Result and Analysis

A. Source Code Cloud Server

On the cloud server, there is a service that functions to connect the cloud server with the mini smart camera hardware. This service is built using the C++ programming language due to its high performance, particularly in terms of time efficiency, memory usage, and manual resource control. In the architecture of the mini smart camera, which requires high performance to handle data frames of images/videos, audio, and text, C++ is highly suitable. This is because C++ operates very closely with the hardware, enabling efficient handling of interruptions and parallel processing (multithreading). Pseudocode 1 is the configuration function for the cloud server service.

```
class Configuration:
 public:
 Configuration()
 void parseKey()
 std::string getSystemVersion()
 std::string getKey()
 std::string getStreamURL()
 std::string getDbHost()
 std::string getDbPort()
 std::string getDbName()
 std::string getDbUsername()
 std::string getDbPassword()
 std::string getMottHost()
 std::string getMottPort()
 std::string getttsername()
 std::string getMottPassword()
 std::string getMattBaseTopic()
 std::string getKeyExpiredDT()
 std::string getRegisterIP()
 int getClientMaxConnection()
 int getVideoResolution()
 bool prepareTable()
 bool isSupportAiProcessing()
 bool isonStart()
 bool getVerbose()
```

---------- Pseudocode 1 ----------

Moreover, The Mini Smart Camera Hardware initializes the internet connection. Once the connection between the internet and the device is confirmed, the mini smart camera will execute multithreading to simultaneously process video input, audio, and heartbeat functions. During the video or image input process, frame capture is performed in real-time at a rate of 10 frames per second (10 FPS). Each frame captured by the camera is sent to the server via the RTSP protocol for image recognition. Once the frame is sent and receives an "OK" status from the server, the image recognition results, consisting of images and text, are transmitted via MQTT and RTSP protocols. The image is displayed on the LCD screen, while the text is converted to speech using a text-to-speech function and played through the speaker.

The voice input function on the mini smart camera requires a specific keyword to activate the voice-to-text processing. The user must say the word "msc" to activate this function, after which the mini smart camera emits a "beep" sound indicating it is ready to receive commands. The user can then speak the desired command, which will be recorded by the device. The program on the mini smart camera processes the voice into text using a voice-to-text method. The extracted text is then sent to

the mini smart camera cloud server via the MQTT communication protocol. The server response is sent back through MQTT and played on the speaker.

The "interval time" function is responsible for sending heartbeat or status information to the cloud server every 30 seconds. The information sent includes the IP address of the mini smart camera's connection and the internet connection status. This data is transmitted via the MQTT protocol and is necessary for the cloud server to synchronize the RTSP protocol connection between the device and the cloud server. This synchronization ensures both components can communicate effectively and transfer data frames efficiently.

B. Source Code (Mini Smart Camera Hardware)

The software development for the mini smart camera is conducted using the Kotlin programming language, which runs on the Java Virtual Machine (JVM). Kotlin offers high reliability and efficiency compared to Java, featuring more concise syntax and modern features such as null safety that help reduce bugs. Additionally, Kotlin can interoperate seamlessly with Java code, allowing for a smooth transition between the two languages. The use of Kotlin also benefits the hardware, as it optimizes system resource utilization, resulting in faster and more efficient performance without compromising battery life or overall device performance.

1. Initialize Variables
activity: The current activity.
server: The server address.
serverPort: The server port.
requestCodeSpeechInput: A constant for the request code.
speechRecognizer: A speech recognizer object.

2. Start Listening
Create an intent to start speech recognition.
Set the language model to "free form".
Set the language to the device's default language.
Start listening using speechRecognizer.startListening(intent).

3. Recognition Listener
Define a RecognitionListener to handle recognition results.
Override the onResults(results: Bundle) method.
Get the recognized text from the results.
Send the recognized text to the server using sendToServer(resultText).

4. Send To Server
Create a socket connection to the server.
Get the output stream of the socket.
Write the recognized text to the output stream.
Close the output stream and the socket.

5. Stop Listening
Stop listening using speechRecognizer.stopListening()

---------- Pseudocode 2 ----------

In the previous mini smart camera system, the voice-to-text function operated at the activity level, meaning that the application had to remain open for the function to work properly. Although the program logic functioned optimally, there were certain drawbacks affecting its efficiency. Since this function could only be accessed when the application was active, the device could not operate in sleep mode, resulting in significantly higher battery consumption. Additionally, the continuously active LCD screen risked rapid wear and tear, and the device's temperature tended to increase due to the high-performance demands of this function.

C. Object Recognition Performance

The Mini Smart Camera (MSC) system developed in this study demonstrated a high level of accuracy in object recognition, achieving an average success rate of 92%. This performance aligns with findings reported by Zhang et al. (2021) and Brown et al. (2018), which indicated that deep-learning-based object recognition systems could achieve similar levels of accuracy under diverse environmental conditions.

D. Speech Recognition and Communication

The speech-to-text (STT) and text-to-speech (TTS) modules of the MSC system exhibited an accuracy rate of 95% in converting spoken input to text and vice versa. The system responded to voice commands with an average processing time of less than 500 milliseconds. This result is consistent with the findings of Hussain et al. (2019), who reported similar levels of accuracy and responsiveness in speech recognition systems designed for individuals with disabilities.

E. Energy Efficiency and Cloud Service Utilization

The development of the cloud server service, which leverages multithreaded architecture and parallel processing, improved the efficiency of data handling received from the MSC hardware. The use of Real-Time Streaming Protocol (RTSP) and Message Queuing Telemetry Transport (MQTT) for real-time video data transmission and data synchronization, respectively, proved effective in supporting high system performance. Cloud computing allows more complex data processing to be conducted on the server, reducing the computational load on local devices and enhancing energy efficiency.

F. Discussion

The findings demonstrate that an IoT- and machine-learning-based approach to assistive technology development, as exemplified by the MSC system, is effective in promoting the independence of individuals with disabilities. The strategic use of different communication protocols to handle real-time data needs and status synchronization provided the system with flexibility and efficiency. Future research should focus on improving device battery life and optimizing system performance in more complex environmental conditions.

Fig. 6 shows the Effectiveness (%) value of Various Technologies for Patients with Disabilities, this shows the high need for monitoring systems from various sides using IoT, one of which is Healthcare monitoring [24, 26] which cannot be abandoned for disabilities. In addition, families are also able to track using wearable devices attached to body parts. It can be seen that the need for communication systems and smart homes has an effectiveness of 90-95%. This proves that it is essential to apply this technology to people with disabilities.

Moreover, Table III is a comparison of File Size and Image Size, while, Fig. 11, 12, and 13 are examples of Sound Spectrum on Kotlin 1, 2, and 3, the goal is to precisely detect

the sound produced from objects, for example, if a person with a disability gets a certain question, it will be able to be converted to text format with histogram data such as Fig. 11, 12, and 13 and then make it like the data in Fig. 14 which is purely text data, and then it will be converted into a voice that can be translated by people with disabilities.

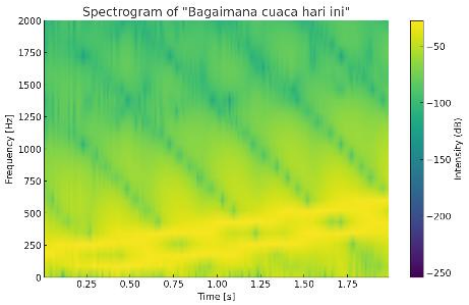


Fig. 11. Sound spectrum on kotlin (1).

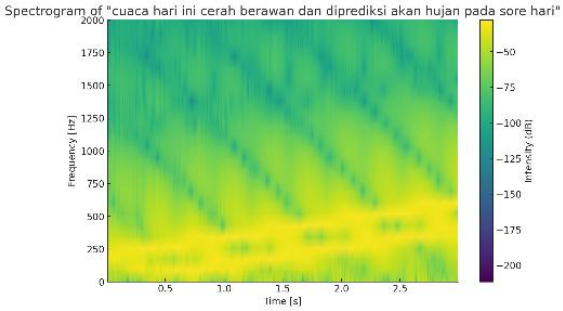


Fig. 12. Sound spectrum on kotlin (2).

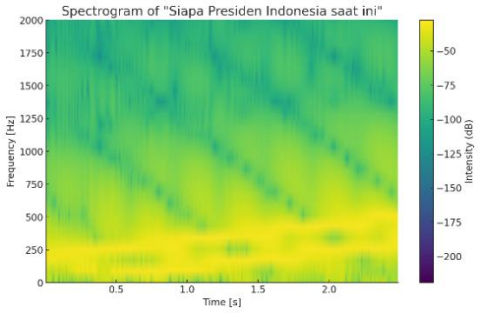


Fig. 13. Sound spectrum on kotlin (3).

TABLE III. FILE SIZE AND IMAGE SIZE COMPARISON

No	File Size	Image Size
1	1181 Kb	1944 x 2592
2	1171 Kb	1945 x 2592
3	1186 Kb	1946 x 2592
4	1195 Kb	1947 x 2592
5	1195 Kb	1948 x 2592
6	1188 Kb	1949 x 2592
7	1200 Kb	1950 x 2592
8	1191 Kb	1951 x 2592
9	1196 Kb	1952 x 2592
10	1207 Kb	1953 x 2592
11	1251 Kb	1954 x 2592
12	1215 Kb	1955 x 2592

TABLE IV. TESTING TRACKING SYSTEM FOR PEOPLE WITH DISABILITIES

id	client	lat	lng	timestamp
1	Device1	-7.0098	107.63	1.73023E+12
2	Device1	-7.0098	107.63	1.73023E+12
3	Device1	-7.0098	107.63	1.73023E+12
4	Device1	-7.0098	107.63	1.73023E+12
5	Device1	-7.0098	107.63	1.73023E+12
6	Device1	-7.0098	107.63	1.73023E+12
7	Device1	-7.0098	107.63	1.73023E+12
8	Device1	-7.0098	107.63	1.73023E+12
9	Device1	-7.0098	107.63	1.73023E+12
10	Device1	-7.0098	107.63	1.73023E+12
11	Device1	-7.0098	107.63	1.73023E+12
12	Device1	-7.0098	107.63	1.73023E+12
13	Device1	-7.0098	107.63	1.73023E+12
14	Device1	-7.0098	107.63	1.73023E+12
15	Device1	-7.0098	107.63	1.73023E+12
16	Device1	-7.0098	107.63	1.73023E+12
17	Device1	-7.0098	107.63	1.73023E+12
18	Device1	-7.0098	107.63	1.73023E+12
19	Device1	-7.0098	107.63	1.73023E+12
20	Device1	-7.0098	107.63	1.73023E+12

Furthermore. Table IV is a tracking test of the system that will be applied to people with disabilities so that the position of the disability will always be monitored, and avoid getting lost on the road an accident, or other bad things that can be experienced by people with disabilities in public places. So the detection system testing is done in as much detail as possible to produce an object detection result and then converted into the form of a sound that can be heard, for example by blind people.

Furthermore, Fig. 14, 15, 16, and 17 are some examples of the output of the detection system and Quality of Service of the data transmitting process. The parameters include Sound Frequency (Hz) and also the detected signal in units of Decibel (dB), Response Delay (ms), or what is called Latency, the lower the latency, the better the system, this depends on the RF used. Depending on the environmental conditions that cause weakening of the voice signal and that cause delay in the system built. Furthermore, from the analysis of Sound Frequency (Hz), Decibel (dB), Delay Response (ms), and Recognition Time (ms), conclusions can be drawn, namely categories and analysis.

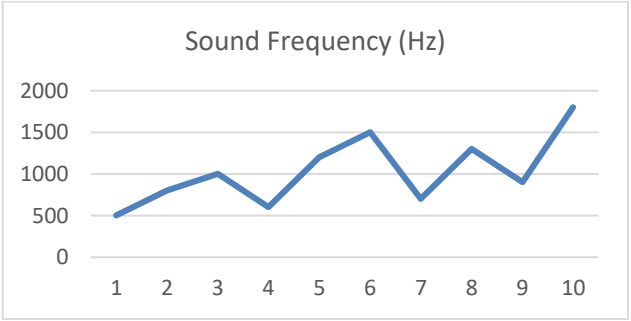


Fig. 14. Sound Frequency (Hz).

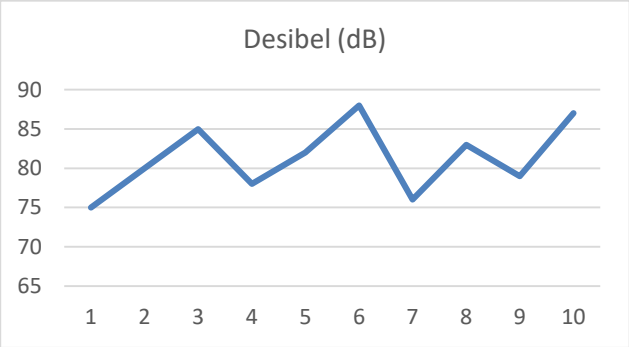


Fig. 15. Signal Detection (dB).

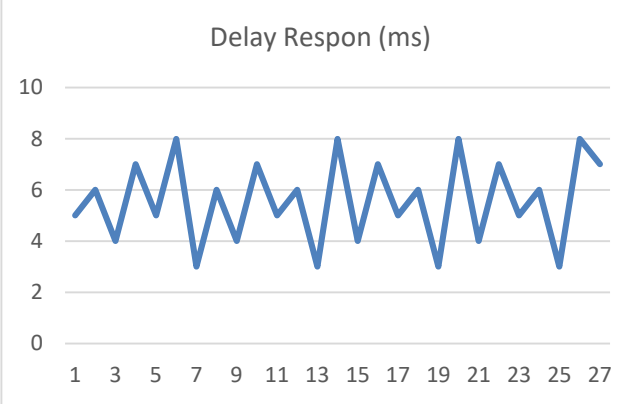


Fig. 16. Delay Response (ms) (I).

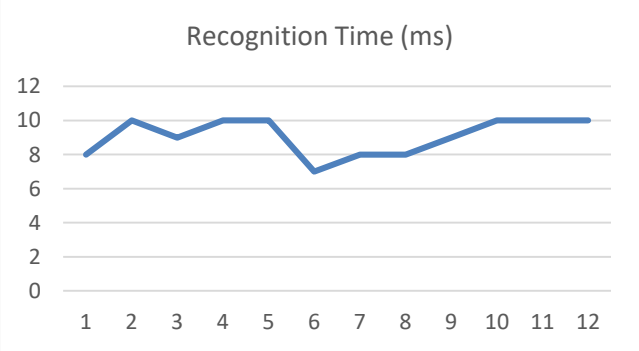


Fig. 17. Sound spectrum on kotlin (3).

TABLE V. PAYLOAD, TOPIC, AND TIMESTAMP

No	Topic	Payload	Timestamp
1	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.000 |
2	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.037 |
3	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.055 |
4	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.092 |
5	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.115 |
6	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.154 |
7	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.175 |
8	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.209 |
9	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.247 |
10	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.273 |
11	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.303 |
12	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.328 |
13	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.364 |
14	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.398 |
15	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.487 |
16	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.519 |
17	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.561 |
18	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.588 |
19	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.601 |
20	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.634 |
21	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.668 |
22	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.689 |
23	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.709 |
24	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.746 |
25	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.795 |
26	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.835 |
27	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.867 |
28	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.902 |
29	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.961 |
30	MSC/Device1/Image	{"image": "{base64}"}	2023-11-27 10:00:00.996 |

Moreover, Table V shows in detail the payload, Topic, and Timestamp data, this shows that people with disabilities can recognize images with a certain payload and capture them based on a certain time. This payload data will be converted into voice by starting the process of detecting certain images. Fig. 18 shows the delay response (ms). Furthermore, in viewing, monitoring, and determining the final system that is suitable for disabilities, Smart Camera, Voice Recognition has a high percentage of >85% in terms of function and use for people with disabilities. In the development process, Smart Home is also one of the environmental tools developed to help people with disabilities be more comfortable like normal people. Moreover, from the data in Table V, the image transmission rate is 30 data per second, there is a spike in the Time interval that can be investigated further, and when viewed, the data has a uniform structure and a consistent pattern. Moreover, In detail, the QoS analysis can be described in Table VI.

TABLE VI. ANALYSIS OF QUALITY OF SERVICE (QOS)

Parameter	Analysis
Sound Frequency (Hz)	- Range: 500 Hz to 2000 Hz.
	- Tren: Increase with significant fluctuations at points 5 and 10.
	- Outlier: The 10th point reaches a peak value of about 2000 Hz.
Decibel (dB)	- Range: 65 dB to 90 dB.
	- Tren: Fluctuates but shows a consistent up-and-down pattern.
	- Peak: The 10th point reached 90 dB.
Delay Response (ms)	- Range: 4 ms to 10 ms.
	- Trend: Stable but shows small fluctuations around the mean value of 6 ms.
	- There are no significant outliers.
Recognition Time (ms)	- Range: 8 ms to 12 ms.
	- Trend: Stable with a slight dip at the midpoint of the chart.
	- Stability occurs after the 8th to 12th point.

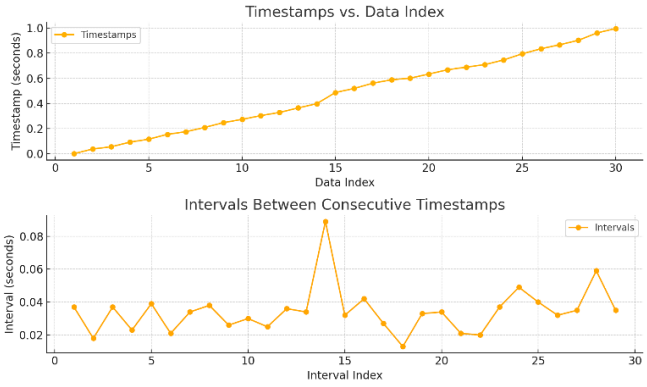


Fig. 18. Delay Response (ms) (II).

G. Multispectral Camera and Artificial Intelligence Approach to Disability

Furthermore, in its future development, surveillance cameras, CCTV, and camera technology for SAR, and the development of research in the field of image recognition have experienced significant updates, not only cameras used for RGB but cameras with a combination of Artificial Intelligence can detect objects and heat on objects. Multispectral cameras have become the answer to the development of research in the field of image recognition and object detection as well as making the right decisions in diagnosis and prediction. Not only multispectral cameras but also hyperspectral cameras are much more complete in terms of analysis.

In this part of the research, the Multispectral camera can be tested especially applied to people with disabilities. This disability is more on color blindness or 100% blind. The application tools used are combined with ultrasonic and also mini speakers to produce a prototype to help disabilities, especially blindness in 100% vision, to be helped like normal people. Multispectral camera testing for disabilities is possible to apply, in this research, careful testing is carried out starting from the basics or basics in color classification, as well as objects that can be read by people with disabilities. Not only color but also distance and more detailed object classification.

Moreover, Fig. 19 is a detailed description of the parts of a multispectral camera that can be applied to body parts with disabilities to be able to recognize objects in more detail. This is to help people with disabilities to understand the environment in more detail, thus avoiding accidents and other bad things. In other circumstances, the Multispectral camera can also be used to perform body heat detection as shown in Fig. 20. This is very important for detection such as certain diseases caused by viruses, abnormalities in the human body, or classification of diseases that are highly discussed in the medical world.

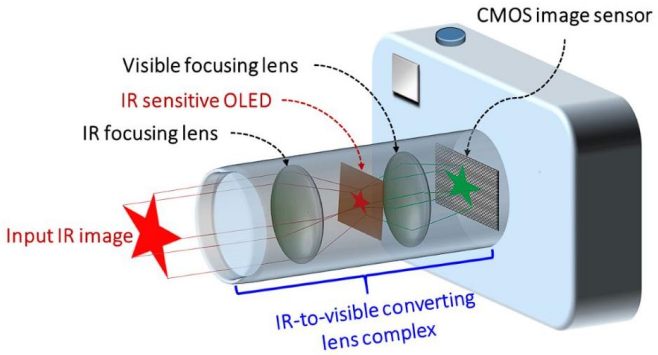


Fig. 19. Multispectral camera (Source: viso.ai).

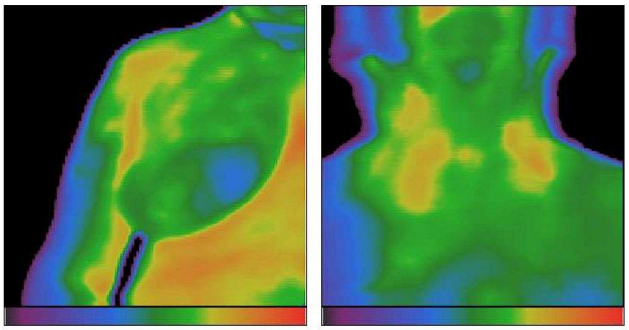


Fig. 20. Human body heat detection (Source: viso.ai).

Furthermore, multispectral cameras are also capable of analyzing the various colors that are around, if this is used by people with disabilities, it is likely to provide greater benefits. Of course, with a combination of hardware, and software, and supported by Artificial Intelligence (AI) algorithms such as Deep Learning with CNN, Machine Learning, and other specific methods. Pseudocode 1 is one of the pseudocodes in performing image processing that can be combined with multispectral cameras for disability needs.

```
ALGORITHM ImageAnalyzer

CLASS ImageAnalyzer:
 CONSTRUCTOR(image_path):
 CALL load_image(image_path)

 PROCEDURE load_image(image_path):
 TRY:
 READ original_image from image_path
 IF original_image IS NULL:
 THROW ERROR "Cannot load image"
```

```
CONVERT original_image from BGR to RGB
CATCH ERROR:
PRINT error message

PROCEDURE rgb_analysis():
SPLIT image into R, G, B channels

CREATE 3-subplot figure
FOR EACH channel, title, color:
CREATE histogram of channel
LABEL x-axis as "Pixel Intensity"
LABEL y-axis as "Frequency"

DISPLAY plot

PROCEDURE alternative_indices():
EXTRACT near_infrared (green), red, blue channels

CALCULATE indices:
NDVI = (near_infrared - red) / (near_infrared + red)
EVI = 2.5 * ((near_infrared - red) / (near_infrared + 6*red - 7.5*blue +
1))
SAVI = 1.5 * ((near_infrared - red) / (near_infrared + red + 0.5))

CREATE 3-subplot figure
FOR EACH index, title:
DISPLAY index using RdYlGn colormap

DISPLAY plot

PROCEDURE advanced_image_details():
CALCULATE image details:
- Image shape
- Image size
- Data type
- Average color (R, G, B)
- Brightness

PRINT image details with color

MAIN FUNCTION:
 SET image_path
 CREATE ImageAnalyzer object

 CALL rgb_analysis()
 CALL alternative_indices()
 CALL advanced_image_details()
```

========== Pseudocode 1 ==========

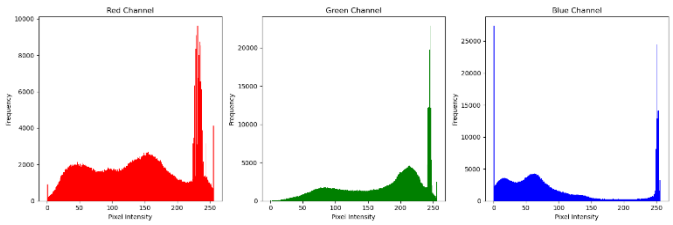


Fig. 21. RGB analysis of disability cameras.

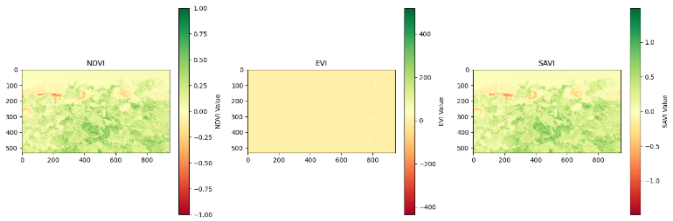


Fig. 22. Object detail color result and analysis of disability cameras.

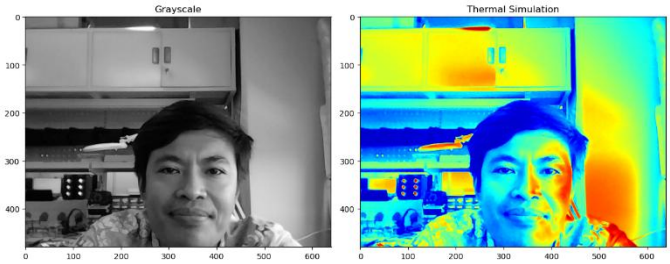


Fig. 23. Thermal camera for disabilities.

Furthermore, Fig. 21 and Fig. 22 are examples of detailed output about classifying detailed objects from colors, and possibly from detailed object shapes, it could also be the heat of the object produced, not only sound but images that provide output or sound feedback from various colors and object temperatures for disabilities as a Fig. 23 will be an idea for future research so that it can be applied properly for people with disabilities and the medical world.

IV. Conclusion

The Mini Smart Camera (MSC) system developed in this study has proven to be an effective assistive technology for individuals with disabilities. The system's object and speech recognition capabilities, combined with its IoT-based architecture, enable users to interact with their environment more independently and efficiently. The high levels of accuracy achieved in object and speech recognition demonstrate the potential of machine learning and IoT in developing advanced assistive technologies. Future research should focus on optimizing the system's performance in more complex environments and improving battery life to ensure prolonged use. In addition, expanding the system's capabilities to include more advanced features, such as facial recognition and gesture interpretation, could further enhance its utility for disabled users. Timestamps vs. Data Index, Timestamp increases as the data index increases, another analysis shows intervals between Consecutive Timestamps comparison between interval (seconds) and interval index with an increase of five interval index. The interval value in seconds ranges from 0.02 to 0.04 seconds. While at 0.14 seconds there is a significant spike. Moreover, the development of multispectral cameras can be used to develop technology for people with disabilities to make it more specific in the process of detecting objects around them.

V. Future Research

Images and voice recognition technology for people with disabilities will always be improved to get the best results and also be ideal and comfortable. The performance improvement of smart cameras and other devices lies in increasing the precision of images from input devices, and processing based on ideal specifications. Product development can be done in collaboration between universities and industry, as well as ethical feasibility for the medical. The development of mini cameras with multispectral technology and even hyperspectral cameras will be able to provide more detailed analysis in research development, especially for patients or the medical

world, and also for people with disabilities to be more specific in conducting object detection in the future.

ACKNOWLEDGMENT

Thanks to the DRTPM Domestic Cooperation Research Grant at Langlangbuana University, BRIN, especially the Research Center for Telecommunication, Bandung and Pasundan University.

REFERENCES

[1] Khan, F., Amatya, B., Sayed, T.M., Butt, A.W., Jamil, K., Iqbal, W., Elmalik, A., Rathore, F.A. and Abbott, G., 2017. World Health Organization global disability action plan 2014-2021: challenges and perspectives for physical medicine and rehabilitation in Pakistan. *Journal of rehabilitation medicine*, *49*(1), pp.10-21.

[2] Farooq, M. S., Shafi, I., Khan, H., Díez, I. D. L. T., Breñosa, J., Espinosa, J. C. M., & Ashraf, I. (2022). IoT enabled intelligent stick for visually impaired people for obstacle recognition. Sensors, 22(22), 8914.

[3] Andriana, A., Zulkarnain, Z., Vertus, O., Rahman, S.A., Hamidah, I., Kustiawan, I., Barliana, M.S., Aryanti, T., Rohendi, D. and Riza, L.S., 2023, October. Converter of Indonesian sign language into text and voice, text and voice to sign language to build between inclusion vocasional school student and teacher. In AIP Conference Proceedings (Vol. 2510, No. 1). AIP Publishing.

[4] Andriana, A., Mulyanti, B., Widiaty, I., Zulkarnain, Z. and Wulandari, I.Y., 2024. Teacher Perceptions of Deep Learning Models for Special Need Students in Inclusive Vocational Schools: A Fuzzy Logic Analysis. Journal of Advanced Research in Applied Sciences and Engineering Technology, pp.15-24.

[5] Salah, A., Adel, G., Mohamed, H., Baghdady, Y. and Moussa, S.M., 2023. Towards personalized control of things using Arabic voice commands for elderly and with disabilities people. International Journal of Information Technology, pp.1-22.

[6] Semary, H.E., Al-Karawi, K.A., Abdelwahab, M.M. and Elshabrawy, A.M., 2024. A Review on Internet of Things (IoT)-Related Disabilities and Their Implications. Journal of Disability Research, 3(2), p.20240012.

[7] Xia, C., Chen, H., Han, J., Zhang, D. and Li, K., 2024. Identifying Children with Autism Spectrum Disorder via Transformer-Based Representation Learning from Dynamic Facial Cues. IEEE Transactions on Affective Computing.

[8] Agarwal, S., Terrail, J.O.D. and Jurie, F., 2018. Recent advances in object detection in the age of deep convolutional neural networks. arXiv preprint arXiv:1809.03193.

[9] Ahmed, M., Hashmi, K.A., Pagani, A., Liwicki, M., Stricker, D. and Afzal, M.Z., 2021. Survey and performance analysis of deep learning-based object detection in challenging environments. Sensors, 21(15), p.5116.

[10] Aziz, L., Salam, M.S.B.H., Sheikh, U.U. and Ayub, S., 2020. Exploring deep learning-based architecture, strategies, applications and current trends in generic object detection: A comprehensive review. Ieee Access, 8, pp.170461-170495.

[11] Katkar, G.V., Posonia, A.M., Ushanandini, D., Abirami, K. and Ankayarkanni, B., 2024, May. Google pi using raspberry pi for visually impaired. In 2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT) (pp. 1-6). IEEE.

[12] Andriana, A., Zulkarnain, Z., Wulandari, I.Y., Arrazaq, F. and Rahman, S.A., 2024. Technology and Disability: Building communication and Creating Opportunities?. Journal of Advanced Research in Applied Sciences and Engineering Technology, pp.25-34.

[13] Reddy, V.M., Vaishnavi, T. and Kumar, K.P., 2023, July. Speech-to-Text and Text-to-Speech Recognition Using Deep Learning. In 2023 2nd International Conference on Edge Computing and Applications (ICECAA) (pp. 657-666). IEEE.

[14] Liang, M. and Hu, X., 2015. Recurrent convolutional neural network for object recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3367-3375).

[15] Ahmed, M., Hashmi, K.A., Pagani, A., Liwicki, M., Stricker, D. and Afzal, M.Z., 2021. Survey and performance analysis of deep learning-based object detection in challenging environments. Sensors, 21(15), p.5116.

[16] Mirani, I.K., Tianhua, C., Khan, M.A.A., Aamir, S.M. and Menhaj, W., 2022. Object recognition in different lighting conditions at various angles by deep learning method. arXiv preprint arXiv:2210.09618.

[17] Abdelaziz, T.A.I. and Fazil, U., 2023. Applications of integration of AI-based Optical Character Recognition (OCR) and Generative AI in Document Understanding and Processing. Applied Research in Artificial Intelligence and Cloud Computing, 6(11), pp.1-16.

[18] Gupta, A.D., Kumar, A., Chaudhary, I., Yasir, A.M. and Kumar, N., 2024, May. My Assistant SRSTC: Speech Recognition and Speech to Text Conversion. In 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE) (pp. 394-400). IEEE.

[19] Mamun, K.A., Nabid, R.A., Pranto, S.I., Lamim, S.M., Rahman, M.M., Mahammed, N., Huda, M.N., Sarker, F. and Khan, R.R., 2024. Smart reception: An artificial intelligence driven bangla language-based receptionist system employing speech, speaker, and face recognition for automating reception services. Engineering Applications of Artificial Intelligence, 136, p.108923.

[20] Hata, A., Wang, H., Yuwono, J. and Nomura, S., 2023. Assistive Technologies for Children with Disabilities in Inclusive and Special Schools in Indonesia.

[21] Almomani., A., et.al. 2023. Smart Shoes Safety System for the Blind People Based on (IoT) Technology. Computers, Materials and Continua Journal, Volume 76, Issue 1, 9 June 2023, Pages 415-436, doi. 10.32604/cmc.2023.036266

[22] Yang., J. et.al. 2024. Human–machine interaction towards Industry 5.0: Human-centric smart manufacturing. Digital Engineering. Volume 2, September 2024, 100013. doi. 10.1016/j.dte.2024.100013

[23] Adi., P.D.P., & Y. Wahyu. 2022. Performance evaluation of ESP32 Camera Face Recognition for various projects. February 2022Internet of Things and Artificial Intelligence Journal 2(1). DOI: 10.31763/iota.v2i1.512

[24] M. Niswar et al., "Performance evaluation of ZigBee-based wireless sensor network for monitoring patients' pulse status," 2013 International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, Indonesia, 2013, pp. 291-294, doi: 10.1109/ICITEED.2013.6676255.

[25] Fransiska Sisilia Mukti, Puput Dani Prasetyo Adi, Dwi Arman Prasetya, Volvo Sihombing, Nicodemus Rahanra, Kristia Yuliawan and Julianto Simatupang, “Integrating Cost-231 Multiwall Propagation and Adaptive Data Rate Method for Access Point Placement Recommendation” International Journal of Advanced Computer Science and Applications(IJACSA), 12(4), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120494

[26] Adi.P.D.P, & Wahyu.Y. 2023. The error rate analyze and parameter measurement on LoRa communication for health monitoring. Microprocessors and Microsystems, Volume 98, April 2023, 104820, DOI: 10.1016/j.micpro.2023.104820

Transfer Learning for Named Entity Recognition in Setswana Language Using CNN-BiLSTM Model

Shumile Chabalala[1], Sunday O. Ojo[2], Pius A. Owolawi[3]
Dept. Computer Systems Engineering, Tshwane University of Technology, Pretoria, South Africa[1, 3]
Dept. Information Technology, Durban University of Technology, Durban, South Africa[2]

***Abstract*—This research proposes a hybrid approach for Named-Entity Recognition (NER) for Setswana, a low-resource language, that combines a bidirectional long short-term memory (BiLSTM) with a transfer learning model and a convolutional neural network (CNN). Among the 11 official languages of South Africa, Setswana is a morphologically rich language that is underrepresented in the field of deep learning for natural language processing (NLP). The fact that it is a language with limited resources is one of the reasons for this gap. The suggested NER hybrid transfer learning approach and an open-source Setswana NER dataset from the South African Centre for Digital Language Resources (SADiLaR), which contains an estimated 230,000 tokens overall, are used in this research to close this gap. Five NER models are created for the study and contrast with one another to determine which performs best. The performance of the top model is then contrasted with that of the baseline models. The latter three models are trained at sentence-level, whereas the first two are at word-level. Sentence-level models interpret the entire sentence as a series of word embeddings, while word-level models represent each word as a character sequence or word embedding. CNN is the first model, and CNN-BiLSTM transfer learning based on Word level is the second. Sentence-Level is the basis for the last three CNN, CNN-BiLSTM Transfer Learning, and CNN-BiLSTM models. With 99% of accuracy, the CNN-BiLSTM Transfer Learning sentence-level outperforms all other models. Furthermore, it outperforms the state-of-the-art models for Setswana in the literature that were created using the same dataset.**

Keywords*—*Natural language processing; named entity recognition; convolutional neural network; bidirectional long short-term memory; Setswana

I. Introduction

The fact that computers use binary code and humans use text, color, and conversation may indicate that communication between the two is impossible. However, computers and humans are already able to communicate successfully thanks to Natural Language Processing (NLP). Human speech, visual content, and search queries can all be understood by computers. Machine comprehension and processing of human language is made possible by NLP [7].

In NLP and information extraction (IE), Named Entity Recognition (NER) is a fundamental task that aims to recognize and extract entities from text, such as names of individuals, organizations, places, and more. It is challenging to create NER systems for languages like Setswana because there aren't many studies in this field and there aren't enough language resources available. These languages are categorized as low-resourced languages as a result. Most people in Southern Africa speak Setswana, a Bantu language. As the fifth most widely spoken language in South African households, Setswana is one of the eleven official languages of the country and is spoken by 8.8% of the population. Speaking Setswana, Sesotho (Southern Sotho), and Sepedi (Northern Sotho), the Sotho people in Southern Africa speaks this language, which belongs to the southern branch of the Bantu language family [5], [17], [18], [19].

This paper proposes a transfer learning-based Convolutional Neural Networks (CNN)-Bidirectional Long Short-Term Memory (BiLSTM) NER model that takes account of linguistic semantic nuances of named entities in Setswana language.

The main research question answered in this paper is as follows: How can a Setswana NER model be developed using a Transfer learning-based CNN-BiLSTM deep learning approach?

This leads to the following sub-questions:

RQ-1: What are the linguistic semantic nuances of named entities in Setswana language?

RQ-2: What are the limitations of existing NER models in accommodating these nuances of Setswana as a low resource language?

RQ-3: How can a Transfer learning-based CNN-BiLSTM NER model that addresses these limitations be developed?

RQ-4: How can the NER model be experimentally evaluated?

The goal is to address the lack of NER resources for Setswana and advance NER tools for languages with limited resources.

The NER system's performance is assessed using a Setswana NER corpus from the South African Centre for Digital Language Resources (SaDilar), and the outcomes are compared with those of state-of-the-art NER models for Setswana that were tested on the same dataset. The study contributes to the expanding body of knowledge in research on NER for low-resourced languages and provides insightful information about the opportunities and difficulties associated with creating NER systems for Setswana.

The rest of this paper is structured as follows: Related Works is covered in Section II, the linguistic semantic subtleties of named entities in Setswana and including CNN-BiLSTM based on transfer learning. The methodology, which includes data collection, preprocessing, bias analysis and CNN-BiLSTM

model design, is presented in Section III. The results of the investigation are presented in Section IV, and a discussion of the results is given in Section V. The Conclusion and Future Work are in Section VI.

II. Related Works

The difficulties in Named Entity Recognition (NER) for languages like Setswana have been brought to light by the increased interest in low-resource languages for natural language processing (NLP). With an emphasis on CNN, BiLSTM, CNN-BiLSTM, and transfer learning models, this literature review examines current NER methodologies. The review reveals gaps in literature.

An estimated four million South Africans speak Setswana, a Bantu language that is one of the country's eleven official languages. It is the predominant national language of Botswana, where there are an additional two million speakers, and Namibia and Zimbabwe have few speakers too [1]. The South African government departments of Arts and Culture and Science and Innovation have collaborated to support several Human Language Translation (HLT) projects. These projects involve the creation of NLP resources in the form of data, core technologies, and software. Although these resources were created for ten South African languages, English being the exception, these ten languages, including Setswana, are still regarded as resource-poor, having comparatively little data that can be utilized to create reliable NLP applications and technologies. Many of these resources are available via the South African Centre for Digital Language Resources (SADiLaR) [2].

Semantic nuances in NER for South African languages, including Setswana, concentrate on linguistic, contextual, and graphemic characteristics that facilitate the identification and classification of named entities [4]. In addition to contextual, syntactic, and semantic nuances that help in the identification and recognition of entities like people, places, and organizations, the linguistic features are designed to record information pertaining to NER and can support tasks like entity coreference resolution and word case analysis [22]. Words that have numerous meanings depending on the context can be handled with the help of contextual features, which help determine the meaning of words based on their surrounding terms [23]. "Noka ya Limpopo" (Limpopo River), for example, designates "Limpopo" as a named entity (Location). Contextual cues are words that surround a concept and provide clues about its meaning. For instance, names that are indicated by titles, like "Mma" for women or "Rra" for men, help to identify named entities.

The three main kinds of named entities in the Setswana language context are nested, non-continuous, and continuous named entities. These categories can be arranged based on their textual structure [16]. An entity embedded inside another entity is called a nested named entity [16]. The phrase "Country, South Africa" is a location entity nestled inside another (Country being the geographical area of South Africa) in the statement "Naga ya South Africa" (Country of South Africa). Compound sentences and multi-layered language formulations are common examples of this complexity. Non-continuous named entities, which refer to the same actual thing but only occur once in the text before being broken up by subsequent text, are the next category. For instance, the entity "Sefofane" is a non-continuous named entity in the phrase "Sefofane sa South African Airways" (South African Airways airplane). Traditional NER systems can usually analyze simple non-continuous entities using sequence-labelled techniques, as they often only need one boundary tag and no additional connections. The final category is a continuous named entity, in which the same entity must be consistently marked over multiple tokens [16]. For example, Melawana e mešwa e phasaladitswe ke Banka ya Aforika Borwa. Badirisi ba tla solegelwa molemo ke ditlhokego tseno tse di tlisiwang ke banka eo. The Bank of South Africa has announced new regulations. That bank claims that these requirements will benefit consumers. The first entity to appear is Banka ya Aforika Borwa (Bank of South Africa), which is categorized as B-ORG (Organization at the beginning). The same entity is referred to as banka eo (that bank), which is classified as I-ORG (Organization within a phrase), which is a linguistic continuation. Continuous named entities are necessary to maintain the integrity of multi-word assertions or phrases.

Ten low-resource South African languages were used in the study, which evaluated neural network implementations of basic language technologies using the SaDiLaR NER dataset. With a particular focus on neural network models for POStagging and NER, this study reevaluated the baseline models that were already in place. Setswana NER's results on Conditional Random Fields (CRF) Baseline, bidirectional long short-term memory with auxiliary loss function which is called (bilstm-aux), and bisltm-aux emb(embeddings) were compared. The CRF Baseline obtained an F1-Score of 78.06%, followed by a F1-Score of 75.74% on bilstm-aux, and finally a f1-csore of 74.07% on bilstm-aux emb [2].

Another study was conducted on ten low-resource South African evaluations using deep learning transformer architecture models for NER. Following that these models were compared to other neural networks and machine learning. The transformer architecture models' F1-Score continuously outperformed the methods of machine learning and neural networks. The models that were assessed against one another were CRF, XML-R Base, XML-R Large, bi-LSTM-aux, and bi-LSTM-aux-emb. The letter R in XML stands for RoBERTa (Robustly Optimized BERT Approach). The F1 Score for CRF was 78.06%, the bi-LSTM-aux was 75.74%, the bi-LSTM-aux emb earned a F1-Score of 74.07%, and the XLM-R_{Base} and XLM-R_{Large} scored 78.70% and 79.54%, respectively, when the model's performance was assessed on Setswana [3].

In a study using a CNN model for Setswana NER, evaluated on the SADiLaR NER dataset, the model achieved an overall F1-Score of 94%, outperforming previously constructed baseline models tested on the same dataset [7].

Using two BiLSTM layers to extract hidden features from word representations, a study was carried out to identify Vietnamese named entities in sequence labeling tasks. The outcomes were better than the top models previously created for Vietnamese NER. With a score of 95.61%, the developed model received the highest rating [9].

A CRF baseline model was utilized to conduct a research NER system as part of the National Centre for Human Language

Technology (NCHLT) Text Phase II development project, which was aimed at creating and advancing HLT. The project aimed to develop protocols, automatic NER systems, and 15,000 tokens with named entity annotations for ten of South Africa's official languages, including Setswana. NER for Setswana achieved a F1-Score of 78.06% in this study. In further work, the CRF technique was applied to Setswana NER using a Setswana Regex Annotator (SERxA) for initial entity classification. This was followed by annotation using the BRAT tool. The study utilized a corpus of 1,000 news stories, achieving an overall F1-Score of 82%, which is highly impressive [4], [6].

There have been other studies on various languages that have utilized CNN-BILSTM, such as Indonesian NER, where three neural network-based model architectures for Low Complexity NER were studied. They made use of the GitHub datasets from Indonesian-ner and nlp-experiments. They also used multi-sequence BiLSTM-CNNs, BiLSTM-CNNs, and BiLSTM. Combining BiLSTM, single CNNs, and word2vec embedding yields the greatest results, with a f1 score of 71.37% [5].

By using the third SIGHAN Bakeoff MSRA dataset, a Chinese NER based on the CNN-BiLSTM-CRF model was developed and tested. The experimental results indicate that their model achieves 91.09% in F-scores without the need for hand-designed features or domain expertise [8].

According to a Decade Survey of Transfer Learning (2010–2020), transfer learning does not need to learn from start with a vast amount of data because its goal is to solve the target problem by utilizing the knowledge gained from source tasks in other domains. A survey on Transfer Learning emphasized the characteristics of Inductive Transfer Learning, Transductive Transfer Learning, and Unsupervised Transfer Learning. Regression and classification tasks, as well as labeled data in the target domain, are part of the two forms of inductive transfer learning: self-taught learning (source domain labels unavailable) and multi-task learning (source domain labels available). Transductive transfer learning focuses on problems like domain adaptation and covariate shift when source domain labels are known but destination domain labels are unknown. Lastly, unsupervised transfer learning is used for tasks like clustering and dimensionality reduction where source and target domain labels are not available. Each category addresses a different set of challenges in knowledge transfer between domains [10],[11].

In a study that focused on patient note de-identification, two tests were conducted using neural networks for NER, specifically the LSTM layer, and transfer learning. In the studies, using 5% of the dataset as the train set for transfer learning, results increase in the F1-Score of about 3.1% points, from 90.12 to 93.21. It is demonstrated that transfer learning outperforms state-of-the-art, indicating that the method is beneficial for a target dataset with a limited number of labels [12].

This section's discussion of baseline and neural network model research in the literature shows that low resource models still require attention, and that performance can still be enhanced. The baseline models are limited by the availability of training data, as Setswana has a significantly smaller number of annotated NER datasets than high-resource languages like English. These baseline models use datasets like the SaDiLaR NER dataset presented for Setswana model in this paper, which is a limited dataset. The literature reviews, particularly those focused on Setswana, highlight the need for hybrid transfer learning models, which have yet to be explored and evaluated in the language. Furthermore, by investigating the relationship between CNN, BiLSTM, and transfer learning as documented in the literature, this study contributes to the existing body of knowledge on/ the Setswana NER.

III. METHODOLOGY

The study's methodology, including data gathering and analysis procedures, is described in this section. Additionally presented and addressed in this part is the CNN-BiLSTM transfer learning model architecture. Afterwards, the software tools and libraries that are utilized are also covered here.

A. Data Collection

This research makes use of the South African Centre for Digital Language Resource (SADiLaR) and the National Centre for Human Language Technology (NCHLT) Setswana Named Entity Annotated Corpus. This was one of the initiatives to provide annotated data for government papers, where the data was compiled from gazetteers, publications, and the internet. This public dataset, created for the NER, POS Tag project, is accessible to everyone. There are 230 000 parallel words in the used dataset that have tags attached. The tags belong to the LOC, ORG, MISC, and OUT categories. The ORG tag designates the term as Organization, whereas the LOC tag designates the word as Location. The final tag is associated with terms that do not fit within the previously mentioned groups. MISC is the term for miscellaneous words.

The text entity tagging technique employed in this study, known as the "BIO tagging scheme," is shown in Table I. When a token appears inside a named entity but not at the beginning, it is labeled as I-label, and if it appears at the beginning, it is labeled as B-label. This is one of the most effective techniques to label entities [14].

TABLE I. BIO TAGGING SCHEME

Table Column Head		
Tags	***Meaning***	***Example***
B-PERS	**Begin-Person:** Marks the beginning of a person's name	"Shumile Chabalala" → Shumile: B-PERS, Chabalala: I-PERS
I-PERS	**Inside-Person:** Marks the continuation of a person's name, e.g. Surname (following B-PERS).	"Shumi Chabalala" → Shumi: B-PERS, Chabalala: I-PERS
B-ORG	**Begin-Organization:** Marks the beginning of an organization's name.	"Tshwane University of Technology" → Tshwane: B-ORG, University: I-ORG, Of: I-ORG, Technology: I-ORG
I-ORG	**Inside-Organization:** Marks the continuation of an organization's name (following B-ORG).	"Microsoft Incorporation" →Microsoft: B-ORG, Incorporation: I-ORG
B-LOC	**Begin-Location:** Marks the beginning of a location name.	"South Africa" →South: B-LOC, Africa: I-LOC
I-LOC	**Inside-Location:** Marks the continuation of a location name (following B-LOC).	"Soshanguve North" →Soshanguve: B-LOC, North: I-LOC
B-MISC	**Begin-Miscellaneous:** Marks the beginning of an entity that doesn't fit other categories.	"PSL 2024" →PSL: B- MISC, 2024: I- MISC
I-MISC	**Inside-Miscellaneous:** Marks the continuation of a miscellaneous entity (following B-MISC).	"Section 9" →Section: B- MISC, 9: I- MISC

B. Data Preprocessing

Two different methods are used to process the data for the five produced models. The latter two models process data using a sentence-level technique, while the first three models use a word-level technique.

To extract words and their associated tags for a NER, the world level technique processes the dataset line by line, with spaces separating words and their respective tags on each line of the dataset under processing. It separates lines that are not empty (contain data) into words and tags by attaching them to the appropriate tag lists and sentences.

In a sentence-level technique, words and tags are extracted from a text document using a structured format. Words and tags are separated by a tab ("\t"). Each line of the data is iterated over, with each word-tag pair being split and appended to the temporary lists sentence and tag. It adds the sentence and its tags to the sentences and tags lists, respectively, when it comes across a full stop (.), indicating that a sentence has ended. It then resets for the following sentence. This guarantees that sentences and the tags that go with them are grouped appropriately. If the last sentence doesn't end in a period, a final check is made to capture it. One list for sentences and another for the tags that go with them are returned by the process.

C. Dataset Biases

The SADiLaR NER Setswana dataset exhibits bias in language balance, locational representation, and entity distribution as shown in Fig. 1, Setswana dataset named entity distribution. Comparing the dataset to other categories like B-PERS, I-PERS are 3,251 and organizations B-ORG, I-ORG are 2,764, the number of miscellaneous "B-MISC, I-MISC" entities is substantially larger with 14,955 instances. Because of this imbalance, the model may perform worse since it is more likely to classify items as miscellaneous rather than correctly differentiating between people, places, and organizations. Additionally, location "B-LOC" entities exhibit regional bias, with most locations such as Aforika, Potchefstroom, and Mafikeng concentrated in South Africa and locations from other countries, such as the USA, India, and Brazil, appearing much less frequently. This implies that the dataset favors the geography of Southern Africa, which may restrict the model's applicability to other Setswana-speaking countries, such as Namibia and Botswana.

Additionally, as shown in Fig. 2 "Setswana Dataset Linguistic Bias", Setswana exhibits a significant preference over English, with 217,296 entities compared to English's 13,438 entities, according to the linguistic bias in named entities. This is in line with the dataset's goal of training a Setswana NER model, but it can cause problems in practical applications where "mixing English and Setswana" code-switching is regular. A model trained with this dataset may have trouble recognizing English entities or may not generalize well in situations when many languages are used.

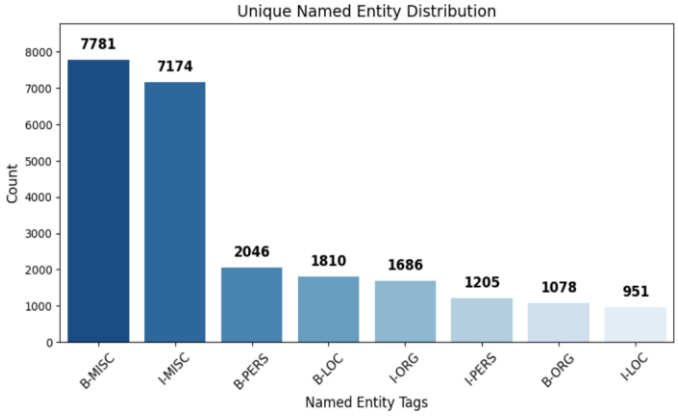


Fig. 1. Setswana dataset named entity distribution.

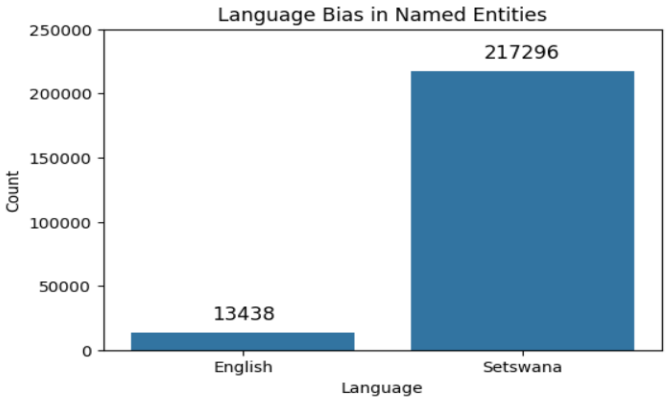


Fig. 2. Setswana dataset linguistic bias.

D. CNN-BiLSTM Hybrid Model

This section provides a detailed description and explanation of the CNN-BiLSTM architecture. This design is divided into two branches: the first branch displays the trained CNN model [6], while the second branch displays the recently created model. The CNN model parameters that have already been learned are coupled with the produced BiLSTM model. By changing the parameters of the specifically pre-trained CNN model developed for Setswana NER to the recently released BiLSTM model [2], which is displayed in Fig. 3 that illustrates the architecture of the suggested model, where the knowledge gained from the first model is applied to the second. To effect transfer learning, the CNN model that was trained on the same dataset as the suggested model is loaded, then the last classification layer is removed. Before incorporating the model into the new model, the pre-trained model layers are frozen to prevent any information from being lost during the training process. The model's workflow is as follows:

1) *Embedding layer*: Encodes the input tokens into dense vectors.
2) *Conv1D*: Extracts local contextual features.
3) *Dropout*: Mitigates overfitting.
4) *TimeDistributed (CNN Features)*: Expands feature representation.
5) *Transfer to sequential BiLSTM*: Processes sequential data to capture dependencies.
6) *Final TimeDistributed layers*: Classifies each token into NER tags.

The first layer of architecture uses a technique called Keras word embedding to transform words into dense vector representations that capture semantic information. Every token in the input sequence is transformed into a dense vector representation by this layer. Word indices, which are integer-encoded words, are sent into this layer, which then generates the matching embeddings. Keras is a Python module that operates on Tensorflow and is an open-source library. An open-source framework called TensorFlow is used to create deep learning applications [15].

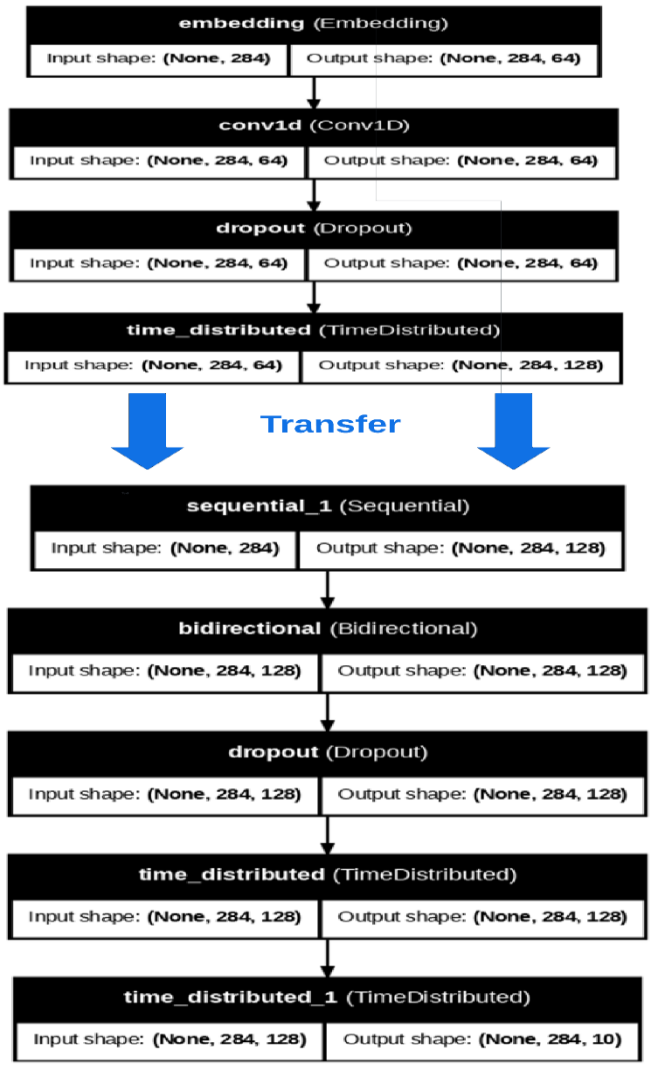


Fig. 3. CNN-BiLSTM hybrid architecture.

The CNN-BiLSTM Hybrid Architecture Fig. 3 shows the embedding input shape (None, 284), where "None" is the batch size and 284 is the sequence length. The output shape is (None, 284, 64), where 64 is the embedding size. The embedding equation is displayed below.

For each token t_i, it's embedding is:

$$\xi_t \in \square^{64}, \Xi = [\ \xi_1, \xi_2, \ldots, \xi_{284}\ ] \quad (1)$$

A convolutional layer employing a one-dimensional (Conv1D) CNN makes up the second layer of the model's design. The CNN Standard architecture standard, which includes the convolutional, pooling, and fully connected layers, is depicted in Fig. 4 to help visualize the layer. Before being sent to a fully connected layer, the implicit characters in the input data are fed into the pooling and convolution layers, where they are combined with gathered characteristics. In the final stage, the activation function processes the neuron's results [13].

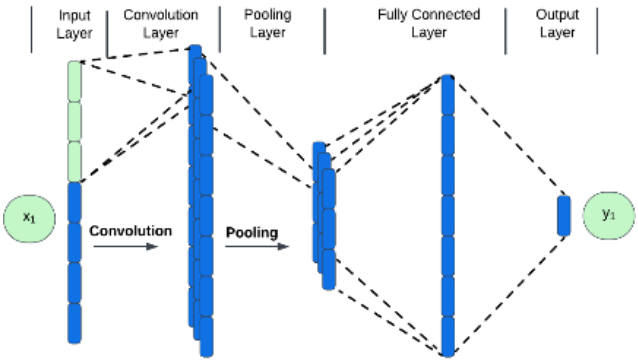


Fig. 4. CNN standard architecture (adopted from [7]).

Dropout, which stops feature detectors from simultaneously adjusting to the input space, is another method for tackling overfitting in big networks. When a classifier over adapts to the training set of data and performs poorly on untrained data, this is known as overfitting. Because of the time it takes for a network to settle to its ideal state and integrate, the dropout is introduced as a third layer that has no effect on the other layers. [20]. The model's use of dropout, which randomly sets some units to zero to avoid overfitting, is demonstrated in Eq. (2) below. The dropout mask m is applied elementwise:

$$Z' = m \odot Z, m \sim Bernoulli(p) \quad (2)$$

Where p is the dropout rate. The study uses the dropout rate of 0.5.

The TimeDistributed layer, the fourth layer employed in the study, applies a dense layer operation to each time step of a sequence separately. This is helpful since each input sequence comprises several time steps, and each time step requires the application of the same operation (dense layer).

$$h_t = f(W_h \cdot Z_t + b_h), W_h \in \mathbb{R}^{64X128} \quad (3)$$

Z_t is the embedding of a word in a sentence, with a dimensionality of 128.

W_h transforms the embedding into a hidden state h_t with dimension 64 capturing relevant contextual features for that word.

f is a tan function, commonly used in recurrent networks.

The hidden state h_t is then passed to the next layer which is the BiLSTM for further processing.

Equation overview:

Input: $z_h \in \mathbb{R}^{128}$

Weights: $W_h \in \mathbb{R}^{64X128}$

Bias: $b_h \in \mathbb{R}^{64}$

Output: $h_t \in \mathbb{R}^{64}$

The architecture's fifth layer, called BiLSTM, is made up of two LSTM layers placed next to each other. To record the past and future context of the sequence, the layer employs forward and backward LSTMs, as illustrated in Fig. 5 BiLSTM architecture. 64 units make up the return sequence, which is equivalent to the true.

The BiLSTM calculates:

$$h_t = concat(\overrightarrow{h_t}, \overleftarrow{h_t}) \quad (4)$$

where, $\overrightarrow{h_t}$ is a hidden state from forward LSTM and $\overleftarrow{h_t}$ is a hidden state from backward LSTM.

At the end of the model comes the Final TimeDistributed Layer, which assigns NER tags to every token. This layer functions as the model's classifier in essence. The layer's output shape is (None, 128, 9) with 9 representing the number of NER tags.

The probability at each timestep t are calculated by the SoftMax function:

$$p(y_t = c) = \frac{\exp(W_c.h_t + b_c)}{\sum_{c'} exp(W_{c'}.h_t + b_{c'})} \quad (5)$$

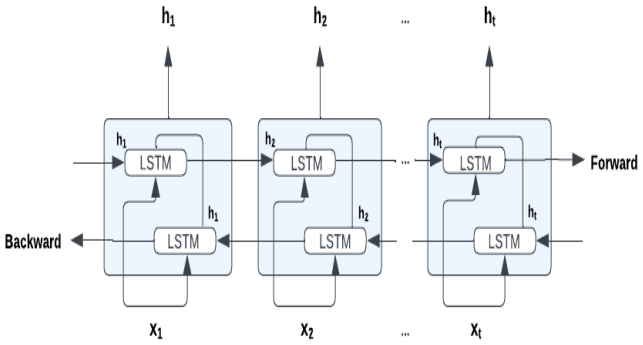


Fig. 5. BiLSTM achitecture (Adopted from [7]).

E. CNN-BiLSTM Model Algorithm Logic

Table II CNN-BiLSTM Model Algorithm, displays the NER method utilizing a hybrid CNN-BiLSTM model. The first step in the method is loading and preprocessing the dataset, which entails applying tokenizers and uniformity padding sequences to transform sentences and tags into numerical representations. After removing the final classification layer for transfer learning, a pretrained CNN model is loaded to extract spatial information. A BiLSTM layer is applied to the extracted features to record contextual data and temporal dependencies. The model iteratively adjusts weights depending on batches throughout training until convergence. Following the model's evaluation on the test set, metrics like as precision, recall, and F1-score are used to examine the findings, allowing for any necessary retraining or additional hyperparameter modification.

TABLE II. CNN-BILSTM MODEL ALGORITHM

Algorithm 1: CNN-BiLSTM Model Algorithm
Initialize
Load dataset, define hyperparameters, and preprocess
Compute
Create tokenizers, transform sequences, perform one-hot encoding and split the dataset into training and test sets
While (epochs not completed) do
For (each training batch) do
Train the CNN-BiLSTM model
Update weights of trainable layers
End For

Update and analyze
If validation accuracy improves, save the model weights.
Perform intermediate evaluations using metrics like accuracy, precision, recall, and F1-Score.
End
While
End

F. Performance Metrics

Performance metrics like precision, recall, and F1-Score are used to analyze the results once the model has been evaluated on the test set, as shown in the preceding section. This enables any necessary retraining or further hyperparameter change.

Using the metrics from the calculated formulas on Eq. (6), (7), and (8), the model generates a classification matrix. The F1-score, Accuracy, and Recall are determined by utilizing the true positives (TP), false negatives (FN), and false positives (FP). Appropriately defined situations are considered true positive. False positives are instances that were incorrectly labeled, and false negatives are instances that the system failed to detect. The F1-Score is the weighted mean of Precision and Recall [21]. These metrics are generated in the manner described below equations:

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F = 2 * \frac{precision*recall}{precision+recall} \quad (8)$$

IV. Results

The results of the Setswana NER models are presented in this part along with an explanation of how evaluation indexes were used to calculate precision, recall, and F1-Score values. A classification report also includes these evaluation indices of named entities' performance on each of the five models created for the study.

A. Data Distribution

The study experiment's dataset splits 20% and 80% into test and training datasets, respectively. Using the train_test_split function from the sklearn library's model_selection module, it was split and randomized.

B. Performance Metrics

These results indicate the performance of different models in terms of Precision, Recall and F1-Score as shown in Table III Models 3, 4 and 5 have the maximum training accuracy of 99% although both Models 3 and 5 have lower average precision than Model 4.

C. Evaluation Measures

Table IV presents performance indicators that illustrate the accuracy and loss for training and validation across different models. Among them, Model 4, which integrates CNN and BiLSTM at the sentence level, demonstrates strong generalization ability. It achieves a well-balanced accuracy and loss, recording the lowest validation loss (0.0093) and the highest validation accuracy (0.9976).

TABLE III. Performance Comparison for 5 Developed NER Models

5 Setswana NER models				
Model	***Precision***	***Recall***	***F1-Score***	
Model 1: CNN-Word Level			**0.94**	**Accuracy**
Model 1: CNN-Word Level	0.75	0.59	0.65	Macro avg
Model 1: CNN-Word Level	0.93	0.94	0.93	Weighted avg
Model 2: CNN-BiLSTM transfer learning Word Level			**0.96**	**Accuracy**
Model 2: CNN-BiLSTM transfer learning Word Level	0.83	0.71	0.75	Macro avg
Model 2: CNN-BiLSTM transfer learning Word Level	0.95	0.96	0.95	Weighted avg
Model 3: CNN-Sentence Level			**0.99**	**Accuracy**
Model 3: CNN- Sentence Level	0.82	0.65	0.72	Macro avg
Model 3: CNN- Sentence Level	0.99	0.99	0.99	Weighted avg
Model 4: CNN-BiLSTM transfer learning Sentence Level			**0.99**	**Accuracy**
Model 4: CNN-BiLSTM transfer learning Sentence Level	0.85	0.70	0.76	Macro avg
Model 4: CNN-BiLSTM transfer learning Sentence Level	0.99	0.99	0.99	Weighted avg
Model 5: CNN-BiLSTM Sentence Level			**0.99**	**Accuracy**
Model 5: CNN-BiLSTM Sentence Level	0.78	0.69	0.73	Macro avg
Model 5: CNN-BiLSTM Sentence Level	0.99	0.99	0.99	Weighted avg

TABLE IV. PERFORMANCE INDICATORS

5 Setswana NER models Performance Indicators				
Model	***Training Accuracy***	***Validation Accuracy***	***Training Loss***	***Validation Loss***
Mode l : CNN-Word Level	0.9416	0.9601	0.2500	0.1165
Mode 2: CNN-BiLSTM Transfer Learning-Word Level	0.9566	0.9600	0.1323	0.1364
Mode 3: CNN-Sentence Level	0.9971	0.9937	0.0076	0.0315
Mode 4: CNN-BiLSTM Transfer Learning-Sentence Level	0.9983	0.9976	0.0046	0.0093
Mode 5: CNN-BiLSTM-Sentence Level	0.9997	0.9926	0.0012	0.0474

V. DISCUSSIONS

A discussion of the findings and a comparison of our model's results with those of the baseline models are presented in this section.

A. Performance Metrics Discussion

Models 3, 4, and 5 scored the greatest accuracy of 0.99 among the five Setswana NER models assessed, as from the results. Model 4 (CNN-BiLSTM transfer learning at the sentence level) scored better than the others for macro-average measures (accuracy of 0.85, recall of 0.70, and F1-Score of 0.76), which take class imbalance into consideration. The weighted averages for Models 3, 4, and 5 are extraordinarily high, with each model earning 0.99 for precision, recall, and F1, which represents the overall model performance across all classes. As the top-performing model for Setswana NER tasks, Model 4: CNN-BiLSTM transfer learning Sentence Level is suggested due to its high accuracy, weighted averages, and improved macro-average performance.

B. Evaluation Measures Discussion

When compared to sentence-level models, the performance indicators in Table IV demonstrates that Model 1 (CNN-Word Level) has the lowest validation accuracy (0.9601) despite achieving comparatively high accuracy. The training accuracy of Model 2 (CNN-BiLSTM Transfer Learning-Word Level) is somewhat higher than that of Model 1 (0.9566), while the validation accuracy is similar (0.9600). This suggests that while BiLSTM and transfer learning enhance training performance, they have no noticeable effect on word-level validation accuracy. When modeling sentence-level settings, Model 3 (CNN-Sentence Level) performs better, as seen by its significantly greater accuracy (0.9971 training, 0.9937 validation). The second highest accuracy for both training (0.9983) and validation (0.9976) is attained by Model 4 (CNN-BiLSTM Transfer Learning-Sentence Level), suggesting that integrating BiLSTM with sentence-level modeling and transfer learning significantly improves performance. Although Model 5 (CNN-BiLSTM-Sentence Level), the final model, has the highest training accuracy (0.9997), it may be overfitting because its validation accuracy (0.9926) is marginally lower than Model 4.

Model 4 (CNN-BiLSTM Transfer Learning-Sentence Level) is the most successful model when accuracy and loss are balanced since it obtains the lowest validation loss (0.0093) and the highest validation accuracy (0.9976), indicating good generalization capabilities.

C. Comparison with Previous Work

The models demonstrate different strengths in recall, F1-Score, and overall efficacy for Setswana NER, according to the performance comparison with baseline models that are already in the literature, as indicated in Table V, performance comparison with other existing work in Setswana NER. With an F1-Score of 78.06%, the Conditional Random Fields (CRF) baseline performs well overall, making it a solid benchmark. The BiLSTM-aux models have somewhat lower F1-Scores; their moderate efficacy is shown by their 74.07% BiLSTM-aux emb score. The CNN NER model in [7] shows inconsistent predictions, performing poorly in F1-score (62%) but well in recall (94%). In contrast, the CNN-BiLSTM model achieves a competitive F1-score (71%) while also improving recall (96%) [7]. But the CNN-BiLSTM transfer learning sentence-level model that was suggested works better than the others, with the highest recall (99%) and a high F1-Score (70%). The CNN-BiLSTM transfer learning sentence-level model is therefore suggested as the top-performing model for Setswana NER because of its competitive overall performance and superior recall.

TABLE V. PERFORMANCE COMPARISON WITH OTHER EXISTING WORK IN SETSWANA NER

Comparison Analysis				
Model	***Dataset***	***Recall***	***F1-Score***	
Conditional random fields (CRF) baseline NER for Setswana (Loubser, M. and Puttkammer, M.J., 2020)	NCHLT Setswana Named Entity Annotated Corpus	80.86%	75.47%	78.06%
bilstm-aux (Loubser, M. and Puttkammer, M.J., 2020)	“	74.14%	77.42%	75.74%
bilstm-aux emb (Loubser, M. and Puttkammer, M.J., 2020)	“	73.45%	74.71%	74.07%
CNN NER Model for Setswana NER (Chabalala, S., Owolawi, P. and Ojo, S., 2023)	“	77.00%	62.00%	94.00%
CNN-BiLSTM NER Model for Setswana (Chabalala, S., Owolawi, P. and Ojo, S., 2024)	"	83.00%	71.00%	96.00%
Our CNN-BiLSTM transfer learning Sentence Level	"	85.00%	70.00%	99.00%

VI. Conclusion

For the Setswana Language NER, the study suggested five models. The first two models, CNN and CNN-BiLSTM Transfer learning, are all based on word terms followed by the final three models, CNN, CNN-BiLSTM Transfer learning, and CNN-BiLSTM hybrid, which are all based on sentence-level. The evaluation was conducted on the South African Centre for Digital Language Resources (SADiLaR) NER dataset, and the top-performing model was ultimately compared to the state-of-the-art Setswana NER models that had already been published in the literature and were created using the same dataset. The top-performing model is Model 4 (CNN-BiLSTM Transfer Learning-Sentence Level), which attains the best macro averages, outstanding weighted averages, and high accuracy 99%. All classes, including minority ones, are balanced in terms of generality. As a result, sentence-level models perform better than word-level models since they can gather more contextual data, which enhances their efficiency. This model outperforms the state-of-the-art models in terms of accuracy (99%) and recall (85%), indicating its high precision and capacity to accurately identify most entities. Despite having a little lower macro average F1-Score (70%) than other models, its total performance, especially the accuracy and recall combination, makes it the best.

In addition to performing better than the other four models in this work, including models from literature, the suggested model has certain drawbacks because of its short dataset size, since deep learning algorithms often require huge datasets. Even though this work attempted to address this constraint through transfer learning, it was insufficient to completely address it; therefore, to further address this limitation, we recommend future research on the two domains indicated below as well as on the other fields.

Investigating cross-lingual transfer learning techniques may be valuable given the limitations of the current dataset and the generally scarce resources for the Setswana language. This study focused exclusively on NER in Setswana, with an emphasis on the South African Centre for Digital Language Resources (SaDilar) NER dataset, allowing for comparison with previous research conducted on the same dataset. Therefore, using tagged data from resource-rich languages could significantly improve the model's functionality in Setswana NER.

Error analysis: To identify any trends or problems the model shares, an error analysis can be conducted to inform future enhancements.

Consequently, it may be beneficial for future study to apply transformers, multilingual models like BERT, RoBERTa, or XLM-R, include larger and more varied datasets, and further optimize hyperparameters.

Dataset bias: The study has demonstrated that the dataset contains biases. Future studies should examine rebalanced datasets by adding more diverse named entities to alleviate these biases, particularly in underrepresented categories such as B-LOC, B-PERS, and B-ORG. Adding more foreign locales and English-Setswana mixed items to the dataset would also increase its robustness.

Acknowledgment

I would want to thank my family from the bottom of my heart for their unwavering support during this work.

References

[1] W. G. Bennett, M. Diemer, J. Kerford, T. Probert, and T. Wesi, ‘Setswana (South African)’, Journal of the International Phonetic Association, vol. 46, no. 2, pp. 235–246, 2016.

[2] M. Loubser and M. J. Puttkammer, ‘Viability of neural networks for core technologies for resource-scarce languages’, Information, vol. 11, no. 1, p. 41, 2020

[3] R. Hanslo, "Deep Learning Transformer Architecture for Named-Entity Recognition on Low-Resourced Languages: State of the art results," in 2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS), Sofia, Bulgaria, 2022.

[4] R. Eiselen, ‘Government domain named entity recognition for South African languages’, in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), 2016, pp. 3344–3348.

[5] S. Sukardi, M. Susanty, A. Irawan, and R. F. Putra, ‘Low complexity named-entity recognition for Indonesian language using BiLSTM-CNNs’, in 2020 3rd International Conference on Information and Communications Technology (ICOIACT), 2020, pp. 137–142.

[6] B. Okgetheng and G. Malema, ‘Named Entity Recognition for Setswana Language: A conditional Random Fields (CRF) Approach’, in Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval, 2023, pp. 240–244.

[7] S. Chabalala, P. Owolawi, and S. Ojo, ‘A Convolutional Neural Network Model for Setswana Named Entity Recognition’, in International Conference on Artificial Intelligence and its Applications, 2023, pp. 165–171.

[8] Y. Jia and X. Xu, ‘Chinese named entity recognition based on CNN-BiLSTM-CRF’, in 2018 IEEE 9th international conference on software engineering and service science (ICSESS), 2018, pp. 1–4.

[9] N. C. Lê, N.-Y. Nguyen, A.-D. Trinh, and H. Vu, ‘On the Vietnamese name entity recognition: A deep learning method approach’, in 2020 RIVF International Conference on Computing and Communication Technologies (RIVF), 2020, pp. 1–5.

[10] S. J. Pan and Q. Yang, ‘A survey on transfer learning’, IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345–1359, 2009.

[11] Niu, Y. Liu, J. Wang, and H. Song, ‘A decade survey of transfer learning (2010--2020)’, IEEE Transactions on Artificial Intelligence, vol. 1, no. 2, pp. 151–166, 2020.

[12] Lee, J.Y., Dernoncourt, F. and Szolovits, P., 2017. Transfer learning for named-entity recognition with neural networks. arXiv preprint arXiv:1705.06273.

[13] Shan, L., Liu, Y., Tang, M., Yang, M. and Bai, X., 2021. CNN-BiLSTM hybrid neural networks with attention mechanism for well log prediction. Journal of Petroleum Science and Engineering, 205, p.108838.

[14] Zhang, Z. Chen, D. Liu, and Q. Lv, ‘Building Structured Patient Follow-up Records from Chinese Medical Records via Deep Learning’, in 2022 2nd International Conference on Bioinformatics and Intelligent Computing, 2022, pp. 65–71.

[15] J. Brownlee, Deep learning with Python: develop deep learning models on Theano and TensorFlow using Keras. Machine Learning Mastery, 2016.

[16] I. Keraghel, S. Morbieu, and M. Nadif, ‘A survey on recent advances in named entity recognition’, arXiv preprint arXiv:2401. 10825, 2024.

[17] B. Okgetheng and G. Malema, ‘Named Entity Recognition for Setswana Language: A conditional Random Fields (CRF) Approach’, in Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval, 2023, pp. 240–244.

[18] S. A. Stats, ‘Community survey 2016 in brief’, Statistics South Africa. Pretoria: SSA, 2016.

[19] R. Letsholo and K. Matlhaku, ‘The syntax of the Setswana noun phrase’, Marang: Journal of Language and Literature, vol. 24, pp. 22–42, 2014.

[20] J. van de Kerkhof, "Convolutional neural networks for named entity recognition in images of documents," Stockholm Sweden, 2016

[21] A. Goyal, V. Gupta, and M. Kumar, 'Recent named entity recognition and classification techniques: a systematic review', Computer Science Review, vol. 29, pp. 21–43, 2018.

[22] V. Harrison and M. Walker, 'Neural generation of diverse questions using answer focus, contextual and linguistic features', arXiv preprint arXiv:1809. 02637, 2018.

[23] T. T. H. Hanh, A. Doucet, N. Sidere, J. G. Moreno, and S. Pollak, 'Named entity recognition architecture combining contextual and global features', in International Conference on Asian Digital Libraries, 2021, pp. 264–276.

Planning and Design of Elderly Care Space Combining PER and Dueling DQN

Di Wang, Hui Ma*, Yu Chen
College of Art and Design, Jilin Jianzhu University, Changchun, 130118, China

***Abstract*—With the continuous development of the aging phenomenon in society, people's attention to the planning of elderly care spaces is increasing. Currently, many scholars have used various spatial planning models to plan and design elderly care spaces. However, the resource utilization rate and comfort of the elderly care spaces designed by these models are low, and the models still need to be optimized. This study first integrates the Prioritized Experience Replay mechanism with the Dueling deep Q-network algorithm, and constructs a spatial planning model based on the fused algorithm, to use this model to plan elderly care spaces reasonably. The study first conducts comparative experiments on the fusion algorithm, and the outcomes indicate that the fusion algorithm has the best prediction performance, with a minimum prediction error rate of only 0.9% and a prediction speed of up to 8.7bps. In addition, the denoising effect of the algorithm is the best, and the performance of the algorithm is much higher than that of the comparative algorithm. Further analysis of the spatial planning model based on this algorithm shows that the average time required for elderly care space planning is only 1.3 seconds, and the comfort level of the planned elderly care space reaches 98.7%, the resource utilization rate reaches 89.7%, and the planned elderly care space can raise the living standard of the elderly by 67.7%. From the above information, the spatial planning model raised in the study can validly enhance the resource utilization and comfort of elderly care spaces, and raise the living standard of the elderly.**

Keywords*—*Elderly care space; planning and design; prioritized experience replay; dueling deep q-network algorithm; spatial planning

I. Introduction

With the continuous aggravation of the aging problem in society, the attention to the rationalization of elderly care space planning and design is constantly increasing [1]. Reasonable planning and design of elderly care spaces can create a comfortable, safe, and natural living environment for the elderly, improving their quality of life. In addition, it can also provide sufficient outdoor activity space for the elderly, promoting physical and mental health [2]. With the sustained growth of computer technology, numerous scholars have conducted research on spatial planning and design methods using intelligent algorithms and physical models [3-4]. However, these spatial planning and design methods still have problems such as slow spatial planning speed and low utilization of spatial resources, and optimization of planning models is needed. The Dueling Deep Q-Network (Dueling DQN) algorithm can improve its learning efficiency through Q-values, thereby accelerating spatial planning speed. The Prioritized Experience Replay (PER) mechanism can accelerate the convergence speed of the algorithm through its priority sampling method. In order to improve the spatial planning speed of the elderly care space planning model, enhance the resource utilization rate in the elderly care space, and ensure the quality of life and mental health of the elderly, this study organically combines the PER mechanism with the Dueling DQN algorithm, and constructs an elderly care space planning and design model based on the combined PER-Dueling DQN algorithm. The innovation of this study lies in optimizing the uniform sampling in the Dueling DQN algorithm through the priority sampling technique in the PER mechanism, in order to reduce the adverse effects of uniform sampling on the calculation accuracy of the Dueling DQN algorithm. The contribution of the research lies in the fact that through reasonable planning of elderly care spaces, it can improve the quality of life of the elderly, promote equal social welfare and public services, drive the development of the elderly care industry, and better cope with population aging.

II. Related Work

With the rapid development of computer technology, many scholars have conducted research on spatial planning and design methods using intelligent algorithms and mathematical models. For example, Gu Y et al. raised a spatial planning model grounded on the least squares parameter estimation method to plan and design the state space of the simulated product portfolio. The model was tested in practical situations, and the outcomes indicated that the spatial planning time of the model was only 3.2 seconds [5]. SS VC et al. developed a space design model grounded on metaheuristic algorithm to address the problem of high memory resource utilization in current engineering space design models. The model was compared with traditional space design models, and the results showed that the model could reduce resource utilization by 23.1% [6]. In addition, to deal with the issue of high computational complexity in diffusion grounded spatial model planning, Karras T et al. raised a spatial design model grounded on fractional network preprocessing method, which was used for detection in practical situations. The outcomes indicated that the computational complexity of this model was reduced by 12.6% compared to traditional spatial planning models [7]. However, there are still problems with slow spatial planning speed and low utilization of spatial resources in the above spatial planning models, and the model still needs to be optimized [8].

Dueling DQN algorithm is an algorithm that can improve computational efficacy and steadiness by separating state value and dominance function [9]. This algorithm is commonly applied in various models due to its ability to validly reduce the quantity of parameters and high applicability. For example, Chi

J et al. designed a deep reinforcement training model grounded on the Dueling DQN algorithm to solve the matters of slow convergence speed and low training accuracy in the deep reinforcement training process. The model was compared with traditional models, and the outcomes indicated that the convergence speed of the model was improved by 19.5% [10]. However, there are still problems with poor learning efficiency and slow convergence speed caused by the uniform sampling mechanism in this algorithm [11]. PER mechanism can improve the convergence speed and learning efficiency of algorithms by sampling based on priority [12]. This mechanism is also commonly utilized in various areas to raise the convergence speed and computational precision of models [13].

To sum up, although many researchers have carried out studies on spatial planning models, these models still suffer from slow planning speed and low utilization of spatial resources, and further improvements are needed. Therefore, this study organically combines the PER mechanism with the Dueling DQN algorithm, and constructs a retirement space planning and design model based on the combined PER Dueling DQN algorithm, to use this model to plan retirement spaces reasonably and guarantee the living standard and mental health of the elderly.

III. Methods and Materials

A. Dueling DQN Algorithm Based on Per Optimization

Elderly care space refers to a social space that provides safe, comfortable, and convenient living and living environments for the elderly [14]. The main purpose of this space is to meet the material and spiritual needs of the elderly and help them maintain physical health [15]. With the continuous development of the aging phenomenon, people's attention to elderly care space is constantly increasing. Thoughtful design and strategic planning of spaces for elderly care can guarantee financial stability for individuals as they age [16]. However, many spatial planning models currently suffer from insufficient coordination and low utilization of spatial planning resources, and further optimization of the models is needed. The Dueling DQN algorithm is an improved Q-learning algorithm that can improve learning efficiency by decomposing Q-values and effectively handle complex spatial states and decision problems [17]. The basic process of this algorithm is shown in Fig. 1.

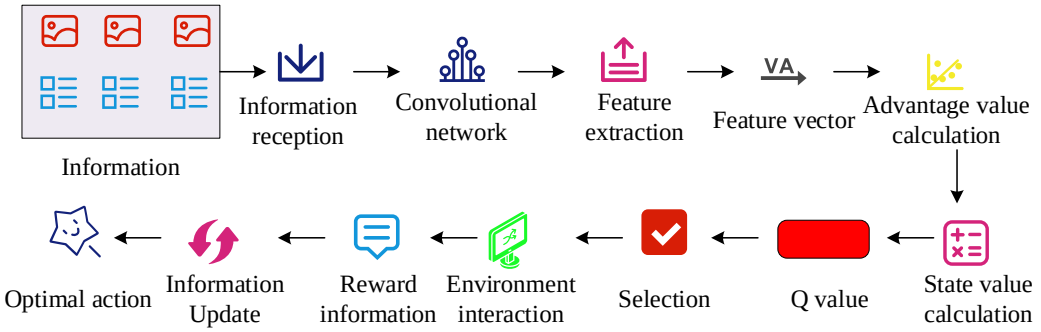


Fig. 1. Basic process of dueling DQN algorithm.

In Fig. 1, the algorithm first needs to receive input state information, which includes multiple images or chart data. Then, the convolutional network is applied to identify the features of the state information and generate feature vectors. Then, the evaluation network and target network in the dual network are used to calculate the value of state information and the advantage value of each action. The calculated state value and action advantage value are added to obtain the Q value of each action. Using the computed Q value, the best course of action is chosen, typically opting for the action with the highest Q value for implementation. After determining the execution action, it engages with its surroundings to acquire rewards and updated state data. The neural network's parameters undergo modification grounded on the obtained new state information and reward mechanism, so that the algorithm can find the optimal action execution. The calculation method of Q value is shown in Eq. (1).

$$Q(S,A,w,\alpha,\beta) = V(S,w,\alpha) + \left(A(S,A,w,\beta) - \frac{1}{|A|}\sum_{a\in A} A(S,\alpha,w,\beta)\right) \quad (1)$$

In Eq. (1), S means the information state, A means the action, w means the network parameters of the common part, α means the network parameters unique to the value function, β means the network parameters unique to the advantage function, $Q(S,A,w,\alpha,\beta)$ means the calculated Q value, $V(S,w,\alpha)$ means the value function part related to the state function, and $A(S,A,w,\beta)$ means the advantage function related to both the state and action. The calculation method of the state value function is shown in Eq. (2).

$$V(s) = f(s;\theta) \quad (2)$$

In Eq. (2), $V(s)$ means the action function in state s, f means the neural network used to calculate the state value, and θ means the parameters of the state value function used to optimize the neural network. The calculation formula for the action advantage function is shown in Eq. (3).

$$A(s,a) = Q(s,a) - V(s) \quad (3)$$

In Eq. (3), $Q(s,a)$ means the Q value of taking action a in state s, and V means the value function. When performing feature extraction operations in the convolutional layer, it is necessary to calculate the computational and parameter quantities of the convolutional layer in order to

optimize the network structure and design a reasonable convolutional layer structure and parameters. The calculation method for the parameter quantity of the convolutional layer is shown in Eq. (4).

$$N = k \times k \times C_{in} \times C_{out} \tag{4}$$

In Eq. (4), N means the number of parameters in the convolutional layer, k means the size of the convolutional kernel, C_{in} means the number of input channels, and C_{out} means the number of output channels. The calculation method for the computational complexity of the convolutional layer is shown in Eq. (5).

$$N' = W_{out} \times H_{out} \times k \times k \times C_{in} \times C_{out} \tag{5}$$

In Eq. (5), N' means the computational complexity, and W_{out} and H_{out} mean the width and height of the output feature map. The optimal action can be found through the above calculation. However, due to the uniform sampling mechanism, the Dueling DQN algorithm may reduce its learning efficiency and effectiveness, and further optimization of the algorithm is needed. The PER mechanism is a commonly-used technique in reinforcement learning, which can improve the computational speed and learning efficiency of algorithms by prioritizing sampling the most valuable experiences for learning [18]. The basic structure of PER mechanism is in Fig. 2.

As shown in Fig. 2, the PER mechanism consists of five parts: experience pool, time difference error calculation, sampling probability calculation, sampling process, and policy update. The experience pool is a data storage pool that contains the experience data obtained by the algorithm during each interaction with the environment during reinforcement learning, including the current state, actions taken, rewards obtained, and next state. The calculation of time difference error refers to the calculation of the difference between the current Q value and the target Q value. The larger the Q value and temporal error of the experience, the higher the importance of the experience in algorithm learning. The sampling probability of each experience is calculated again, the sampling probability of each experience is calculated, and the experience with high sampling probability is prioritized for training by PER, thereby reducing unnecessary learning time and accelerating algorithm convergence to get the best solution. Finally, the Q value parameter or strategy parameter is adjusted based on the calculated sampling probability and time difference error to improve the algorithm's computation speed and learning efficiency. The calculation method of time difference error in this algorithm is shown in Eq. (6).

$$TD = Q(s,a) - (r + \gamma * V(s') - V(s)) \tag{6}$$

In Eq. (6), r means the immediate reward obtained from executing action a from state s, γ means the discount factor, and $V(s')$ and $V(s)$ mean the value estimation functions of the current state and the next state. The calculation method of sampling probability is shown in Eq. (7).

$$E[f(x)] = 1/m \sum_{i=1}^{m} f(xi) \tag{7}$$

In Eq. (7), m means the total number of sampling times, xi means the sampled samples, and $f(xi)$ means the function value of the samples. To improve the drawbacks of slow computation speed and low learning efficiency caused by uniform sampling methods in the Dueling DQN algorithm, this study utilizes the PER mechanism to optimize the Dueling DQN algorithm. The basic process of the optimized PER-Dueling DQN algorithm is in Fig. 3.

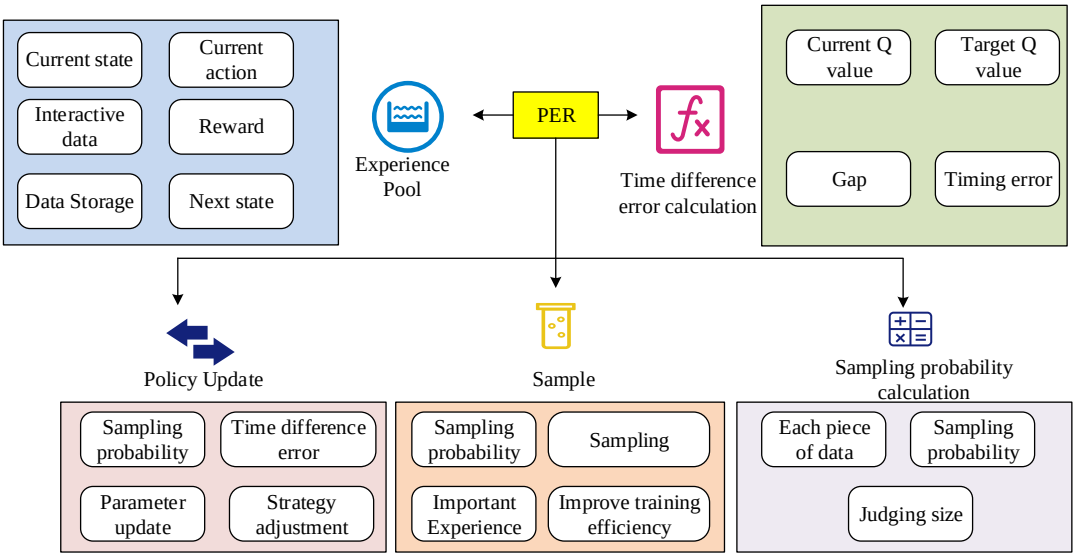


Fig. 2. Basic structure diagram of PER mechanism.

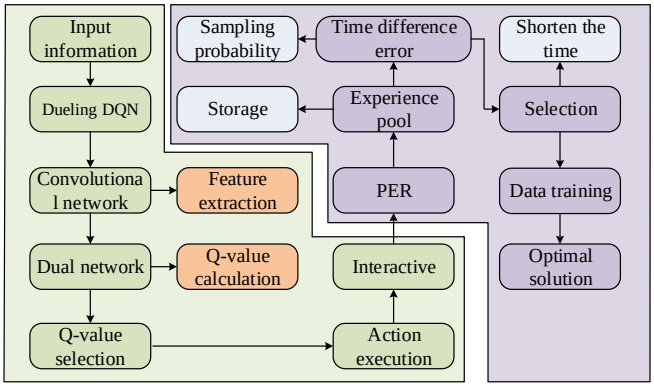


Fig. 3. Basic process of PER-dueling DQN algorithm.

As shown in Fig. 3, after receiving various input information, the PER-Dueling DQN algorithm first uses the convolutional network in the Dueling DQN algorithm to extract features, and then uses the dual network in the algorithm to calculate the Q value of each piece of information. The action with the highest Q value is chosen for execution. After the action is executed, when the algorithm interacts with the environment, the experience pool in the PER mechanism is used to store the experience obtained after each action interaction. The time discrepancy error and sampling likelihood of each action are computed, and the empirical data with high sampling likelihood and minimal temporal error are chosen for training. This approach minimizes unnecessary training duration and enhances both the computational speed and training efficiency of the algorithm.

B. *Retirement Space Planning Model Based on PER-Dueling DQN Algorithm*

In response to the problems of poor spatial planning effectiveness and long planning time in current elderly care space planning models, this study uses the PER-Dueling DQN algorithm proposed in the previous section to optimize the spatial planning model, aiming to raise the planning efficiency of the model and reduce the planning time of the model. The basic framework of the elderly care space planning model grounded on PER-Dueling DQN algorithm is shown in Fig. 4.

From Fig. 4, the model first needs to clarify the planning goals and objectives, namely the quality of life, medical security, and leisure activities for individuals after retirement. Secondly, it is necessary to evaluate the existing assets, income, and expenses of individuals or families to determine retirement funds. Then, based on the determined retirement funds, a suitable pension plan should be developed, including pension investment portfolio, pension insurance, and retirement pension plan. Then, the pension plan is followed, and corresponding adjustments are made according to the actual situation. Finally, spatial planning and design are carried out. The planning and design of elderly care spaces first need to collect relevant information about space use, management, and functions. Then, the collected information and planning objectives are input into the PER-Dueling DQN algorithm to find the optimal planning path, that is, the optimal spatial planning and design scheme. After finding the optimal design scheme, it is necessary to verify the scheme to determine its practical feasibility. Finally, the plan is recorded and a planning and design report is generated, which is submitted to relevant personnel for review and evaluation. If the evaluation is qualified, the elderly care space planning and design will be carried out according to the plan. If the evaluation is not qualified, the plan selection will be repeated until it is qualified. In spatial planning and design, it is necessary to calculate the volume of various objects in space, as well as the strength of objects and structures in space, in order to design the structures in space. The calculation method for the volume of various objects is shown in Eq. (8).

$$\begin{cases} V1 = c^3 \\ V2 = \pi r^2 h \\ V3 = 3/4\pi r^3 \end{cases} \tag{8}$$

In Eq. (8), $V1$ means the volume of the cube, c means the side length of the cube, $V2$ means the volume of the cylinder, r means the radius, h means the height, and $V3$ means the volume of the sphere. The strength calculation method for beam objects in space is shown in Eq. (9).

$$\begin{cases} \sigma = M\max / Wz \\ \tau = V''S'' / Wit \end{cases} \tag{9}$$

In Eq. (9), σ means the maximum stress of the beam material, $M\max$ means the maximum bending moment, $V(s')$ and Wz mean the section modulus, τ means the material shear stress, V'' means the material shear force, S'' means the material gross section area, and Wit means the gross section moment of inertia. The calculation method for the strength of a pole in space is shown in Eq. (10).

$$\begin{cases} F1 = Pq / \varepsilon \\ F2 = \delta 1 / \delta 2 \end{cases} \tag{10}$$

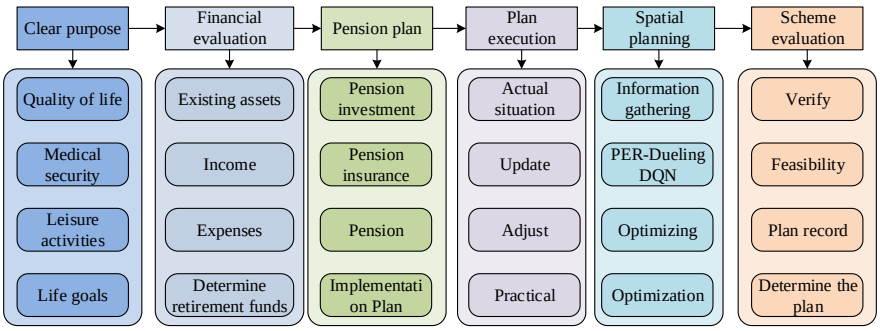


Fig. 4. Retirement space planning model based on PER-Dueling DQN algorithm.

In Eq. (10), $F1$ means the tensile strength of the rod, Pq means the yield strength of the rod, ε is the safety factor, $F2$ is the bending strength, δ is the bending moment, and $\delta 1$ means the bending moment resistance. In the planning of elderly care space, compound interest terminal value can help planners better plan elderly care funds. By selecting the initial investment amount and investment time reasonably, it ensures that the funding needs of the elderly care space are met. The calculation formula is shown in Eq. (11).

$$Fv = Pv \times (1+r)^n \tag{11}$$

In Eq. (11), Fv means the terminal value, Pv means the present value, r means the interest rate, and n is a term number. The application principle of PER-Dueling DQN algorithm in elderly care space planning is shown in Fig. 5.

In Fig. 5, in the planning and design of elderly care space, this algorithm first needs to receive the spatial data collected in the early stage, as well as the spatial planning purpose and requirements, and preprocess these data to denoise the data information. Then the eval network and target network in the PER-Dueling DQN algorithm, namely the evaluation network and target network, are initialized. Spatial information is input into the eval network to calculate the Q value of each action, that is, the Q value of each planning and design scheme. The action with the highest Q value is chosen for execution, and the current action, state, and the next state after completing the action are stored in the PER's experience pool. Once all actions have been executed and states recorded, the temporal difference error for each piece of data in the experience replay buffer is computed to establish its priority. The data with larger errors has higher priority. A batch of data is selected by priority from the experience pool and input into the eval network, where the current Q value is calculated. At the same time, these data are input into the target network to calculate the target Q value. The mean square deviation between the current Q value and the target Q value is used as the loss function. Based on this loss function, the parameters of the eval network are backpropagated and updated, and the parameters of the target network are synchronously updated. The above steps are repeated until the preset training coefficients are reached or the optimal solution is found, that is, the spatial optimal planning and design scheme. The optimal solution is output and recorded. The calculation method for the target Q value is shown in Eq. (12).

$$Y = \chi + \gamma Qt(s', \max a')Qe(s', a', \theta', \theta'') \tag{12}$$

In Eq. (12), χ represents the immediate reward, Qt is the target network, Qe represents the estimation network, θ' represents the parameters of the estimation network, and θ'' represents the parameters of the target network. Through the above process, the optimal planning and design scheme for the elderly space is obtained, in order to meet the material, spiritual, and social needs of the elderly, and help them maintain physical health, independent living, and mental vitality.

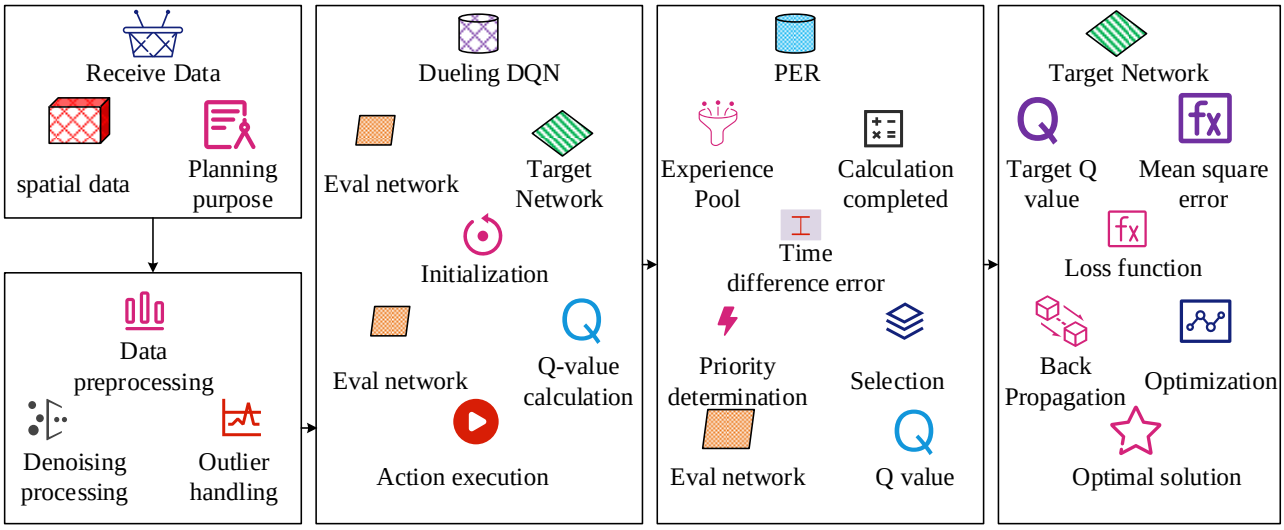


Fig. 5. Application principle of PER-dueling DQN algorithm.

IV. Results

A. Performance Analysis of PER-Dueling DQN Algorithm

In order to analyze the superiority of the prediction performance of the PER-Dueling DQN algorithm, this study conducted comparative experiments between the PER-Dueling DQN algorithm and the Genetic Algorithm-Back Propagation (GA-BP) algorithm, the Particle Swarm Optimization-Support Vector Machine (PSO-SVM) algorithm, and the Dueling DQN algorithm before optimization using the PER mechanism. The environment configuration during the experiment is in Table I.

TABLE I Experimental Environment Configuration

Experimental environment	Index	Style
Hardware environment	CPU	Intel Core i9
	EMS memory	64GB
Software environment	OS	Windows 10
	Python edition	Python 4.0
	Python environment	Anaconda 3

During the experiment, the ImageNet dataset was chosen as the experimental dataset, which contains 22000 categories of image data. The predictive performance of the four algorithms was analyzed through the above experimental environment configuration and experimental dataset. Firstly, the validity of the dataset was verified using k-fold cross validation. The dataset in ImageNet was evenly divided into (a, b, c, d, e) 5 parts, and four parts were used for testing. The remaining dataset was used for validation. The selection of the dataset was validated by testing each part of the dataset. The results are shown in Table II.

According to Table II, each dataset in the ImageNet dataset had a relatively low impact on the PER-Dueling DQN algorithm, indicating that the selection of this dataset was reasonable. The prediction performance of the four algorithms was compared, and the results are shown in Fig. 6.

TABLE II RESULTS OF RATIONAL TESTING OF DATA SETS

Training dataset	Test dataset	PER-Dueling DQN prediction accuracy	PER-Dueling DQN denoising accuracy
a,b,c,d	e	98.7%	93.2%
a,b,c,e	d	97.9%	92.8%
a,b,d,e	c	97.9%	92.7%
a,c,d,e	b	98.1%	93.1%
b,c,d,e	a	98.6%	93.5%

The red dashed line in Fig. 6 represents the areas where the predicted values differ from the actual values. From Fig. 6, among the four algorithms, only the PER-Dueling DQN algorithm had similar forecasted values to the true values. However, GA-BP algorithm, PSO-SVM algorithm, and Dueling DQN algorithm had different degrees of error between their forecasted values and true values after predicting the data, with Dueling DQN algorithm having the largest error between the forecasted values and true values. By comparing the prediction errors and speeds among the four algorithms, the results are presented in Fig. 7.

According to Fig. 7 (a), among the four algorithms, the PER-Dueling DQN algorithm had the lowest prediction error, only 0.9%, while the GA-BP algorithm, PSO-SVM algorithm, and Dueling DQN algorithm had prediction errors of 1.6%, 2.1%, and 3.2%, respectively. The prediction errors of the latter three algorithms were much higher than those of the PER-Dueling DQN algorithm. As shown in Fig. 7 (b), the average prediction speeds of PER-Dueling DQN algorithm, GA-BP algorithm, PSO-SVM algorithm, and Dueling DQN algorithm for data were 8.7 bps, 5.7 bps, 4.8 bps, and 2.1 bps, respectively. PER-Dueling DQN algorithm had the fastest prediction speed, while the unoptimized Dueling DQN algorithm had the slowest prediction speed, with a difference of 6.6 bps between the two. Finally, the denoising effects of the four algorithms were compared, and the results are shown in Fig. 8.

In Fig. 8, the PER-Dueling DQN algorithm had the best denoising effect among the four algorithms. After using this algorithm for denoising, the noise information in the data could be almost completely removed. However, after denoising the data, GA-BP algorithm, PSO-SVM algorithm, and Dueling DQN algorithm still contained a large amount of noise information. The denoising performance of the last three algorithms was inadequate, potentially leading to inaccurate predictions when forecasting future data. From the above experimental results, the PER-Dueling DQN algorithm raised in this study had the best denoising effect, the shortest prediction time, the lowest prediction error, and the best prediction effect. Therefore, this study uses the PER-Dueling DQN algorithm to construct a spatial planning model, in order to predict the various performance of elderly space planning through the excellent prediction effect of the PER-Dueling DQN algorithm. By adjusting the elderly space planning scheme based on the prediction results, the comfort and resource utilization of the elderly space can be improved.

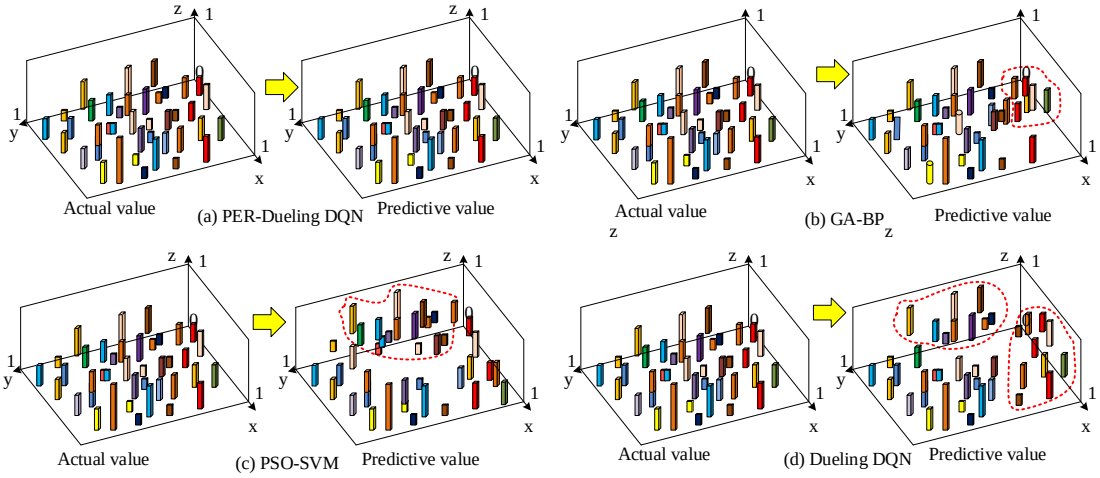
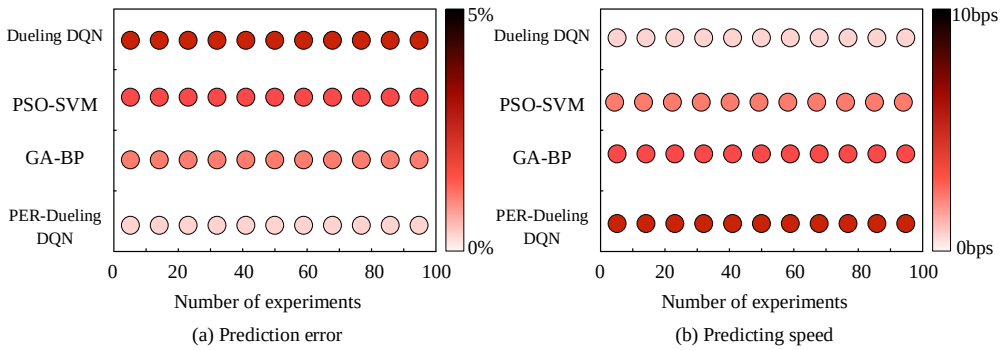
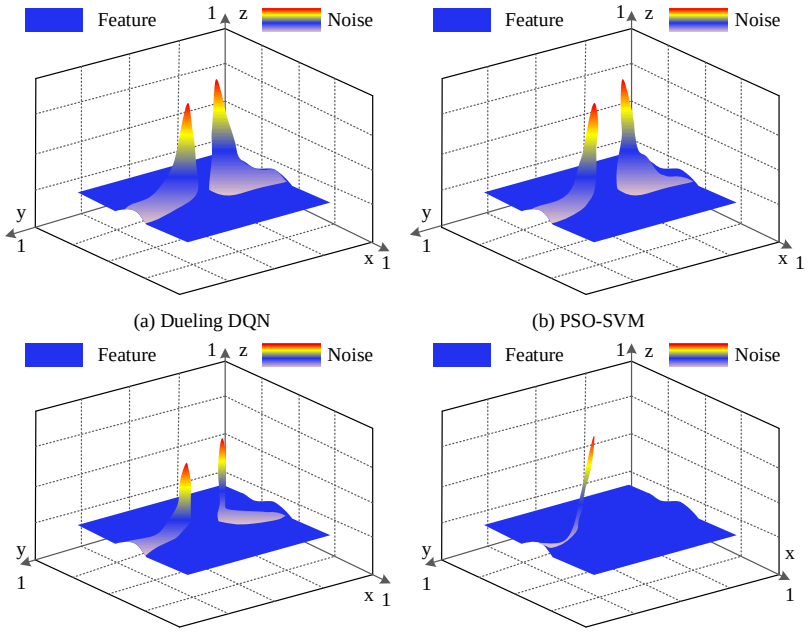


Fig. 6. Comparison of algorithm prediction performance.



(a) Prediction error (b) Predicting speed

Fig. 7. Comparison of algorithm prediction error and prediction speed.



(a) Dueling DQN (b) PSO-SVM

(c) GA-BP (d) PER-Dueling DQN

Fig. 8. Comparison of denoising effects of algorithms.

B. *Empirical Analysis of PER-Dueling DQN Planning Model*

After verifying the superiority of the prediction performance of the PER-Dueling DQN algorithm, the planning effect of the spatial planning model based on this algorithm was validated. The PER-Dueling DQN model was compared with commonly-used spatial planning models based on Convolutional Neural Networks Simulated Annealing (CNN-SA), Ant Colony Optimization Tabu Search (ACO-TA), and Discrete Grey Wolf Optimization (DGWO). During the experiment, a planning scheme for elderly care space in a certain family was selected as the source of the experimental dataset, which includes information on the size, scale, funding sources, and elderly care goals and needs of the elderly care space. The elderly space planning scheme was optimized using four models, the performance of the elderly space design scheme optimized by the four models was compared, and the advantage of the raised model was verified. Firstly, the planning time and spatial comfort of the four planning models were compared, and the results are shown in Fig. 9.

From Fig. 9 (a), the PER-Dueling DQN spatial planning model took an average of 1.3 seconds to plan the elderly care space, and the comfort level of the elderly care space designed by this model reached 98.7%. The spatial planning time of this model was short, and the planned space had a high comfort level, which was conducive to the living of the elderly. As shown in Fig. 9 (b), the average time taken by the CNN-SA spatial planning model for spatial planning reached 2.2 seconds, which was higher than that of the PER-Dueling DQN model. Moreover, the comfort level of the elderly care space planned by the CNN-SA model was 85.6%. From Fig. 9 (c) and Fig. 9 (d), the average time spent on spatial planning by the ACO-TA and DGWO spatial planning models was much higher than that of the PER-Dueling DQN model. The time spent by the ACO-TA and DGWO models was 2.9 seconds and 3.7 seconds, respectively. The comfort levels of the elderly care space planned by the two models were 78.9% and 68.6%, respectively. Comparing the safety and spatial resource utilization efficiency of the elderly care spaces planned by the four models, the results are shown in Fig. 10.

According to Fig. 10 (a), the PER-Dueling DQN spatial planning model had the highest safety of the elderly care space, reaching 98.3%, which can effectively ensure the safety of the elderly. The safety of the elderly care space designed by the CNN-SA model, ACO-TA model, and DGWO model was 90.2%, 82.6%, and 75.8%, respectively. The safety of the elderly care space planned by the latter three models was much lower than that of the PER-Dueling DQN model. From Fig. 10 (b), among the four models, the PER-Dueling DQN model had the highest utilization rate of spatial resources, reaching 89.7%, while the DGWO model had the lowest utilization rate of spatial resources, only 65.4%. Finally, the impact of elderly care spaces on the elderly was compared, and the data are in Table III.

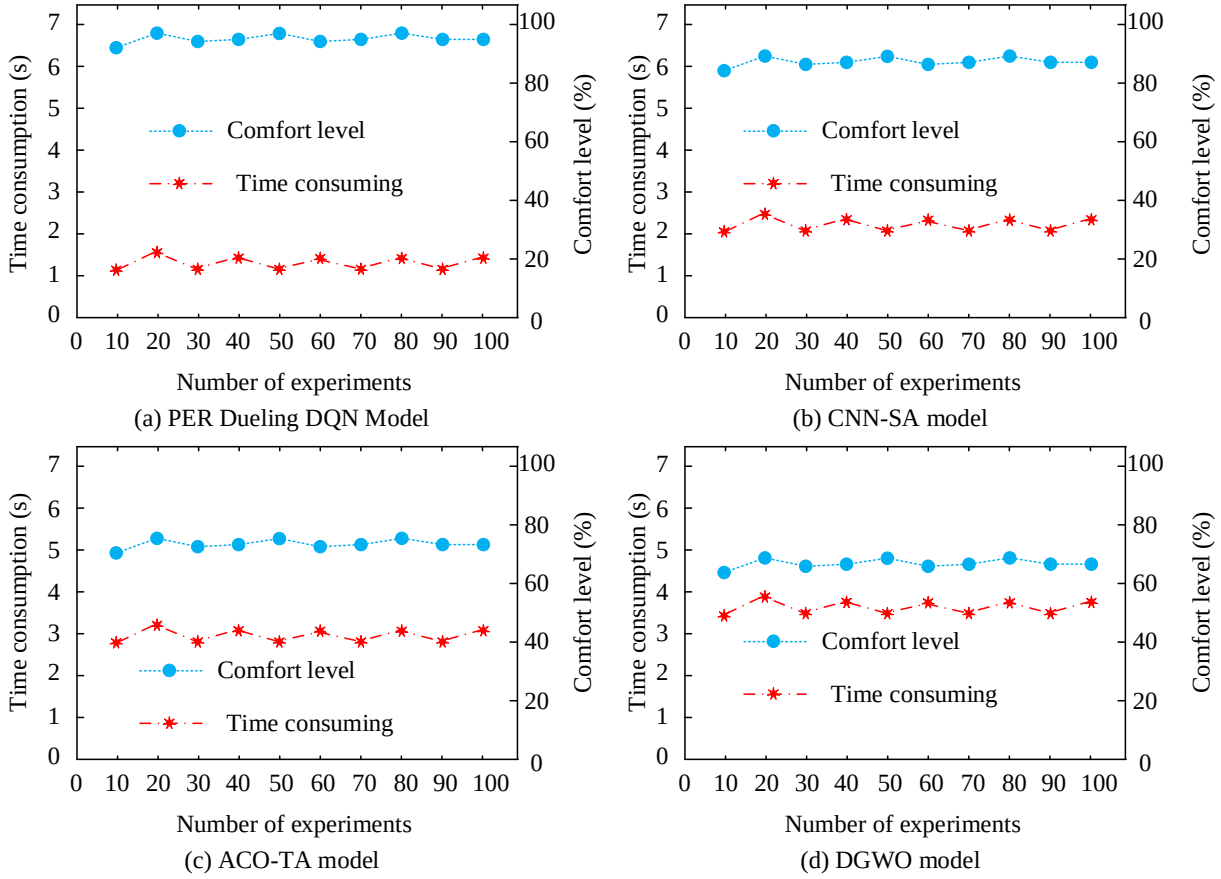


Fig. 9. Model planning time and spatial comfort.

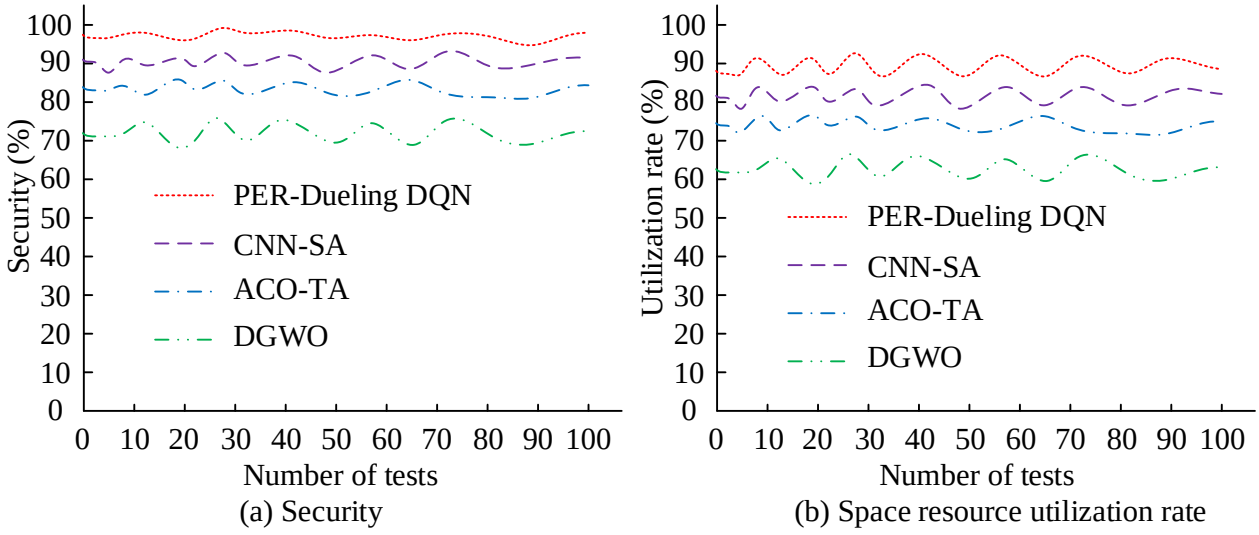


Fig. 10. Comparison of safety and resource utilization in elderly care spaces.

TABLE III ANALYSIS OF THE EFFECTS OF FOUR MODELS ON THE ELDERLY

Model		Quality of life	Mental health	Convenient living	Sense of happiness	Social skills
PER-Dueling DQN	Increase percentage	67.7%	82.4%	88.7%	90.3%	96.4%
	Meet expectations	Y	Y	Y	Y	Y
CNN-SA	Increase percentage	60.5%	74.7%	82.3%	85.8%	89.7%
	Meet expectations	Y	N	Y	N	N
ACO-TA	Increase percentage	50.3%	70.3%	78.7%	80.6%	82.7%
	Meet expectations	N	N	N	N	N
DGWO	Increase percentage	48.7%	68.5%	75.3%	73.8%	72.1%
	Meet expectations	N	N	N	N	N
Expected increase		>50.7%	>75.6%	>80.5%	>88.6%	>90.2%

According to Table III, after using four spatial planning models for elderly care space planning and design, only the PER-Dueling DQN spatial planning model could achieve the expected performance indicators of elderly care space for the elderly. After using the PER-Dueling DQN model, the planned elderly care space could raise the living standard of the elderly by 67.7%, mental health by 82.4%, convenience of life by 88.7%, happiness by 90.3%, and social skills by 96.4%, which was much higher than the expected improvement of 50.7%, 75.6%, 78.7%, 80.6%, and 82.7%. The elderly care space designed using the CNN-SA model could only raise the living standard and convenience of the elderly to the expected level, while other indicators were slightly lower than expected. Although the performance indicators of the elderly care space planned by the ACO-TA model and DGWO model improved, the improvement level of both models was far from meeting the expected requirements. Based on the above information, the PER-Dueling DQN spatial planning model proposed in this study can offer sensible planning and design for elderly care spaces, ensuring a high standard of living for the elderly and enhancing their physical health.

V. DISCUSSION

In response to the problems of poor spatial planning rationality and low spatial comfort in current elderly care space planning models, this study used the PER mechanism to optimize the Dueling DQN algorithm, proposed a PER-Dueling DQN algorithm, and constructed an elderly care space planning model based on this algorithm. To confirm the superiority of the model, comparative experiments were carried out on the optimization algorithm first. The study compared the PER-Dueling DQN algorithm with the GA-BP algorithm, PSO-SVM algorithm, and Dueling DQN algorithm. The results showed that among the four algorithms, the PER-Dueling DQN algorithm had the strongest denoising ability, and its prediction error was only 0.9%, far lower than the other three algorithms. The prediction speeds of PER-Dueling DQN algorithm, GA-BP algorithm, PSO-SVM algorithm, and Dueling DQN algorithm were 8.7 bps, 5.7 bps, 4.8 bps, and 2.1 bps, respectively. PER-Dueling DQN algorithm has the fastest prediction speed, which is similar to the experimental results of Bai Z team [19]. The reason for this result may be that the PER-Dueling DQN algorithm can improve the training speed and efficiency of the algorithm through the priority sampling method in the PER mechanism, thereby ensuring the denoising effect of the algorithm and improving its prediction accuracy. Further experimental analysis was conducted on the spatial planning models based on the PER-Dueling DQN algorithm, CNN-SA spatial planning model, ACO-TA spatial planning model, and DGWO spatial planning model. The results showed that among the four spatial planning models, the PER-Dueling DQN model had the shortest spatial planning time, only requiring 1.3 seconds, and the highest resource utilization rate of the elderly care space planned by the PER-Dueling DQN model reached 89.7%. The elderly care space planned by the PER-Dueling DQN model could raise the living standard of the elderly by 67.7%, improve their mental health by 82.4%, and increase their sense of happiness by 90.3%, which fully met the expected level of improvement. Although the other three models also improved, they could not all meet expectations, which is consistent with the results of Williams R A et al. [20]. The reason for the results may be that the algorithm in the PER-Dueling DQN model has excellent predictive ability, which can continuously optimize the space through the predicted results until the optimal result is reached, thereby improving the performance of the elderly care space.

VI. CONCLUSION

In order to solve the problem of low spatial planning rationality in current elderly care space planning, this study combined PER mechanism and Dueling DQN algorithm, and designed an elderly care space planning model based on the combined algorithm. The study first compared the PER-Dueling DQN algorithm with other related algorithms, and the results showed that the performance of the PER-Dueling DQN algorithm was superior to other algorithms. The elderly space planning model based on the PER-Dueling DQN algorithm could improve the quality of life, mental health status, and happiness of the elderly. From the above results, the suggested elderly care space planning model can effectively improve the comfort of the elderly care space, thereby ensuring the physical and mental health status and quality of life of the elderly. However, although the PER-Dueling DQN algorithm used in this study had excellent prediction performance, it contained many hyperparameters, making parameter tuning difficult. It had high computational complexity and required a large amount of resources. Moreover, in some complex or dynamically changing environments, the PER-Dueling DQN algorithm may

not be able to effectively adapt and achieve optimal algorithm performance. In the future, automated parameter tuning techniques such as Bayesian optimization algorithm and genetic algorithm can be utilized to intelligently select and optimize hyperparameters of algorithms, reducing the difficulty and cost of parameter tuning and lowering computational complexity. Moreover, it can be combined with other reinforcement learning algorithms to form a hybrid algorithm, further improving the performance and stability of the algorithm.

FUNDINGS

The research is supported by: Key Project of Science and Technology Research of Education Department of Jilin Province, Research on Service design of Urban Nursing Center based on Extension data Mining Technology, (No. JJKH20220277KJ). This is a key project of the Jilin Provincial Education Science '14th Five-Year Plan' for 2023, This Research on the Construction of an Innovative Practical Teaching System for Environmental Design Major through Multidisciplinary Integration under the Background of 'Mass Entrepreneurship and Innovation'. (No. ZD23009).

REFERENCES

[1] Bardaro G, Antonini A, Motta E. Robots for elderly care in the home: A landscape analysis and co-design toolkit. International Journal of Social Robotics, 2022, 14(3): 657-681.

[2] Rahmawati E A, Wahyunengseh R D, Mulyadi A W E. Evaluation of elderly-friendly open space and public service buildings in madiun city using importance performance analysis (IPA). JPPI (Jurnal Penelitian Pendidikan Indonesia), 2024, 10(3): 148-162.

[3] Amri I, Giyarsih S R. Monitoring urban physical growth in tsunami-affected areas: A case study of Banda Aceh City, Indonesia. Geojournal, 2022, 87(3): 1929-1944.

[4] Serat Z, Fatemi S A Z, Shirzad S. Design and Economic Analysis of On-Grid Solar Rooftop PV System Using PVsyst Software. Archives of Advanced Engineering Science, 2023, 1(1): 63-76.

[5] Gu Y, Zhu Q, Nouri H. Identification and U-control of a state-space system with time-delay. International Journal of Adaptive Control and Signal Processing, 2022, 36(1): 138-154.

[6] SS V C, HS A. Nature inspired meta heuristic algorithms for optimization problems. Computing, 2022, 104(2): 251-269.

[7] Karras T, Aittala M, Aila T, Laine S. Elucidating the design space of diffusion-based generative models. Advances in neural information processing systems, 2022, 35(5): 26565-26577.

[8] Li C, Conejo A J, Liu P, Omell B P, Siirola J D, Grossmann I E. Mixed-integer linear programming models and algorithms for generation and transmission expansion planning of power systems. European Journal of Operational Research, 2022, 297(3): 1071-1082.

[9] Chraibi A, Ben Alla S, Touhafi A, Ezzati A. A novel dynamic multi-objective task scheduling optimization based on Dueling DQN and PER. The Journal of Supercomputing, 2023, 79(18): 21368-21423.

[10] Chi J, Zhou X, Xiao F, Lim Y, Qiu T. Task Offloading via Prioritized Experience-based Double Dueling DQN in Edge-assisted IIoT. IEEE Transactions on Mobile Computing, 2024, 23(12): 14575-14591.

[11] Huang L, Ye M, Xue X, Wang Y, Qiu H. Intelligent routing method based on Dueling DQN reinforcement learning and network traffic state prediction in SDN. Wireless Networks, 2024, 30(5): 4507-4525.

[12] Zhang L, Feng Y, Wang R, Xu Y, Xu N, Liu Z. Efficient experience replay architecture for offline reinforcement learning. Robotic Intelligence and Automation, 2023, 43(1): 35-43.

[13] Zhou C, Huang B, Hassan H, Fränti P. Attention-based advantage actor-critic algorithm with prioritized experience replay for complex 2-D robotic motion planning. Journal of Intelligent Manufacturing, 2023, 34(1): 151-180.

[14] Carlsson H, Pijpers R, Van Melik R. Day-care centres for older migrants: spaces to translate practices in the care landscape. Social & Cultural Geography, 2022, 23(2): 250-269.

[15] Kikuta J, Kamagata K, Takabayashi K, Taoka T, Yokota H, Andica C, Aoki S. An investigation of water diffusivity changes along the perivascular space in elderly subjects with hypertension. American Journal of Neuroradiology, 2022, 43(1): 48-55.

[16] Boeing G, Higgs C, Liu S, Giles-Corti B, Sallis J F, Cerin E. Using open data and open-source software to develop spatial indicators of urban design and transport features for achieving healthy and sustainable cities. The Lancet Global Health, 2022, 10(6): 907-918.

[17] Cao J, Wang X, Wang Y. An improved Dueling Deep Q-network with optimizing reward functions for driving decision method. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 2023, 237(9): 2295-2309.

[18] Saglam B, Mutlu F, Cicek D, Kozat, S. Actor Prioritized Experience Replay (Abstract Reprint). Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(20): 22710-22710.

[19] Bai Z, Fan X, Jin X, Zhao Z, Wu Y, Oenema O. Relocate 10 billion livestock to reduce harmful nitrogen pollution exposure for 90% of China's population. Nature Food, 2022, 3(2): 152-160.

[20] Williams R A. From racial to reparative planning: Confronting the white side of planning. Journal of Planning Education and Research, 2024, 44(1): 64-74.

All Element Selection Method in Classroom Social Networks and Analysis of Structural Characteristics

Zhaoyu Shou[1], Zhe Zhang[2], Jingquan Chen[3], Hua Yuan[4], Jianwen Mo[5]
School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, China[1, 2, 3, 4, 5]
Guangxi Wireless Broadband Communication and Signal Processing Key Laboratory[1]

***Abstract*—To deeply investigate the complex relationship between learners' structural characteristics in classroom social networks and the dynamics of learning emotions in smart teaching environments, an innovatively improved RP-GA. All Element Selection Method based on genetic algorithm is proposed. The method calculates the importance of factors based on the random forest model and guides the population initialization together with random numbers to achieve the differentiation and efficiency of factor selection; and utilized the Partial Least Squares regression model in conjunction with a cross-validation optimization model to enhance the accuracy of fitness evaluation, efficiently tackling the issues of premature convergence and low prediction accuracy inherent in traditional genetic algorithms for factor selection. Based on this method, the elements affecting learning emotions are precisely screened, and the intrinsic links between elemental changes and structural properties are deeply analyzed. Experiments show that RP-GA selects a small and efficient number of key elements on public datasets and significantly improves the prediction performance of classifiers such as SVM, NB, MLP, and RF. The proposed learning sentiment all-essential selection method provides effective conditions for classroom network structure characterization and future learning sentiment computation.**

***Keywords*—Genetic algorithms; element selection; random forest; partial least squares; classroom network**

I. Introduction

With the integration and development of teaching practices and emerging information technologies, smart teaching has become an important driver of global education reform and has enabled interactive learning environments, data-driven decision support, and personalized learning experiences, among others, fundamentally reshaping the process of teaching and learning [1]. As an important carrier of information mapping the dimensions of student interactions, learning effectiveness, cognitive processes, and affective learning, the interaction mechanism between its structural properties and student performance has become a hot research topic [2]. Learning emotion, which refers to the learning-related emotional experiences that students produce in intelligent teaching and learning scenarios, covers the basic emotional states (e.g., concentration, confusion, anxiety, excitement, etc.) of the learning process. It is the emotional experience that accompanies students in the cognitive process related to learning content comprehension and knowledge construction, as well as the social-emotional responses generated in teaching activities such as teacher-student interaction and student-student interaction. These rich and diverse learning emotions do not exist in isolation, but are intertwined and influenced by each other with the structural characteristics of the classroom social network. Acquiring heterogeneous information data on students' academic performance, learning behaviors and social relationships through the smart teaching classroom provides a strong data foundation for analyzing the mutual influence relationship between learning emotions and the structural characteristics of classroom social networks, and lays a solid theoretical foundation and scientific basis for the computation and dissemination of learning emotions in the smart classroom scenario [3,4]. Classroom social networks with positivity can facilitate the transmission of positive emotions, while networks full of negativity may lead to an increase in negative emotions [5] The computation and transmission of learning emotions depend on the structural properties of classroom networks, and the structural properties of classroom networks are dynamically influenced by a variety of factors. Therefore, it is important to select the elements that have a significant effect on learning emotions to construct the classroom network, and then analyze the relationship between the changes in the elements and the structural characteristics of the network.

However, existing research on classroom social networks is constructed based on a series of established factors and provides only a preliminary analysis of their network structural characteristics [6-8]. But it failed to delve into how multiple factors affect the dynamics of classroom social networks, as well as how the computation and dissemination of learning emotions intertwine and interact with the structural properties of classroom social networks, a dynamic mechanism that is still under-researched in the existing research. In order to deeply explore the intrinsic connection between the structural characteristics of classroom social networks and learning emotions, this paper proposes a method called Random Partial-Genetic algorithm (RP-GA) for the selection of all elements of classroom social networks. In order to verify the reliability and validity, six public datasets were selected for systematic experiments, and the performance of the RP-GA algorithm on different datasets was comprehensively evaluated by using scientifically reasonable evaluation indexes and strictly controlling experimental variables. Meanwhile, in order to fit the actual needs of smart teaching scenarios, this study specifically collects the SCLE-Dataset of students' learning emotions under smart classrooms. This dataset covers multi-dimensional information of students in the learning process, including but not limited to academic performance, classroom interaction behavior and so on. The RP-GA algorithm is used to mine and analyze the SCLE-Dataset in depth, and the elements that have a significant impact on learning emotions are selected

comprehensively and meticulously. Based on the selected elements, an all-element classroom network model is constructed, which fully considers the dynamic coupling mechanism between the elements and the structural characteristics of the classroom network. Using graph theory, complex networks and other theories and methods, we deeply analyze the dynamic change law between the elements and the network structure characteristics. The results of this research provide solid theoretical support for the future accurate calculation of the propagation of learning emotions in the classroom network, help educators better understand the process of the generation and development of students' learning emotions, and provide a scientific basis for optimizing teaching strategies and creating a positive learning atmosphere.

The main contributions of this study can be summarized in the following two points:

1) An innovative approach to full-factor selection is proposed in RP-GA. The RP-GA algorithm is an innovative extension of genetic algorithm (GA), based on the innovative improvements of factor importance and random number co-guided population initialization and R^2-based PLS regression fitness assessment, which significantly improves the search efficiency and predictive accuracy of the classifier compared with the existing algorithms. In classroom social network-specific scenarios, RP-GA is able to traverse and evaluate numerous factors affecting learning emotions more efficiently, providing a powerful tool for accurately identifying the elements affecting learning emotions.

2) Construct a multi-layer heterogeneous classroom network model based on all elements, while analyzing key structural properties such as network diameter, average degree, clustering coefficient and so on. In-depth exploration of the role of the mechanism that influences the dynamic impact of changes in the emotional elements of learning on the structural characteristics of the network. The construction and analysis of the all-element classroom network model not only provides a solid theoretical foundation for the effective use of network structural characteristics to guide and regulate the learning affective state in the field of educational technology, but also provides a scientific basis for future research on learning affective computation and dissemination.

II. Related Work

In smart teaching scenarios, educational researchers are faced with the challenge of accurately screening out the elements affecting learning emotions from massive data [9]. As a powerful global optimization search algorithm [10], genetic algorithm shows great potential in model optimization and parameter tuning with its robustness, adaptivity and parallel processing capability. In recent years, a variety of improved algorithms based on genetic algorithms have shown remarkable achievements in the field of factor selection research. Izabela et al. [11] proposed a Genetic Algorithm with Aggressive Mutation and Reduced Factors (GAAMmf), which scales down the number of factors while performing the factor selection. Aram et al. [12] proposed an Alternating Sorting Method Genetic Algorithm (ASMGA), which combines genetic algorithm and maximum bounded factor selection in a hybrid packing-filtering algorithm. Deng et al. [13] proposed a Factor Threshold Guided Genetic Algorithm (FTGGA), which first applies the ReliefF algorithm to filter the redundant factors, and then further evaluates the retained subset of factors by FTGGA, which exhibits higher classification accuracy and a smaller subset of factors on a 12-gene microarray dataset. However, these methods mainly focus on the optimization of the algorithm itself, ignoring the important impact of factor quality on algorithm performance. Studies have shown that factor selection and factor importance assessment are critical to the performance of genetic algorithms, and researchers have begun to explore factor importance assessment methods in depth, providing an effective means of assessing the contribution of factor importance to the predictive power of a model. Razmjoo et al. [14] proposed two incremental ranking methods for factors for classification tasks with the aim of developing a factor importance ranking strategy for effective removal of irrelevant factors from the classification model. Kaneko et al. [15] proposed a new Cross-Validated Ranking Factor Importance (CVPFI) method, which achieves stable computation even with a small number of samples and is capable of assessing the importance of strongly correlated factors. Du et al. [16] computed the importance of influencing factors and improved the prediction accuracy of the risk of conflict by constructing a Random Forest model that contains multiple decision trees. Although these methods perform well in specific situations, they fail to effectively utilize factor importance to guide the initial population construction of the genetic algorithm, resulting in inefficient convergence of the algorithm. In order to break through the limitations of existing research, this paper proposes the RP-GA full-factor selection method. The RP-GA algorithm is an innovative extension of the Genetic Algorithm (GA) with two key innovative improvements. Firstly, the algorithm is based on factor importance and random numbers jointly guiding population initialization. Previous studies have failed to fully integrate factor importance into the initial population construction of genetic algorithms, while the RP-GA algorithm, by combining factor importance and random numbers, can increase the proportion of high-quality solutions in the population, accelerate the convergence speed of the algorithm, and thus improve the quality of the final solution. Secondly, R^2-based PLS regression is used for fitness assessment. This innovative fitness assessment can more accurately measure the fitness of individuals compared to traditional methods, providing more effective guidance for the evolution of the algorithm. With these two innovations, the RP-GA algorithm is expected to provide a more efficient and accurate solution for accurately screening the elements affecting learning emotions in smart teaching scenarios.

In the field of classroom social network research, researchers have found that network structural properties [17] (e.g., network diameter, mean degree, clustering coefficient, etc.) are closely related to the exchange of information or the propagation of learning emotions in classroom networks. Tang et al. [18] established a database based on multidimensional data such as student identity, seating relationship, and social relationship, constructed a classroom social network through the seating similarity between learners, and used the CRITIC algorithm and the CRITIC algorithm and entropy weight method to obtain the

combination weights, and proposed GRA-TOPSIS multi-decision fusion algorithm to mine the key student nodes with negative influence. The algorithm can objectively evaluate learners based on the classroom social network and provide a theoretical basis for learning emotion calculation and dissemination. Xie et al. [19] showed that the network density, average degree, and clustering coefficients of the classroom social network have a significant impact on the learning emotion calculation and dissemination of students. Classroom social networks with smaller network diameters and higher network densities are conducive to the dissemination of learning emotions among students; networks with higher clustering coefficients, although conducive to the rapid dissemination of learning emotions within a small group, may form information silos and impede cross-group dissemination. The current field of classroom network research has yielded some results, but most of them follow an established model. Without selecting the factors affecting learning emotions, researchers often consider all the attributes of students, including academic performance, classroom interaction behavior and emotional traits, as potential factors affecting learning emotions, and construct a classroom network model based on them. Subsequently, we focus on analyzing the dynamic changes of students' learning emotions over time within the framework of the network model, in an attempt to reveal the evolution of learning emotions at different stages of teaching and learning. However, this research paradigm has significant methodological flaws, the core of which lies in the failure to analyze the factors of the classroom network in depth and identify the key elements that affect learning emotions. Theoretically, the influence of various student attributes on affective learning is not uniformly distributed, but has a complex hierarchical and causal relationship. Conducting research without distinguishing between critical and non-critical elements will introduce a large amount of redundant information, increase the complexity and computational cost of the research model, and the interfering factors will obscure the real association between affective learning and the structure of the classroom network, thus reducing the validity and reliability of the research.

To address the above limitations, this study will use the innovative RP-GA algorithm to accurately identify the key elements affecting learning emotions in smart classroom scenarios, deeply analyze the dynamic changes between the elements and the network structure characteristics, and construct a multi-layer heterogeneous all-element classroom network model. This innovative approach not only improves the accuracy of feature selection and the convergence efficiency of the algorithm, but also more comprehensively grasps the complex impact of network structural characteristics on the propagation of learning emotions, providing new theoretical perspectives and technical support for the practice of smart classroom teaching.

III. Data Set and Experimental Parameters

A. Introduction to Data Sets

In the study of this paper, the historical academic dataset of students collected from Kaggle, an open machine learning database, and the constructed Student Sentiments for Learning in a Smart Classroom dataset (SCLE-Dataset) are used as the experimental datasets, and Table I demonstrates the important information about the use of public datasets in this paper.

The dataset of students' learning emotions under the smart classroom is composed of video data of 67 students studying a course in the smart classroom of a university as well as the results of a questionnaire survey, in which the learning emotions are acquired with reference to the method in the literature [22], and the learning emotions are classified into five grades: very positive, relatively positive, neutral, relatively negative, and very negative. These categories cover not only the intensity but also the positive and negative polarities of learning emotions, providing a multidimensional framework for analyzing the dynamics of learning emotions during the learning process. The SCLE-Dataset was divided into 63 sub-datasets based on the order of the length of time the 63 knowledge points were taught within the course, with each sub-dataset corresponding to an individual knowledge point and consistent factor dimensions. Table II shows the descriptions of the fields in the SCLE-Dataset.

TABLE I IMPORTANT CHARACTERISTICS OF THE PUBLIC DATASETS

Datasets	Number of instances	Number of factors	Name of the target attribute in the dataset
data[20]	4424	36	target
Student-mat[21]	395	32	G3
Student-pro[21]	649	32	G3
xAPI-Edu-Data[21]	480	16	Class
Students Performance[21]	1000	7	Writing score
Turkiye Student[21]	5820	32	Attendance

TABLE II DESCRIPTION OF FIELDS IN SCLE-DATASET

Name	Description	Name	Description	Name	Description
ID	Student number	Classroom test score	Classroom test scores	Interaction	Number of interactions
K point	Knowledge point number	Eng_grade	Grades in English	Use phone	Number of cell phone uses
Class	Date of class	Math_grade	math grade	Take notes	Number of notes taken
Gender	Gender of students	Position preferences	Seat Selection Preference Area	Lean on table	Number of times lying on the table
Character	Student Character	Head on rate	percentage of heads up	Prop up head	Number of headrests
Club	Participation in associations or not	Friend nomination	Number of friend nominations	Yawn	Number of yawns
Competition	Participation in competitions or not	Friend nominated	Number of times nominated by friends	Award rating	Scholarship level
Committee	Serve on a class council or not	Bad for learn	Number of adverse learning nominations	Fail	Number of subjects failed
Dorm	Dormitory number	Good for learn	Number of nominations for enabling learning	Learning emotions	Student Learning Emotions at a Knowledge Point

B. Data Standardization

Data standardization is a key step in the factor selection process, and standardization is the process of transforming each factor column Z_j into a new factor column X_j with a mean of 0 and a standard deviation of 1. It can eliminate scale differences between factors and improve the effectiveness of factor selection and model training. If the factors are not normalized, factors with a large range of scale values may dominate the calculation of factor importance, while factors with a small range of scale values may be ignored. This will lead to inaccurate results in the assessment of factor importance, ignoring the possibility that some small-valued factors are more important to the prediction results.

C. Classifier Parameters

The classifiers used in this paper are Support Vector Machine (SVM), Plain Bayes (NB), Multi-Layer Perceptron (MLP) and Random Forest (RF) classifiers, and the parameter settings of each classifier are shown in Table III.

TABLE III NAMES AND PARAMETERS OF THE CLASSIFIERS USED IN THE EXPERIMENT

Name	Parameters
Support Vector Machine (SVM)	C=1.0, kernel="rbf", gamma="scale", tol=1e3, cache_size=200
Plain Bayes (NB)	var_smoothing=1e-9
Multi-Layer Perceptron (MLP)	hidden_layer_ sizes=50, max_iter=1000, learning_rate_init=0.01, random_state=123
Random Forest (RF)	n_estimators=100, max_depth=None, min_samples_split=2, max_features="auto"

IV. RP-GA ALGORITHM

In this section, two improvements of the RP-GA algorithm over the GA algorithm will be presented: first, the use of factor importance versus random numbers to guide population initialization; and second, the dynamic selection of the number of components (i.e., the dimensionality of the PLSs) of the partial least squares PLS model in the fitness function, which allows the model to adaptively adjust its complexity to match the number and nature of the selected factors. The improved RP-GA algorithm optimizes not only the number of factors but also the prediction performance based on the selected set of factors. Fig. 1 illustrates the overall framework diagram of the RP-GA algorithm:

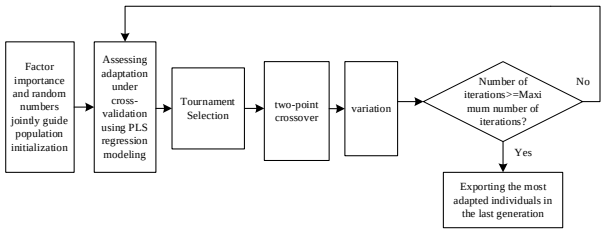


Fig. 1. Flowchart of RP-GA algorithm implementation.

A. Importance Calculation Based on Random Forest Model

Random Forest is an integrated learning method designed to improve the prediction accuracy of a model by constructing multiple decision trees. The main steps include Bootstrap sampling, decision tree construction, model training and prediction. In calculating factor importance, this paper focuses on quantifying the extent to which each factor contributes on average across all decision trees, rather than the relative proportion of each factor's contribution to the totality of all factors. The evaluation method uses mean square error (MSE) as an impurity metric by comparing the reduction in model prediction error (MSE) with and without node splitting using specific factors. Mean Square Error, a key indicator of predictive performance, calculates the average of the squares of the differences between the model's predicted values and the actual observed values, which can be accurately assessed to quantify the level of importance of each factor by comparing the change in MSE before and after factor use.

The steps for calculating the importance of factors are as follows:

1) Calculation of baseline MSE: The prediction of the dataset using the trained model and calculating the mean square error between the predicted and actual values can be expressed as:

$$MSE_{baseline} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (1)$$

Where y_i is the actual target value, $\hat{y}_i$ is the model predicted value, and N is the sample size.

2) Remove factor X_j from the model, then retrain the model and compute a new MSE

$$MSE_{exclude,j} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_{i,j})^2 \quad (2)$$

Where $\hat{y}_{i,j}$ is the predicted value of the model with factor X_j excluded.

3) For each decision tree i, calculate the reduction in mean square error before and after excluding each factor X_j from that decision tree, denoted by $VIM_{ij}^{(MSE)}$

$$VIM_{ij}^{(MSE)} = MSE_{baseline} - MSE_{exclude,j} \quad (1)$$

4) Sum the mean squared error reductions of factor X_j across all decision trees and divide by the number of decision trees, n, to arrive at the average impurity reduction of the factor, i.e., the factor importance score

$$VIM_j^{(MSE)} = \frac{1}{n}\sum_{i=1}^{n} VIM_{ij}^{(MSE)} \quad (4)$$

If $VIM_j^{(MSE)}$ is larger, it indicates that factor has a greater impact on the predictive ability of the model and therefore the factor is more important. Table IV shows the results of the importance ratings of all the factors affecting learning emotions:

TABLE IV RESULTS OF FACTOR IMPORTANCE SCORES

Factor name	score	Factor name	score	Factor name	score
K point	0.183	Eng_grade	0.019	Use phone	0.010
Class	0.021	Math_grade	0.014	Take notes	0.038
Gender	0.022	Position preferences	0.048	Lean on table	0.022
Character	0.023	Head on rate	0.137	Prop up head	0.095
Club	0.010	Friend nomination	0.036	Yawn	0.010
Competition	0.008	Friend nominated	0.031	Award rating	0.005
Committee	0.009	Bad for learn	0.006	Fail	0.026
Dorm	0.065	Good for learn	0.030		
Classroom test score	0.101	Interaction	0.030		

B. Population Initialization Based on Importance and Random Numbers

In traditional genetic algorithms, the individuals of the initial population are usually generated by random generation, although the diversity of the initial population generated in this way is high, a large number of low-quality individuals will be introduced, which reduces the convergence speed and wastes the computational resources. To overcome this limitation, this paper customizes a method for generating a subset of factors that uses factor importance and random numbers to guide the initialization of individuals, with the initial value of each factor depending on its importance in the random forest model. For each factor, if the randomly generated number is less than the importance of the factor, its initial value is randomly determined by the importance of the factor and the specified minimum and maximum bounds. Otherwise, it is set to the minimum boundary value and the important factors are prioritized for model training.

The process in which factor importance and random numbers jointly guide population initialization is illustrated in Fig. 2. The random_threshold and random_value in the figure are random numbers generated between min_bound and max_bound, labeled differently in order to distinguish between the two generated random numbers. The number of factors in the dataset used in this paper is shown in Table 1. The number of individuals in the initialized population is set to 50, striking a balance between the speed of convergence and the maintenance of solution diversity.

When studying the complexity of factors affecting students' learning emotions in classroom networks, the use of factor importance and random numbers together to guide population initialization demonstrated the following significant advantages:

1) Avoidance of over-conditioning: If a fixed threshold is used to compare the importance of factors, the initial values of the factors can be influenced by the distribution of the data, potentially leading to over-adjustment issues, i.e., "biased initialization." By combining the importance of factors with dynamically generated random numbers for initialization, it becomes possible to consider the potential influence of factors on learning emotions in a more balanced manner, making the analysis more objective and scientific.

2) Enhanced exploration and adaptation: In the complex environment of the classroom network, learning emotions are influenced by a wide array of intertwined factors characterized by high degrees of dynamism and uncertainty. By incorporating random numbers into the initialization process, it not only boosts the diversity of individuals in the initial population but also mimics the randomness of learning emotional changes in diverse contexts. This enhances the algorithmic exploration flexibility and scope, enabling the algorithm to comprehensively investigate the interactions between factors and reducing the risk of converging to local optimal solutions.

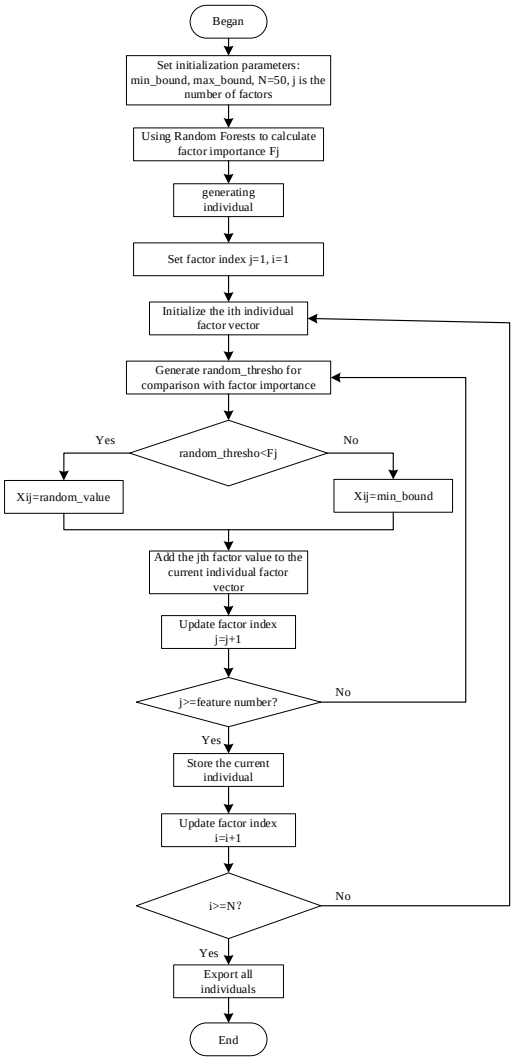


Fig. 2. Flowchart for initializing the population.

C. R^2-Based Adaptation Assessment of PLS Regression

For each individual in the initialized population of individuals that have been guided by the importance of the factors, the factors whose factor weights exceed a set threshold are selected to form a subset of the factors. PLS regression models were applied to a subset of factors in each individual and performance was evaluated within a cross-validation framework. In the fitness function, the number of components (i.e., regression dimensions) of the PLS model is dynamically selected based on the rank of the selected subset of factors. The method improves the accuracy of the fitness assessment by finding the optimal number of PLS components on a selected subset of factors using a partial least squares regression model and evaluating the actual impact of the selected factors on the model performance in terms of maximizing the coefficient of determination (R^2).

In the 3-fold cross-validation framework, the following operations are performed for each fold: the model is trained using the training set; predictions are made on the validation set thus obtaining the predicted value $\hat{y}_i$; the R^2-value on the validation set for each fold is calculated; and the largest R^2-value is selected as the fitness value for that individual, which can be expressed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

where y_i is the actual value of the ith sample in the validation set, $\hat{y}_i$ is the value predicted by the model for the ith sample, and $\bar{y}$ is the mean of the actual values of all samples in the validation set.

In the RP-GA algorithm, PLS regression modeling with cross-validation was chosen as the key innovation based on the following considerations:

1) Adaptive adjustment and optimization of model complexity: In the fitness function of the RP-GA algorithm, the dynamic selection of the number of components of the PLS model (i.e., the number of dimensions of the PLS) allows the model to adaptively adjust its complexity according to the number and nature of the selected factors. Through cross-validation, the optimal number of PLS model components is determined, thus ensuring model accuracy while avoiding over-complexifying the model and improving the optimization efficiency of the algorithm. This adaptive adjustment mechanism enables the RP-GA algorithm to balance model complexity and prediction performance more effectively.

2) PLS regression model: The PLS regression model reduces the impact of data noise on the model by extracting the most representative components, thus improving the robustness of the algorithm. Meanwhile, cross-validation further enhances the stability of the algorithm by dividing the dataset multiple times for training and validation, which reduces the bias caused by unreasonable dataset division. This combined use enables the RP-GA algorithm to show more stable performance when facing different datasets and problems.

V. Experimental Results and Analysis

A. RP-GA Comparison Experiment

Among the current factor selection algorithms, Forward algorithm [23] and PSO algorithm [24] are more widely used. Forward algorithm can effectively and efficiently select key factors by virtue of its step-by-step forward strategy in the process of feature screening. PSO algorithm can quickly find the best in the search space by virtue of the principle of simulating the foraging behavior of bird flocks. The two algorithms have their own unique advantages, and have been widely used in a number of practical application scenarios. Each of them has its own unique advantages and has been widely used in many practical application scenarios. In addition, using the correlation Pearson coefficient as the evaluation criterion for the subset of factors is more stable. In this experiment, we use the Pearson correlation coefficient as the evaluation criterion for the CL-GA

algorithm and as a comparative model, and the following is the formula for the correlation coefficient [21]:

$$R(f_i,c) = \frac{Cov(f_i,c)}{\sqrt{Var(f_i)Var(c)}} \quad (6)$$

where $R(f_i,c)$ denotes the correlation between factor f_i and target c , $Cov(f_i,c)$ denotes the covariance between factor f_i and target c , and $Var(f_i,c)$ denotes the variance between factor f_i and target c . In addition, the forward selection algorithm is also used as a comparison algorithm.

In this paper, we conduct experiments on six publicly available datasets as well as on a specific sub-dataset of the SCLE dataset and compare the performance of the RP-GA algorithm with other algorithms when using SVM, NB, MLP and RF classifiers. We measure the performance of the model using the accuracy ratio (acc), which is defined as the ratio of the number of samples correctly predicted by the classifier to the total number of samples. It is calculated using the formula:

$$acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

From Table V, it can be seen that RP-GA algorithm outperforms CL-GA algorithm, Forward algorithm, and PSO algorithm on all experimental datasets. Differences in the performance of the algorithms not only depend on the characteristics of the algorithms themselves, but may also be closely related to the characteristics of the dataset. In the case of the Student-pro [21] dataset and the Turkiye Student [21] dataset, for example, although both have a total number of factors of 32, there is a significant difference in the number of samples, which has an impact on the algorithm's prediction accuracy. The Student-pro dataset with a sample size of 649 is derived from Portuguese language course grades in Portuguese secondary schools. The smaller sample size allows the algorithm to quickly capture key patterns and achieve higher prediction accuracy. The Turkiye Student dataset, with a sample size of 5820, is derived from Turkish university course evaluations, and the large number of samples brings rich information but also introduces more noise and uncertainty. In addition, the complexity of the university course evaluation data, including subjective student evaluations and diverse course content, puts the algorithm under computational pressure and the risk of overfitting, and it is easy to learn local non-universal patterns, which affects the prediction accuracy. In conclusion, RP-GA better adapts to various datasets, suits small-to-medium-sized samples with clear features, and shows superior experimental performance.

The four datasets data, Student-mat, Student-pro, and Turkiye Student, which have a higher number of factors, were selected to demonstrate the comparison of the number of factors after selection, and from Table VI it can be seen that compared to CL-GA algorithm, Forward algorithm, and PSO algorithm, the RP-GA algorithm reduces the number of factors after selection, improves the training speed and prediction accuracy of the classifier.

TABLE V COMPARING THE PREDICTION ACCURACY OF CL-GA, FORWARD, PSO AND RP-GA ON DIFFERENT CLASSIFIERS

Datasets	Algorithms	Prediction accuracy of classifiers			
		SVM	*NB*	*MLP*	*RF*
data	CL-GA	0.718	0.721	0.755	0.752
	Forward	0.751	0.736	0.715	0.735
	PSO	0.721	0.704	0.723	0.732
	RP-GA	**0.855**	**0.808**	**0.867**	**0.865**
Student-mat	CL-GA	0.807	0.798	0.815	0.849
	Forward	0.807	0.840	0.840	0.840
	PSO	0.816	0.823	0.832	0.832
	RP-GA	**0.832**	**0.874**	**0.874**	**0.899**
Student-pro	CL-GA	0.841	0.831	0.780	0.846
	Forward	0.846	0.862	0.856	0.862
	PSO	0.843	0.851	0.843	0.857
	RP-GA	**0.918**	**0.882**	**0.923**	**0.928**
xAPI-Edu-Data	CL-GA	0.701	0.688	0.694	0.806
	Forward	0.743	0.688	0.688	0.736
	PSO	0.717	0.688	0.679	0.742
	RP-GA	**0.764**	**0.701**	**0.771**	**0.833**
Students Performance	CL-GA	0.643	0.727	0.753	0.697
	Forward	0.697	0.717	0.710	0.717
	PSO	0.691	0.704	0.721	0.721
	RP-GA	**0.770**	**0.780**	**0.780**	**0.810**
Turkiye Student Evaluation	CL-GA	0.402	0.346	0.436	0.442
	Forward	0.410	0.418	0.416	0.416
	PSO	0.423	0.444	0.423	0.432
	RP-GA	**0.531**	**0.480**	**0.546**	**0.536**
SCLE-Dataset	CL-GA	0.667	0.619	0.714	0.571
	Forward	0.619	0.619	0.714	0.667
	PSO	0.623	0.619	0.714	0.651
	RP-GA	**0.700**	**0.700**	**0.750**	**0.700**

TABLE VI FACTOR RESULTS AFTER CL-GA, FORWARD, PSO AND RP-GA SEARCHES

Datasets	Results of factor selection			
	CL-GA	*Forward*	*PSO*	*RP-GA*
data	1,2,8,12,17,18,19,20,22,29,30,31,32,34,35	2,4,7,9,10,13,17,24,25,31,35	1,2,8,9,10,13,17,22,24,25,29,31,35	**3,12,17,20,21,23,24,25,26,28**
Student-mat	3,5,6,7,9,10,12,13,14,19,20,23,24,32	9,10,18,22,24,26,27	1,5,6,7,9,10,12,13,18,22,24,26	**3,8,15,16,17,31**
Student-pro	1,2,3,4,8,10,11,12,13,19,23,24,27,28,29,30,32	3,12,16,18,19,21,22,24,25	1,2,3,4,8,10,12,13,19,21,22,23,24,25	**2,15,25,31,32**
TurkiyeStudent Evaluation	3,4,7,9,11,14,16,17,19,20,29,30	1,2,3,14,17,22	1,2,3,4,7,9,11,14,16,17,29,30	**1,3,4,13,21,31**

B. Analysis of Ablation Experiments

In order to verify the validity of population initialization through factor importance and random number co-guiding in RP-GA and the use of PLS regression model with cross-validation to assess the fitness. The GA algorithm is used as the baseline model for the ablation experiments, and the data preprocessed DATA dataset is used for the experiments, using the classifiers in Table III and keeping the parameters consistent, and ablating the improvement modules one by one to obtain four sets of experimental data, as shown in Table VII. In the table, A represents the improvement point in Chapter IV.B (using factor importance and random number to jointly guide population initialization), and B represents the improvement point in Chapter IV.C (using PLS regression model with cross-validation to assess fitness). From the ablation experimental data, it can be found that baseline improved by 0.091, 0.041, 0.05, and 0.061 after adding improvement point A with SVM, NB, MLP, and RF classifier prediction accuracies of 0.712, 0.720, 0.751, and 0.751, respectively, and the final effect improved by 0.143 after adding improvement point B. The final results were, 0.088, 0.116, 0.114. The above improvement points made significant contributions to the performance enhancement of the RP-GA algorithm in terms of the efficiency of the factor selection and the accuracy of the adaptation assessment, obtaining a large prediction accuracy enhancement, which fully proved the effectiveness of the two-point improvement strategy proposed in this paper. The visualization of the ablation experiment is shown in Fig. 3.

TABLE VII ABLATION EXPERIMENT

A	B	Prediction accuracy of classifiers			
		SVM	*NB*	*MLP*	*RF*
Baseline		0.712	0.720	0.751	0.751
✓	×	0.803	0.761	0.801	0.812
×	✓	0.793	0.758	0.796	0.784
✓	✓	**0.855**	**0.808**	**0.867**	**0.865**

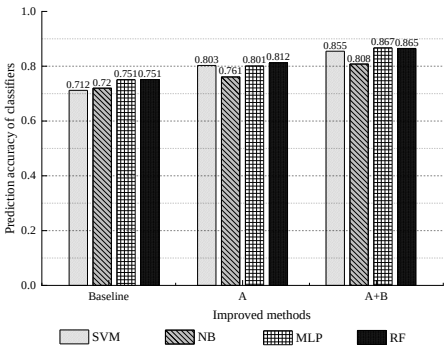


Fig. 3. Intuitive display of ablation experiments.

C. Convergence Analysis

The fitness function is used to measure the degree of superiority or inferiority of an individual (chromosome) in the problem environment. During the iteration of the algorithm, the fitness of individuals in the population changes in each generation. In order to further verify that the factor importance and random number co-guided population initialization proposed in this paper can improve the convergence speed of the RP-GA algorithm, we analyze the average fitness and standard deviation fitness change process of each generation of the population in the iterative process of $RP-GA_{lif}$ (lack of factor importance and random number co-guided population initialization) and RP-GA.

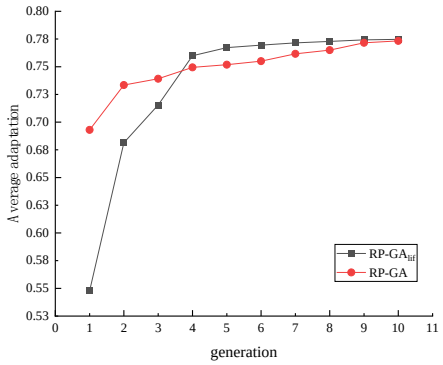


Fig. 4. The changing trend of the average fitness of RP-GA and RP − GA_{lif}.

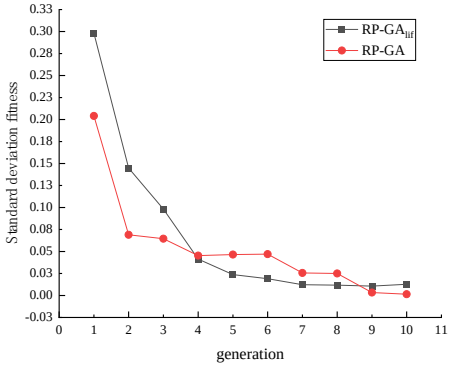


Fig. 5. The changing trend of the standard deviation fitness of RP-GA and RP − GA_{lif}.

From the average fitness curves in Fig. 4, RP-GA converges faster in the early stage, thanks to its method of factor

importance and random number co-guiding population initialization, which enables the initial population to be more reasonably distributed in the solution space. As the iteration proceeds, although the average fitness of RP-GA and $RP-GA_{lif}$ both tend to stabilize, RP-GA converges to a region with higher fitness, indicating that it is better in convergence effect. And RP-GA finally reaches a higher average fitness, which indicates that RP-GA is more advantageous than in finding the optimal solution, and the special population initialization helps to avoid falling into the local optimal solution and easier to find the region of the global optimal solution.

From the standard deviation fitness curve in Fig. 5, the standard deviation fitness of RP-GA decreases faster in the early stage, which is due to the fact that it adopts the factor importance and random number to guide the population initialization, so that the population quickly concentrates in the more optimal direction. With the increase of iteration, the standard deviation of both RP-GA and $RP-GA_{lif}$ tends to stabilize and have smaller values, and the standard deviation of RP-GA is slightly smaller. This indicates that the population convergence of RP-GA is better, the individual adaptations are more consistent, and it can effectively guide the population to converge in the process of evolution, reduce the differences between individuals, and help to find a more optimal solution.

D. Analysis of SCLE-Dataset Experimental Results

SCLE-Dataset is composed of data from different knowledge points of freshman students at a university while studying a course. Considering the existence of different elements in each knowledge point that affect students' learning emotions, this paper evaluated each knowledge point using the RP-GA algorithm. Table VIII shows all the knowledge points elements results and the prediction accuracy of the NB classifier. The elements of all knowledge points were counted statistically, from which the top five elements were filtered as the key elements affecting students' learning emotions, which were gender, head-up rate, preferred area for seat selection, number of friend nominations, and number of favorable study nominations.

TABLE VIII RESULTS OF ELEMENTS FILTERED WITHIN ALL KNOWLEDGE POINTS AND ACCURACY

Knowledge point number	Element name	Accuracy of NB
Point 1	Classroom test score， Award rating	0.85
Point 2	Classroom test score， Award rating	0.85
Point 3	Gender, Friend nominated	0.60
⋮	⋮	⋮
Point 61	Position preferences, Head on rate, Lean on table	0.70
Point 62	Math grade, Head on rate, Use phone, Lean on table	0.95
Point 63	Gender, Committee, Math_grade, Head on rate, Good for learn, Take notes	0.75

E. Characterization of the Structure of the Total Element Classroom Network

In order to further reveal the synergistic effect of key elements and enhance the classroom learning experience, the key elements affecting learning emotions are integrated into the framework of analyzing the structural properties of classroom networks, and the potential influence of key elements on the propagation of students' learning emotions is explored by quantifying and analyzing the structural properties of classroom networks. The structural characteristics of the multi-layer heterogeneous all-element classroom network G_{MHLA} based on SCLE-Dataset influencing the elements of students' learning emotions are analyzed experimentally, and its multi-layer heterogeneous network G_{MHLA} construction method is utilized in the literature [19].

Table IX shows that the all-factor classroom network G_{MHLA} constructed on the basis of the key elements (students' gender, head-up rate, seating information, friends' nomination, and favorable learning nomination) has the properties of low network diameter, low average path length, high network density, high average degree, high average weighting, and low clustering coefficient. These structural properties reflect more connections of nodes in the multilayer heterogeneous all-factor classroom network, shorter paths for learning emotions, and easier cross-cluster dissemination of learning emotions among small groups. Fig. 6 shows a visualization of the structural changes in the classroom network caused by the change of student seating information in the first, fourth, and seventh classes, and it can be found that the denseness of the connections between nodes in the network increases significantly with the increase in the number of classes. Fig. 7 shows a slight increase in the density of connections between the nodes in the first class as the number of points taught increases and the students' head-up rate changes, due to the fact that the students are unfamiliar with each other and have established fewer connections with each other in the first class.

Table X shows the analysis of the structural characteristics of the total classroom network with different classroom seating variations. It can be found that as the number of classes increases, the structural characteristics of the all-factor classroom network change significantly, which is manifested in the reduction of network diameter and average path length, the enhancement of network density, the growth of average degree and average weighting degree, and the decrease of clustering coefficient, and the structural all-factor classroom network is more compact and efficient in the seating relationship of the last classroom, and all these changes together constitute a network environment more favorable to emotion dissemination and network environment for emotion transmission.

Table XI shows the characterization of the whole-factor classroom network structure in terms of changes in head-up rates for different knowledge points in the first class. It can be found that as the number of knowledge points increases, the structural characteristics of the whole-factor classroom network change, which is manifested in the reduction of the average path length, the enhancement of the network density, the growth of the average degree and the average weighting degree, and the decrease of the clustering coefficient.

On the basis of analyzing the structural characteristics of the network, this paper constructs a multidimensional and comprehensive classroom network model using the key

elements screened by the RP-GA algorithm, and analyzes the mechanism of the key elements and the structural characteristics of the classroom network in depth. The results of the analyses show that the dynamics of the key elements in the classroom network significantly influence and change the structural characteristics in the classroom network. This strongly verifies the effectiveness and scientificity of the RP-GA algorithm in identifying the elements affecting learning emotions, and further proves that the elements selected by the RP-GA algorithm provide an important basis and support for the calculation and dissemination of learning emotions in the future.

TABLE IX STRUCTURAL CHARACTERIZATION OF THE MULTILAYERED HETEROGENEOUS ALL-FACTOR CLASSROOM NETWORK

Network	Network diameter	Average path length	Network density	Average degree	Average weighted degree	Minimum degree	Clustering factor
G_{MLHA}	3	2.842	0.271	8.657	20.49	7	0.454

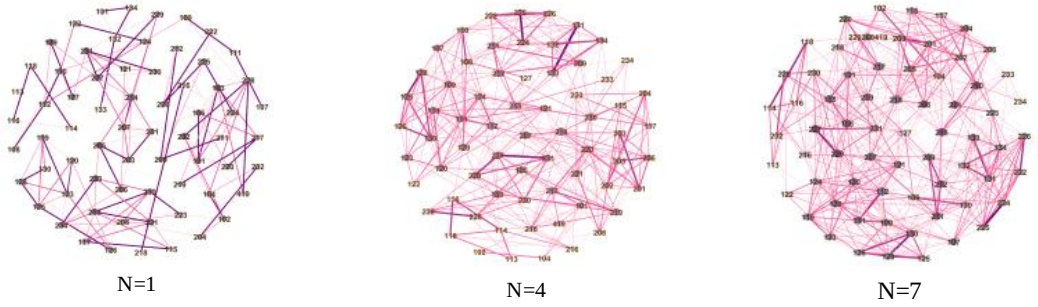


Fig. 6. Network structures under different numbers of classes N.

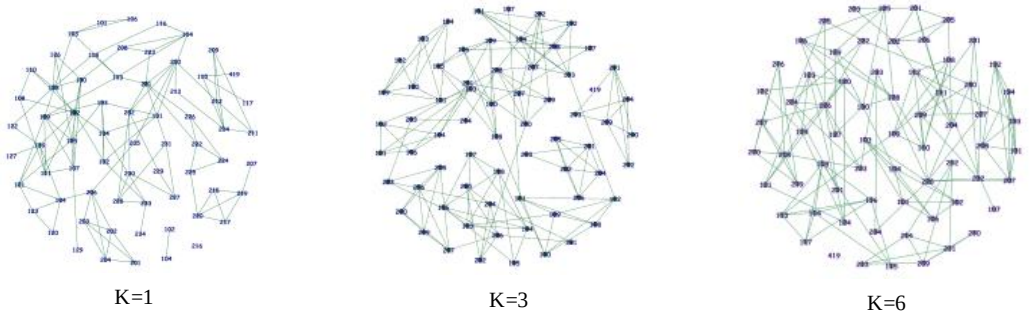


Fig. 7. Network structures under different numbers of knowledge points K.

TABLE X CHARACTERIZATION OF THE NETWORK STRUCTURE AT DIFFERENT NUMBER OF N HOURS OF CLASSES

Network	Network diameter	Average path length	Network density	Average degree	Average weighted degree	Minimum degree	Clustering factor
$G_{MHLA}^{N=1}$	5	3.324	0.153	6.857	15.01	7	0.624
$G_{MHLA}^{N=2}$	5	3.186	0.173	7.089	15.75	7	0.612
$G_{MHLA}^{N=3}$	5	3.169	0.186	7.282	16.45	7	0.610
$G_{MHLA}^{N=4}$	4	3.143	0.195	7.749	17.65	7	0.594
$G_{MHLA}^{N=5}$	4	3.015	0.203	8.102	19.65	7	0.540
$G_{MHLA}^{N=6}$	3	2.943	0.224	8.371	20.23	7	0.471
$G_{MHLA}^{N=7}$	3	2.842	0.271	8.657	20.49	7	0.454

TABLE XI CHARACTERIZATION OF THE NETWORK STRUCTURE AT DIFFERENT NUMBERS OF KNOWLEDGE POINTS K

Network	Network diameter	Average path length	Network density	Average degree	Average weighted degree	Minimum degree	Clustering factor
$G_{MHLA}^{k=1}$	5	3.224	0.122	6.695	14.32	7	0.625
$G_{MHLA}^{k=2}$	5	3.221	0.135	6.701	14.66	7	0.625
$G_{MHLA}^{k=3}$	5	3.221	0.135	6.701	14.66	7	0.625
$G_{MHLA}^{k=4}$	5	3.214	0.140	6.705	14.86	7	0.623
$G_{MHLA}^{k=5}$	5	3.214	0.140	6.743	15.01	7	0.623
$G_{MHLA}^{k=6}$	5	3.121	0.142	6.743	15.01	7	0.616
$G_{MHLA}^{k=1}$	5	3.224	0.122	6.695	14.32	7	0.625

VI. CONCLUSION

In this study, the RP-GA algorithm is innovatively employed to screen key elements influencing learning emotions in smart classroom teaching, and subsequently, a multi - layer heterogeneous all-element classroom network model is built, which, through in-depth exploration of the relationship between elemental variations and network structural features, lays a scientific groundwork for future calculation and dissemination of learning emotions. As the research process was limited by the scope of data acquisition, the collected data could not completely cover all types of smart classroom teaching scenarios, which might have an impact on the comprehensiveness of the RP-GA algorithm in screening the key elements. In factor selection and model optimization, the RP - GA algorithm, with its unique initialization and dynamic adjustment, avoids local optima and boosts the model's prediction accuracy. In this study, we mainly focus on small and medium-sized datasets, for large datasets there may be problems of lower prediction accuracy and higher computational pressure. Even so, this study substantiates the high validity and reliability of the RP-GA algorithm in optimizing classroom network structural characteristics. It not only broadens the theoretical research on classroom networks but also offers a scientific basis and practical guidelines for enhancing teaching quality and evolving learning emotions.

ACKNOWLEDGMENT

The National Natural Science Foundation of China (62177012, 62267003).

Guangxi Natural Science Foundation (2024GXNSFDA010048).

The Project of Guangxi Wireless Broadband Communication and Signal Processing Key Laboratory (GXKL06240107).

REFERENCES

[1] X. Jia. “Research on the role of big data technology in the reform of English teaching in universities,” Wireless Communications and Mobile Computing, 2021..

[2] G. Putnik, E. Costa, C. Alves, et al. “Analysing the correlation between social network analysis measures and performance of students in social network-based engineering education,” International Journal of Technology and Design Education, vol. 26(3), 2016.

[3] A. Charitopoulos, M. Rangoussi and D. Koulouriotis, “On the use of soft computing methods in educational data mining and learning analytics research: A review of years 2010–2018,” International Journal of Artificial Intelligence in Education, vol. 30(3), 2020.

[4] N. Romanov, L. M. Culci, A. I. Daniel, et al. “Artificial intelligence applications and tools IN higher education: an overview,” Proceedings of the SESYR Sustainable Education through European Studies for Young Researchers Jean Monnet Module, 2020.

[5] L. You, “Research on learning sentiment based on group interaction behavior,” Wuhan: Central China Normal University, 2022.

[6] Z. J. Qing, “Intelligent education visualization system based on social network analysis,”2020 International Conference on Robots & Intelligent System (ICRIS). 2020: IEEE, pp. 291-294.

[7] K. Vignery, “From networked students centrality to student networks density: What really matters for student performance?,” Social Networks, vol. 70, pp. 166-186, 2022.

[8] J. K. Yang, “Incorporating network and propagation properties for source identification on social networks,” Hangzhou: Huazhong Dianai Univerity, 2023.

[9] G. X. Dong, Z. Xia, G. Y. Mei, “A study of affective factors affecting college students' autonomous english learning,” Advances in Educational Technology and Psychology, vol. 7(18), pp. 6-17, 2023.

[10] L. X. Deng, H. Y. Chen, H. Y. Liu, H. Zhang, Y. Zhao, “Overview of UAV path planning algorithms,” 2021 IEEE International Conference on Electronic Technology, Communication and Information (ICETCI). 2021: IEEE, pp. 520-523.

[11] R. Izabela, L. Krzysztof, “GAAMmf: genetic algorithm with aggressive mutation and decreasing feature set for feature selection,” Genetic Programming and Evolvable Machines, vol. 24(2), 2023.

[12] K. Y. Aram, S. S. Lam and M. Khasawne, “Cost-sensitive max-margin feature selection for SVM using alternated sorting method genetic algorithm,” Knowledge-Based Systems, vol. 267, 2023.

[13] S. Deng, Y. Li, J. Wang, R. Cao, M. Li. “A feature-thresholds guided genetic algorithm based on a multi-objective feature scoring method for high-dimensional feature selection,” Applied Soft Computing, vol. 148, 2023.

[14] A. Razmjoo, P. Xanthopoulos, Q. P. Zheng, “Feature importance ranking for classification in mixed online environments,” Annals of Operations Research, vol. 276, pp. 315-330, 2019.

[15] H. Kaneko, “Cross-validated permutation feature importance considering correlation between features,” Analytical Science Advances, vol. 3(9-10), pp. 278-287, 2022.

[16] S. K. Du, J. Zhang, Z. J. Han, M. Y. Gong, . “Armed conflict risk prediction and influencing factors analysis based on the random forest model at the grid-month scale: a case study of indochina peninsula,” Journal of Geo-information science, vol. 25(10), pp. 2026-203, 2023.

[17] Z. Kong, Q. Sun, X. Y. Kou, L. F. Wang. “Research on the importance of network nodes based on attribute information and structural characteristics,” Journal of Northeastern University (NatralScience), vol. 43(05), pp. 625-631, 2022.

[18] Z. Y. Shou, M. Tang, H. Wen, et al. "Key student nodes mining in the in-class social network based on combined weighted GRA-TOPSIS method," International Journal of Information and Communication Technology Education (IJICTE), vol. 19(1), pp.1-19, 2023.

[19] Z. Y. Shou, H. Wang, H. B. Zhang, J. L. Xie, J. H. Tang. "The NEDC-GTOPSIS node influence evaluation algorithm based on multi-Layer heterogeneous classroom networks," International Journal of Information and Communication Technology Education (IJICTE), vol. 20(1), pp. 1-24, 2024.

[20] R. Valentim, M. Jorge, B. Luis, et al. "Predict students' dropout and academic success," UCI Machine Learning Repository, vol. 10, 2021, C5MC89.

[21] W. Xiao , P. Ji, J. Hu, "RnkHEU: A hybrid feature selection method for predicting students' performance," Scientific Programming, vol. 2021(1), 2021, 1670593.

[22] Z. Y. Shou, N. Zhu, W. H. Wang, et al. "A method for analyzing learning sentiment based on classroom time-series images," Mathematical Problems in Engineering, vol. 2023(1), 2023, 6955772.

[23] T. Nakanishi, P. Chophuk, K. Chinnasarn, "Evolving Feature Selection: Synergistic Backward and Forward Deletion Method Utilizing Global Feature Importance," IEEE Access,12[2025-01-22].DOI:10.1109/ACCESS.2024.3418499.

[24] X. F. Song, Y. Zhang, D. W. Gong, X. Z. Gao. "A fast hybrid feature selection based on correlation-guided clustering and particle swarm optimization for high-dimensional data," IEEE Transactions on Cybernetics, vol. 52(9), pp. 9573-9586, 2021.

An NLP-Enabled Approach to Semantic Grouping for Improved Requirements Modularity and Traceability

Rahat Izhar[1], Shahid Nazir Bhatti[2], Sultan A. Alharthi[3]
Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand[1]
Department of Software Engineering, College of Computer Science and Engineering, University of Jeddah, Jeddah 21493, Saudi Arabia[2, 3]

***Abstract*—The escalating complexity of modern software systems has rendered the management of requirements increasingly arduous, often plagued by redundancy, inconsistency, and inefficiency. Traditional manual methods prove inadequate for addressing the intricacies of dynamic, large-scale datasets. In response, this research introduces SQUIRE (Semantic Quick Requirements Engineering), a cutting-edge automated framework leveraging advanced Natural Language Processing (NLP) techniques, specifically Sentence-BERT (SBERT) embeddings and hierarchical clustering, to semantically organize requirements into coherent functional clusters. SQUIRE is meticulously designed to enhance modularity, mitigate redundancy, and strengthen traceability within requirements engineering processes. Its efficacy is rigorously validated using real-world datasets from diverse domains, including attendance management, e-commerce systems, and school operations. Empirical evaluations reveal that SQUIRE outperforms conventional clustering methods, demonstrating superior intra-cluster cohesion and inter-cluster separation, while significantly reducing manual intervention. This research establishes SQUIRE as a scalable and domain-agnostic solution, effectively addressing the evolving complexities of contemporary software development. By streamlining requirements management and enabling software teams to focus on strategic initiatives, SQUIRE advances the state of NLP-driven methodologies in Requirements Engineering, offering a robust foundation for future innovations.**

Keywords*—Requirements Engineering (RE); semantic clustering; sentence-BERT; natural language processing (NLP)

I. Introduction

Requirements Engineering is a cornerstone of software development, focusing on the identification, documentation, and management of requirements that guide the design and implementation of software systems [1]. It ensures alignment between user needs and project goals, providing a foundation for system functionality and quality. However, as software systems grow in complexity and scale, managing requirements effectively becomes increasingly challenging. Issues such as redundancy, inconsistencies, and overlapping functionalities not only complicate design processes but also disrupt modularity, traceability, and efficient project execution [2], [3]. These challenges highlight the need for innovative solutions to streamline requirements management, particularly in large-scale projects.

Conventional approaches to managing requirements rely heavily on manual processes, which are often time-consuming, prone to human error, and inadequate for handling large datasets [4]. These traditional methods, while useful for small-scale projects, fail to meet the demands of modern, dynamic software development, where requirements are frequently updated and involve intricate relationships. Automated techniques, particularly those leveraging Natural Language Processing (NLP), have emerged as promising solutions for addressing these limitations. By analyzing textual requirements for semantic relationships, NLP-based methods can uncover patterns and organize requirements efficiently [5]. However, existing methods face limitations in capturing the subtle semantic relationships within diverse or domain-specific datasets, restricting their ability to handle the complexity of real-world requirements engineering tasks.

To address these gaps, this paper presents SQUIRE (Semantic QUIck Requirements Engineering), a novel and structured methodology aimed at automating the grouping of semantically similar requirements into functional clusters. SQUIRE combines state-of-the-art NLP techniques, such as Sentence-BERT (SBERT) embeddings, hierarchical clustering, and a comprehensive preprocessing pipeline, to analyze and group requirements based on their semantic similarity [4], [6], [12]. By focusing on reducing redundancy, enhancing modularity, and improving traceability, SQUIRE provides a practical solution to key challenges in RE, enabling more efficient and effective system design [7].

The methodology represents a significant advancement in the application of NLP to RE, building on recent trends in natural language understanding and clustering algorithms [9], [10], [17], [18]. The authors' work reflects a broader effort in the research community to leverage the capabilities of NLP for addressing critical challenges in RE, including the need for scalability, semantic analysis, and automation [19]. Recent developments in NLP, particularly transformer-based models such as Sentence-BERT (SBERT), have enabled significant improvements in the ability to capture the semantic meaning of textual data, making them well-suited for addressing RE challenges.

The effectiveness of SQUIRE is evaluated across diverse, real-world datasets representing distinct functional domains, including attendance management, e-commerce, and school operations. The methodology's performance is assessed using quantitative metrics such as cohesion, separation, silhouette scores, and the Davies-Bouldin Index [20], [24], [25]. Visualizations using Principal Component Analysis (PCA) further demonstrate the ability of SQUIRE to organize requirements into meaningful clusters, providing clear insights

into the relationships between requirements. These evaluations highlight the robustness of the proposed approach, showcasing its potential to improve modularity and traceability in varied software development contexts.

This research contributes to the ongoing integration of NLP into requirements engineering by demonstrating how advanced language models and clustering techniques can address long-standing challenges in the field [5]. By automating the grouping of requirements, SQUIRE reduces the manual effort required in RE, allowing software teams to focus on higher-value activities such as innovation and strategic planning. Furthermore, the domain-agnostic nature of SQUIRE makes it adaptable to various industries, offering a scalable solution for modern software development needs.

SQUIRE advances the state of requirements engineering by introducing a systematic and scalable methodology for semantic clustering. By leveraging the latest advancements in NLP, it bridges the gap between textual complexity and actionable insights, enabling software teams to design more organized, traceable, and modular systems [22]. This work lays the groundwork for further innovations in automated RE, positioning NLP as a central tool in the evolution of software development practices.

The remainder of this paper is organized as follows: Section II delineates a meticulous review of pertinent literature, elucidating seminal advancements and inherent constraints within NLP-driven requirements engineering. Section III expounds upon the proposed SQUIRE methodology, systematically detailing its preprocessing pipeline, embedding generation mechanisms, and clustering paradigms. Section IV articulates the experimental framework, encompassing the evaluation metrics and empirical assessments employed to ascertain the model's efficacy. Section V presents a rigorous discourse on the obtained findings, critically analyzing the methodological robustness and potential limitations. Lastly, Section VI encapsulates the study's principal contributions and delineates prospective avenues for future research.

II. Related Work

Modern software development has made RE more challenging due to the growing complexity of systems and the volume of requirements. Traditional approaches often rely on manual analysis, which can lead to errors and inefficiencies. To overcome these challenges, researchers have introduced techniques like NLP and clustering algorithms. These methods aim to simplify requirement management by automating processes such as grouping and prioritization. This section reviews recent work in applying these techniques to RE, focusing on their impact, limitations, and future potential.

Radwan et al. [8] proposed the CMHR (Conceptual Mapping for Healthcare Requirements) approach to enhance the analysis of non-functional requirements in healthcare systems. The methodology involves clustering requirements using the K-means++ algorithm based on attributes such as priority and suitability, enabling structured visualization through conceptual mapping. Applied to ventilator requirements, the approach achieved a silhouette score of 0.71 and accurately classified new requirements using a Naïve Bayes classifier. However, the study is limited by its small dataset and focus on a single domain, necessitating further validation across other healthcare systems for broader applicability.

Salman et al. [9] explored semantic clustering of functional requirements (FRs) using Agglomerative Hierarchical Clustering (AHC). The study addresses a gap in software requirements engineering, where functional requirements are typically analyzed manually. Existing works focus primarily on non-functional requirements classification, leaving FR clustering largely unexplored. Previous research has used techniques like Support Vector Machines and ontology-based methods for requirement classification, but functional requirements have lacked similar advancements. This paper introduces a semantic similarity-based approach to group FRs, validated across four software projects. While the proposed method achieved promising results, it relies heavily on vocabulary consistency, limiting its applicability across diverse datasets.

Del Sagrado et al. [10] integrate clustering techniques with the MoSCoW method to automate requirements prioritization in software projects. Their methodology, validated on datasets of varying scales (20, 50, and 100 requirements), demonstrates clustering's efficacy in identifying core requirements. While enhancing decision-making, the approach is constrained by dependency considerations and reliance on subjective estimations, leading to variability in algorithmic performance.

Bakar et al. [11] employ Latent Semantic Analysis (LSA) alongside K-Means and Hierarchical Agglomerative Clustering (HAC) to facilitate software requirements reuse in Software Product Lines (SPL). Evaluations on 27 product review documents reveal HAC's superiority in cluster compactness, whereas K-Means marginally outperforms in external validation. However, HAC is preferred for its grouping precision. Sensitivity to domain-specific data and input parameters remains a limitation, with future work directed at refining clustering optimization.

Das et al. [12] introduce PUBER and FiBER, domain-specific sentence embedding models enhancing similarity detection in natural language requirements. Built on the BERT architecture, PUBER trains on the PURE dataset, while FiBER refines it via fine-tuning. Using cosine similarity, FiBER achieves 88.35% accuracy, surpassing Universal Sentence Encoder and RoBERTa by up to 10%. These models advance classification, similarity detection, and reusability, addressing NLP challenges in software engineering.

The study by Elhassan et al. [13] developed an automated conflict detection model using the Mean Shift clustering algorithm to identify and classify requirement conflicts during elicitation. The methodology involved data transformation based on McCall's quality model and clustering requirements into three categories: conflict-free, partial conflict, and conflicted requirements. Results demonstrated accurate clustering, with conflict-free requirements achieving low standard error (SE) values, validating the model's efficiency. Limitations include a small dataset size (207 observations), restricting generalizability, and challenges in data collection from diverse sources. Future work suggests expanding datasets,

applying the model to varied IS environments, and exploring decision tree algorithms for enhanced detection.

The reviewed studies show that using NLP and clustering in Requirements Engineering has great potential to solve key challenges like redundancy and lack of scalability. However, issues such as handling diverse requirements and adapting to different domains remain. These findings suggest the need for smarter, more flexible methods to address these gaps. By building on existing work, approaches like SQUIRE offer a step forward, combining advanced tools with practical solutions for improving how requirements are managed in real-world software projects.

III. Research Methodology

The proposed methodology, SQUIRE offers an automated and efficient framework for grouping semantically similar requirements into distinct functional clusters. By leveraging NLP techniques, embedding models, and hierarchical clustering, SQUIRE addresses key challenges in requirements engineering, such as redundancy reduction, traceability enhancement, and modular system design.

The methodology comprises five structured stages: Preprocessing, Semantic Embedding Generation, Clustering, Evaluation, and Visualization.

A. Data Preprocessing

The preprocessing stage standardizes and simplifies textual requirements, ensuring consistency and reducing noise to prepare data for embedding generation [27]. This involves a series of transformations applied sequentially: converting text to lowercase to eliminate inconsistencies caused by capitalization, removing punctuation marks to simplify content, and filtering out stopwords such as “the” and “shall” that do not add semantic value [30]. The text is then tokenized into individual words or phrases, enabling detailed analysis, and normalized to unify synonyms or domain-specific terms (e.g., “log in” and “sign in” standardized to “log in”). Table I demonstrates the progressive refinement of requirements through these steps.

TABLE I Example Results of Data Preprocessing Pipeline

Original Requirement	Lowercased	Punctuation Removed	Stopwords Removed	Tokenized
"The system shall allow users to log in."	"the system shall allow users to log in."	"the system shall allow users to log in"	"system allow users log in"	["system", "allow", "users", "log", "in"]
"Users can register themselves for access."	"users can register themselves for access."	"users can register themselves for access"	"users register access"	["users", "register", "access"]

The output of preprocessing is a clean, tokenized, and normalized version of each requirement, ready for embedding generation. This ensures that noise is minimized while retaining the semantic essence of the text.

B. Semantic Embedding Generation

To effectively cluster requirements, each preprocessed input is transformed into a dense vector representation using SBERT, a pre-trained transformer-based model optimized for capturing semantic similarity [23]. SBERT excels at encoding contextual information and relationships between words, making it an ideal choice for analyzing and grouping requirements. Specifically, the all-MiniLM-L6-v2 model is utilized, which generates embeddings with a dimensionality of 384 [31]. These embeddings represent the semantic meaning of textual inputs in a high-dimensional vector space. For instance, a requirement like "system allow users log in" is converted into a numerical vector (e.g., E = [0.23, -0.45, 0.67, …, -0.11]). The proximity of two embeddings in this space directly reflects the semantic similarity between their corresponding requirements.

C. Clustering Requirements in Groups

Using the embeddings generated in the previous step, requirements are grouped into clusters through Agglomerative Clustering, a hierarchical method that iteratively merges data points based on similarity [15]. This process relies on Euclidean Distance as the similarity metric, where smaller distances indicate higher semantic similarity between requirements [16], [29]. To ensure optimal clustering, Ward Linkage is employed, which minimizes intra-cluster variance by selecting merges that reduce the overall variance [28]. The variance adjustment is calculated as:

$$\Delta Variance = Variance\ of\ Cluster\ A + Variance\ of\ Cluster\ B - Variance\ of\ Combined\ Cluster$$

For better interpretability and functional modularity, the number of clusters (k) is predetermined based on the dataset’s size and complexity (e.g., k = 4 for smaller datasets). Each resulting cluster represents a distinct functional module, grouping semantically similar requirements while separating unrelated ones [18], [19]. For instance, requirements like *"System allow users log in"* and *"Users register access"* are grouped into Cluster 1, while others such as *"System calculate attendance"* and *"Users generate reports"* fall into separate clusters, reflecting unique functional distinctions. Examples of these clusters is shown in Table II.

TABLE II Clustering Output Example

Requirement	Cluster
Ensure image sliders work properly and link to the restaurant homepage.	1
Match UI with the Android version: icons, formats, sizing, and alignment.	1
Show current prices for discounted items and notify users about deletions.	2
Show discount value on item card and final total price box for % discount items.	2
Connect QR code feature and create deep links for reading the QR code.	3
Ensure every QR code generator has a deep link and forwards to the app store/play store if not installed.	3

D. Evaluation

The quality of clustering results is evaluated using several key metrics to ensure both intra-cluster cohesion and inter-cluster separation. Cohesion measures the semantic similarity of requirements within a cluster, ensuring that grouped requirements share strong relationships [26]. In contrast,

Separation evaluates the dissimilarity between clusters, ensuring functionally distinct requirements are placed in separate groups [20]. The overall clustering quality is assessed using the Silhouette Score, which combines both cohesion and separation into a single metric. Additionally, the Davies-Bouldin Index is used to measure intra-cluster compactness and inter-cluster separation, where lower values indicate better clustering performance [21]. Together, these metrics provide a comprehensive evaluation of clustering effectiveness.

E. Visualization

For interpretability, high-dimensional embeddings are reduced to two dimensions using Principal Component Analysis (PCA). This dimensionality reduction enables the generation of scatter plots where each point represents a requirement, and clusters are distinguished by color [14]. These visualizations provide insights into the clustering structure and relationships among requirements. The SQUIRE methodology integrates preprocessing, semantic embedding generation, clustering, evaluation, and visualization into a cohesive framework for automating requirements analysis. By leveraging SBERT embeddings and hierarchical clustering with a fixed number of clusters, it ensures accurate grouping of semantically similar requirements while enhancing modular design and traceability.

F. Algorithm: SQUIRE (Semantic QUIck Requirements Engineering)

Input:

- $R = \{R_1, R_2, \dots, R_n\}$: A set of n textual requirements.

Output:

- $C = \{C_1, C_2, \dots, C_k\}$: k clusters of semantically similar requirements.

Steps:

1. Preprocessing

Transform each requirement R_i into a standardized form:

- Convert to lowercase: R_i'= lower(R_i).
- Remove punctuation: R_i''= remove_punctuation(R_i').
- Tokenize text: W_i= tokenize(R_i'').
- Remove stopwords: $W_i' = W_i / Stopwords$

2. Embedding Generation

Generate dense vector embeddings $E = \{E_1, E_2, \dots, E_n\}$ using a pre-trained SBERT model:

$$E_i = SBERT(R_i')$$

Where $E_i \in \mathbb{R}^{384}$

3. Similarity Computation

Compute pairwise similarity between embeddings using the Euclidean distance.

$$d(E_i, E_j) = \sqrt{\sum_{k=1}^{384} (E_{i_k} - E_{j_k})^2}$$

where $E_{i_k} and\ E_{j_k}$ are the k^{th} dimensions of embeddings $E_i\ and\ E_j$.

4. Clustering

Perform hierarchical clustering using Agglomerative Clustering:

- Linkage method: Ward's linkage minimizes intra-cluster variance.
- Clustering criterion:

$$Merge\ clusters\ that\ minimize\ \Delta Variance$$

Result: Assign each requirement R_i to a cluster label C_i.

5. Evaluation

1. Cohesion: Measure intra-cluster similarity:

$$Cohesion = \frac{1}{|C_k|} \sum_{i,j \in C_k} Sim(E_i, E_j)$$

where $Sim(E_i, E_j) = 1 - d(E_i, E_j)$.

2. Separation: Measure inter-cluster dissimilarity:

$$Separation = \min_{i \in C_k, j \in C_l, k \neq l} d(E_i, E_j)$$

3. Silhouette Score: Combines cohesion and separation to measure the overall clustering quality:

$$Silhouette\ Score = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

where $a(i)$ is the average intra-cluster distance and $b(i)$ is the average nearest-cluster distance.

4. Davies-Bouldin Index: Quantifies intra-cluster compactness and inter-cluster separation:

$$DB\ Index = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \frac{\sigma_i + \sigma_j}{\mathrm{d}(C_i, C_j)}$$

Lower values indicate better clustering.

6. Complexity Analysis:

The complexity of different stages in the SQUIRE methodology is analyzed as follows:

1. Preprocessing Complexity (Time & Space):

Time Complexity: $O(n)$ where n is the number of requirements. Each requirement undergoes tokenization, stopword removal, and normalization, which operates linearly with respect to the number of requirements.

Space Complexity: $O(n)$, as each requirement is stored as a processed text sequence before embedding.

2. Embedding Generation Complexity (Time & Space):

Time Complexity: $O(n)$, since the Sentence-BERT (SBERT) model processes each requirement independently, leading to a linear complexity.

Space Complexity: $O(n \times d)$, where d is the embedding dimension (384 in the case of MiniLM-SBERT). The output matrix of embeddings requires storage proportional to the dataset size.

3. Clustering Complexity (Time & Space):

Time Complexity: $O(n^3)$ for Agglomerative Clustering, due to the hierarchical structure requiring pairwise distance computations and iterative merging.

4. Space Complexity: $O(n^2)$, as the clustering algorithm maintains a distance matrix for all requirement pairs.

This complexity analysis provided a clear distinction between time and space requirements at different stages of the methodology.

The SQUIRE algorithm efficiently organizes textual requirements into meaningful clusters by leveraging NLP embeddings and hierarchical clustering [28]. Its structured workflow ensures semantic precision and scalability while minimizing redundancy. With its foundation in advanced language models and practical clustering methods, SQUIRE lays the groundwork for a streamlined approach to handling complex requirements datasets, offering clarity and functionality to modern software engineering practices

IV. Results and Validation of Proposed Model

The validation of the proposed SQUIRE methodology was conducted using four real-world datasets sourced from a software company [32]. These datasets, representing diverse functional domains, included four different domains as shown in Table III. The primary objective of this validation was to assess the model's ability to accurately group semantically similar requirements into functional clusters.

TABLE III DATASETS OVERVIEW

Dataset	Domain	Number of Requirements
Attendance Management	Employee attendance tracking and reporting	26
E-commerce	Online shopping portal requirements	20
Lottery Management	Lottery system functionality	24
School Management	Educational institution operations	23

These datasets represent diverse functional requirements, providing a robust basis for testing the domain-agnostic capabilities of the SQUIRE methodology [32].

The evaluation focused on clustering outcomes, visualized through scatter plots and assessed quantitatively using Cohesion, Separation, Silhouette Score, and Davies-Bouldin Index. The results highlight the methodology's strengths and its performance across different datasets.

The clustering results for each dataset were visualized using Principal Component Analysis (PCA), reducing the 384-dimensional embeddings to two dimensions [23]. The scatter plots for the datasets Fig. 1 illustrate the semantic clusters, with each point representing a requirement and colors distinguishing the clusters.

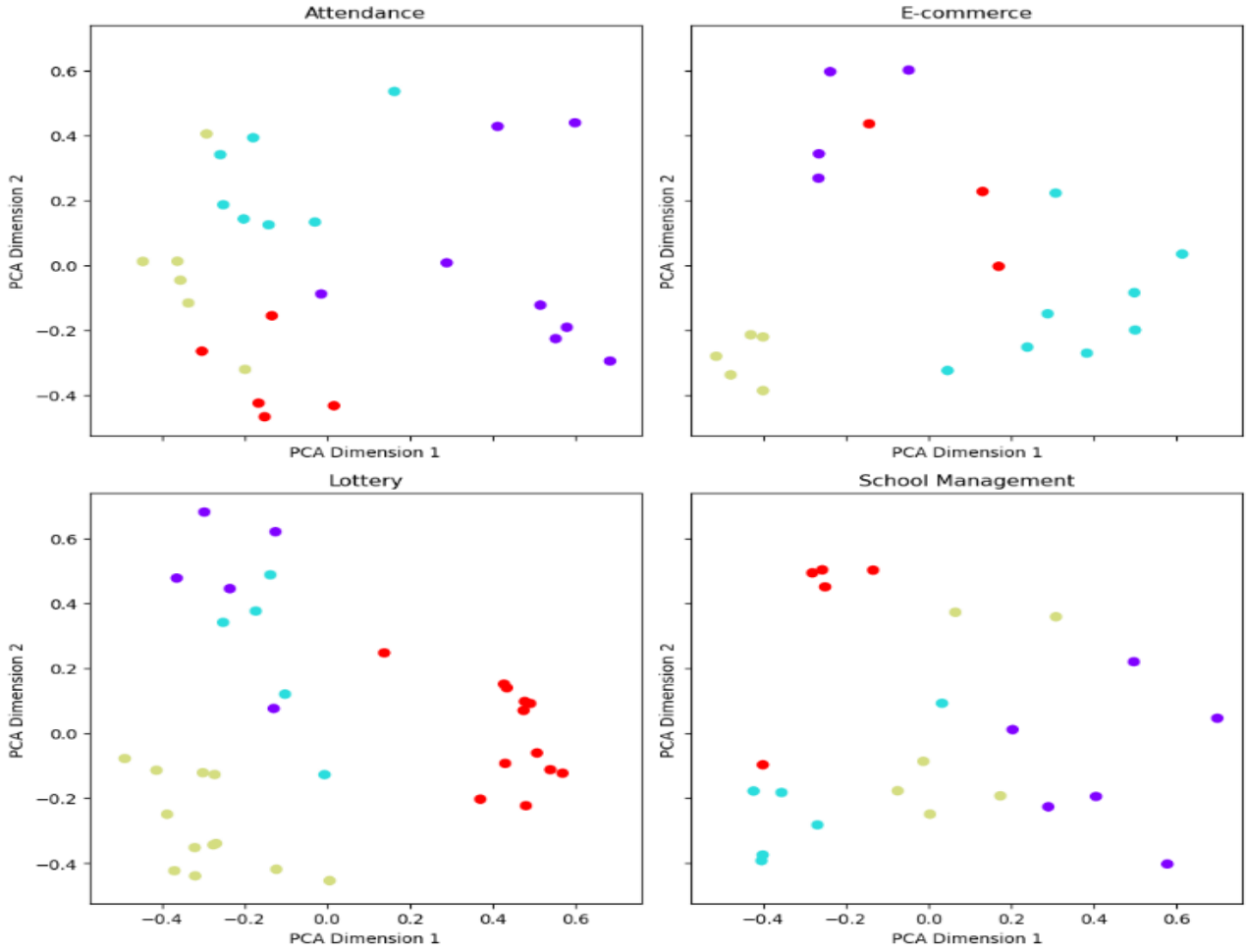


Fig. 1. Distribution of requirements clusters using PCA for each dataset.

The clustering performance is summarized in the Table IV below:

TABLE IV CLUSTERING METRICS FOR DIFFERENT DATASETS

Dataset	Cohesion	Separation	Silhouette Score	Davies-Bouldin Index
Attendance Management	0.2362	1.1306	0.1276	1.7027
E-commerce	0.1360	1.2911	0.1044	1.7195
Lottery Management	0.1096	1.2622	0.1587	1.8307
School Management	0.1940	1.1789	0.1174	1.7759

1) Cohesion: The Attendance Management dataset achieved the highest cohesion (0.2362), indicating well-grouped clusters. The Lottery Management dataset displayed slightly lower cohesion (0.1096), reflecting greater diversity within clusters.

2) Separation: The E-commerce dataset achieved the highest separation (1.2911), highlighting distinct clusters. The Attendance Management dataset had slightly lower separation (1.1306), possibly due to overlapping functional requirements.

3) Silhouette score: The Lottery Management dataset recorded the highest silhouette score (0.1587), indicating a good balance between cohesion and separation.

4) Davies-bouldin index: The Attendance Management dataset exhibited the lowest Davies-Bouldin Index (1.7027), reflecting compact and well-separated clusters. The Lottery Management dataset had the highest index (1.8307), suggesting room for improvement in cluster separation.

The validation results demonstrate the practical utility of the SQUIRE methodology in streamlining requirements engineering. By automating the clustering of semantically similar requirements, SQUIRE significantly reduces the manual effort required for organizing and analyzing requirements, allowing practitioners to focus on higher-value tasks such as decision-making and system design [7]. The methodology's ability to achieve high cohesion and separation across diverse datasets highlights its adaptability to different functional domains, making it a robust solution for handling large-scale and dynamic software projects. Furthermore, the integration of PCA-based visualizations enhances interpretability, providing clear insights into the relationships between requirements and supporting better traceability. Overall, SQUIRE offers a scalable and domain-agnostic approach that addresses critical challenges in requirements engineering, paving the way for more efficient and error-free software development processes.

V. DISCUSSION, LIMITATIONS AND FUTURE WORK

The validation results of the SQUIRE methodology demonstrate its effectiveness in clustering semantically similar requirements into distinct functional groups across diverse datasets. By leveraging Sentence-BERT embeddings and hierarchical clustering techniques, the methodology successfully addresses key challenges in requirements engineering, such as redundancy reduction, modularity, and enhanced traceability [2], [3], [13].

The methodology consistently produced meaningful clusters across four distinct datasets, sourced from a software company, representing domains Attendance Management, E-commerce, Lottery Management, and School Management [32]. These datasets varied in size and complexity, containing 20–26 requirements each, and provided a realistic foundation for testing the robustness and adaptability of the proposed approach. The clustering outputs revealed clear functional distinctions in well-structured domains like Attendance Management, while moderately overlapping clusters were observed in datasets like Lottery Management, which contained diverse and less structured requirements.

The evaluation metrics provided deeper insights into clustering performance:

Cohesion values indicated the strength of relationships within clusters, with the Attendance Management dataset achieving the highest cohesion, reflecting compact and meaningful clusters. Lower cohesion in the Lottery Management dataset suggests room for improvement in handling more heterogeneous requirements.

Separation metrics demonstrated the distinctiveness of clusters across datasets, with E-commerce showing the highest separation due to its well-defined functional boundaries.

Silhouette Score, a balance of cohesion and separation, highlighted the methodology's ability to achieve reasonable clustering quality across all datasets, with the highest score recorded for the Lottery Management dataset.

Davies-Bouldin Index values, indicative of clustering compactness and separation, were lowest for Attendance Management, reinforcing its strong cluster formations, while slightly higher values for Lottery Management reflected less compact clusters.

The visualizations further supported these findings, with distinct and well-separated clusters for structured datasets such as Attendance Management and School Management, while partially overlapping clusters were observed in Lottery Management due to functional overlaps in its requirements. The PCA-reduced scatter plots provide a clear representation of the semantic clustering process, aiding interpretability and further validating the methodology [14].

Despite the strong results, some limitations were observed. The clustering process relied on a fixed number of clusters, which may not always align with the inherent structure of the dataset. This could result in under- or over-clustering, especially in datasets with varied functional complexity. Additionally, the preprocessing pipeline, while robust, could be further enhanced with more domain-specific customizations, such as advanced synonym resolution or enhanced tokenization techniques, to address ambiguities in textual requirements. Finally, the methodology's reliance on static embeddings may limit its adaptability to rapidly evolving datasets, where requirements are frequently updated or redefined.

Future work will focus on enhancing the SQUIRE methodology by integrating advanced preprocessing techniques,

such as lemmatization and domain-specific synonym resolution, to improve the consistency and semantic accuracy of requirements. Adaptive clustering techniques, such as silhouette-based optimization, will be explored to dynamically determine the optimal number of clusters, ensuring better alignment with diverse datasets. Additionally, the use of more advanced models, such as GPT-based embeddings, will be investigated to capture deeper semantic relationships. A user-friendly tool incorporating real-time clustering, visualization, and traceability features will be developed to make the methodology more accessible to practitioners. Validation will be extended to real-world software engineering projects across various industries to evaluate practical applicability and scalability. Finally, interactive 3D visualizations and additional evaluation metrics will be introduced to improve interpretability and provide more comprehensive assessments of clustering quality.

VI. Conclusion

Requirements Engineering (RE) constitutes a pivotal phase in software development, focusing on the elicitation, definition, and management of stakeholder needs. Despite its criticality, traditional approaches frequently falter in managing the complexity, scale, and dynamism of contemporary software systems. Natural Language Processing (NLP) has emerged as a transformative enabler, offering automation in the analysis and organization of textual requirements. The SQUIRE framework, leveraging Sentence-BERT embeddings for semantic clustering, introduces a structured, scalable methodology for refining requirements management. By enhancing traceability, minimizing redundancy, and facilitating modular organization, SQUIRE addresses key inefficiencies in conventional RE practices. Its potential for broad applicability across diverse domains underscores its relevance to evolving software engineering demands. While SQUIRE has demonstrated efficacy, further refinements are necessary to optimize its alignment with stakeholder objectives and its adaptability to increasingly complex, dynamic requirements. Advancing the framework's flexibility and scalability will not only bridge theoretical innovations with practical application but also expand its impact across a wider spectrum of domains, establishing a robust foundation for next-generation RE methodologies.

References

[1] H. Villamizar, T. Escovedo and M. Kalinowski, "Requirements Engineering for Machine Learning: A Systematic Mapping Study," *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, Palermo, Italy, 2021, pp. 29-36, doi: 10.1109/SEAA53835.2021.00013.

[2] L. Karlsson, Å. G. Dahlstedt, and A. Persson, “Requirements engineering challenges inmarket-driven software development An interview study withpractitioners”, Inf and Soft Techn, vol. 49, Dec. 2007, pp.588-604, doi: 10.1016/j.infsof.2007.02.008.

[3] R. Izhar, Kenneth Cosh, “Enhancing Agile Software Development: A Novel Approach to Automated Requirements Prioritization”, 2024 21st International Joint Conference on Computer Science and Software Engineering, 2024.

[4] Sonbol, R., Rebdawi, G., and Ghneim, N. (2022). The use of nlp-based text representation techniques to support requirement engineering tasks: A systematic mapping review. IEEE Access.

[5] Pei, Z., Liu, L., Wang, C., and Wang, J. (2022). Requirements engineering for machine learning: A review and reflection. In 2022 IEEE 30th International Requirements Engineering Conference Workshops (REW), pages 166–175. IEEE.

[6] Ahanger, M.M.; Wani, M.A.; Palade, V. sBERT: Parameter-Efficient Transformer-Based Deep Learning Model for Scientific Literature Classification. *Knowledge* **2024**, *4*, 397-421. https://doi.org/10.3390/knowledge4030022.

[7] Sehrish Alam, Shahid N. Bhatti,” Impact and challenges of requirement engineering in agile methodologies: A systematic review”, International Journal of Advanced Computer Science and Applications, 2017. http://dx.doi.org/10.14569/IJACSA.2017.080455.

[8] Radwan, Aya, et al. “An Approach for Requirements Engineering Analysis Using Conceptual Mapping in Healthcare Domain.” International Journal of Advanced Computer Science and Applications, vol. 12, no. 8, 2021, https://doi.org/10.14569/ijacsa.2021.0120822.

[9] Eyal Salman, H.; Hammad, M.; Seriai, A.-D.; Al-Sbou, A. Semantic Clustering of Functional Requirements Using Agglomerative Hierarchical Clustering. *Information* **2018**, *9*, 222. https://doi.org/10.3390/info9090222.

[10] del Sagrado, J., del Águila, I.M. Assisted requirements selection by clustering. *Requirements Eng* **26**, 167–184 (2021). https://doi.org/10.1007/s00766-020-00341-1.

[11] Bakar, N. H., et al. (2014). Tochs requirements reuse: Identifying similar requirements with latent semantic analysis and clustering algorithms. *International Journal of Software Engineering and Its Applications*.

[12] Das, S., Deb, N., Cortesi, A. et al. Sentence Embedding Models for Similarity Detection of Software Requirements. SN COMPUT. SCI. 2, 69 (2021). https://doi.org/10.1007/s42979-020-00427-1.

[13] Elhassan, H. et al. (2022) ‘Requirements Engineering: Conflict Detection Automation Using Machine Learning’, Intelligent Automation & Soft Computing, 33(1), pp. 259–273. Available at: https://doi.org/10.32604/iasc.2022.023750.

[14] R. Izhar, K. Cosh and S. N. Bhatti, "Enhancing Agile Software Development: A Novel Approach to Automated Requirements Prioritization," 2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE), Phuket, Thailand, 2024, pp. 286-293, doi: 10.1109/JCSSE61278.2024.10613648.

[15] M. P. Naik, H. B. Prajapati and V. K. Dabhi, "A survey on semantic document clustering,"2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, 2015, pp. 1-10, doi: 10.1109/ICECCT.2015.7226036.

[16] Nagwani, N.K. Summarizing large text collection using topic modeling and clustering based on MapReduce framework. Journal of Big Data 2, 6 (2015). https://doi.org/10.1186/s40537-015-0020-5.

[17] Fougères, A.-J., & Ostrosi, E. (2020). Intelligent requirements engineering from natural language and their chaining toward CAD models. https://doi.org/10.48550/arXiv.2007.07825.

[18] Mehta, V., Agarwal, M. & Kaliyar, R.K. A comprehensive and analytical review of text clustering techniques. Int J Data Sci Anal 18, 239–258 (2024). https://doi.org/10.1007/s41060-024-00540-x.

[19] Haji, S.H., Jacksi, K., Salah, R.M. (2022). Systematic Review for Selecting Methods of Document Clustering on Semantic Similarity of Online Laboratories Repository. In: Daimi, K., Al Sadoon, A. (eds) Proceedings of the ICR’22 International Conference on Innovations in Computing Research. ICR 2022. Advances in Intelligent Systems and Computing, vol 1431. Springer, Cham. https://doi.org/10.1007/978-3-031-14054-9_23.

[20] Chen, D.; Wang, J. A Prompt Example Construction Method Based on Clustering and Semantic Similarity. Systems 2024, 12, 410. https://doi.org/10.3390/systems12100410.

[21] A. K. Singh, S. Mittal, P. Malhotra and Y. V. Srivastava, "Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means," *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2020, pp. 306-310, doi: 10.1109/ICCMC48092.2020.ICCMC-00057.

[22] Noor Hasrina Bakar, Zarinah M. Kasirun, Norsaremah Salleh, Feature extraction approaches from natural language requirements for reuse in software product lines: A systematic literature review, Journal of Systems

and Software, Volume 106, 2015, Pages 132-149, ISSN 0164-1212, https://doi.org/10.1016/j.jss.2015.05.006.

[23] R. Izhar, S. N. Bhatti, "Bridging Precision and Complexity: A Novel Machine Learning Approach for Ambiguity Detection in Software Requirements," in IEEE Access, vol. 13, pp. 12014-12031, 2025, doi: https://doi.org/10.1109/ACCESS.2025.3529943.

[24] A. Udomchaiporn, N. Prompoon and P. Kanongchaiyos, "Software Requirements Retrieval Using Use Case Terms and Structure Similarity Computation," 2006 13th Asia Pacific Software Engineering Conference (APSEC'06), Bangalore, India, 2006, pp. 113-120, doi: 10.1109/APSEC.2006.53.

[25] A. Radovanović, J. Li, J. V. Milanović, N. Milosavljević and R. Storchi, "Application of Agglomerative Hierarchical Clustering for Clustering of Time Series Data," 2020 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe), The Hague, Netherlands, 2020, pp. 640-644, doi: 10.1109/ISGT-Europe47291.2020.9248759.

[26] Jie He, Wanqiu Long, and Deyi Xiong. 2022. Evaluating Discourse Cohesion in Pre-trained Language Models. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 28–34, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

[27] Hazem Abdelazim, Mohamed Tharwat and Ammar Mohamed, "Semantic Embeddings for Arabic Retrieval Augmented Generation (ARAG)" International Journal of Advanced Computer Science and Applications(IJACSA), 14(11), 2023. http://dx.doi.org/10.14569/IJACSA.2023.01411135.

[28] Murtagh, F., Legendre, P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?. *J Classif* **31**, 274–295 (2014). https://doi.org/10.1007/s00357-014-9161-z.

[29] I. Dokmanic, R. Parhizkar, J. Ranieri and M. Vetterli, "Euclidean Distance Matrices: Essential theory, algorithms, and applications," in *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 12-30, Nov. 2015, doi: 10.1109/MSP.2015.2398954.

[30] Amato, Alberto and Di Lecce, Vincenzo. "Data preprocessing impact on machine learning algorithm performance" *Open Computer Science*, vol. 13, no. 1, 2023, pp. 20220278. https://doi.org/10.1515/comp-2022-0278.

[31] Bin Wang and C.-C. Jay Kuo. 2020. SBERT-WK: A Sentence Embedding Method by Dissecting BERT-Based Word Models. IEEE/ACM Trans. Audio, Speech and Lang. Proc. 28 (2020), 2146–2157. https://doi.org/10.1109/TASLP.2020.3008390.

[32] rahat-23. "GitHub - Rahat-23/Datasets-For-Requirements." *GitHub*, 2024, github.com/rahat-23/Datasets-for-Requirements.

Data-Driven Technology Augmented Reality Digitisation in Cultural Communication Design

A Case Study from Qinhuai Lantern Festival ICH

Na YIN*
School of International Education, Wuhan Business University, Wuhan 430056, Hubei, China

***Abstract*—The digitalisation of intangible cultural heritage and big data technology provide great potential for the development of intangible cultural heritage in low-carbon reform tourism, which not only increases the accuracy of AR digital design, but also contributes to the management and protection of intangible cultural heritage based on tourism. Aiming at the lack of testing and evaluation process of the current tourism intangible cultural heritage AR digital design process, this paper proposes an intangible cultural heritage AR digital design testing algorithm based on data-driven technology using Qinhuai lanterns and colours as a case study. Firstly, an AR digitization scheme based on Qinhuai lanterns intangible cultural heritage is designed; then, around the scheme, key technical contents of AR digitization design of intangible cultural heritage are analysed; secondly, combining the dragonfly algorithm with the restricted Boltzmann machine model, a test method for the AR digitization design of tourism intangible cultural heritage of low-carbon reform based on the optimization of the structural parameters of restricted Boltzmann machine by the dragonfly algorithm is put forward; lastly, relying on the collected data, the design of the AR digital design model of Qinhuai lanterns and colours tourism, and also analysed the effectiveness of the intelligent testing algorithm proposed in this paper. The results show that the proposed digital design method is effective, while the optimised test method has improved convergence speed and increased accuracy, and the test score prediction accuracy reaches 93.5%.**

Keywords*—*Intangible cultural heritage AR digitization; low-carbon tourism; design test analysis; dragonfly algorithm; restricted Boltzmann machine model

I. Introduction

In recent years, with the rapid development of digital information technology, the use of information technology and intelligent algorithm technology to identify, store, display and disseminate the intangible cultural heritage (ICH) has become an important way of intangible heritage heritage inheritance and protection, which makes the digitisation of intangible heritage a hotspot of concern in the field industry and academia [1]. As a unique cultural phenomenon and spiritual wealth, intangible culture not only meets people's spiritual needs, but also enhances the attractiveness of regional tourism [2]. With the rise of Augmented Reality (AR) technology, the digitisation of ICH has made a qualitative breakthrough and development, which provides new ideas for the digital design and display of intangible genetic culture [3]. At the same time, the development of big data technology, cloud computing technology, and intelligent algorithms has provided strong support for the collection, collation and analysis of intangible heritage [4]. Currently, with the combined application of AR technology and various types of ICH, the focus of the research falls on the AR digital design process, the lack of research on the testing method of the AR digital design effect of ICH, and at the same time, taking into account the factor of low-carbon tourism, the AR digital design algorithms of intangible culture for low-carbon tourism are even more scarce [5]. Therefore, the research of AR technology and data-driven algorithms in the digital design and analysis method of ICH in low-carbon tourism has become the development trend of theoretical research and realistic problem-solving research.

Currently, the digital research of ICH mainly includes ICH digital acquisition technology, storage technology, production technology, dissemination technology, and design and analysis technology [1]. Bai and Boo propose a digital exhibition prototype system based on mixed reality technology for the digitalisation of physical museum exhibitions [6]. Tu and Jiang combine immersive virtual reality technology and proposes a digital experience method for museums [7]. Gireesh investigates 3D digital intangible genetic dissemination methods [8]. Yingji explores the application of AR technology in the digital design and display of intangible genetic culture [9]. Li et al. [10] studied an online museum roaming system based on virtual genetic digitisation technology. By analysing the existing literature research, the following problems exist in the digitization technology of intangible genetic culture based on AR technology: 1) the digitization of intangible genetic culture lacks the research of specific detail technology as well as the performance analysis; 2) the research of intangible genetic culture digitization based on AR technology lacks the research of quantitative design analysis and evaluation system. With the rapid development of artificial intelligence algorithms, the use of high-performance algorithms to solve the AR digitisation design analysis and testing problems has become one of the future directions for the application of AR technology in the development of intangible genetic culture. Considering that the digitisation of ICH is mainly to improve the dissemination of ICH as well as the development of its carriers to promote, that is, to adapt to the current low-carbon reform and tourism ICH promotion of digital applications [11].

Facing the problems existing in the current AR digital design of ICH as well as considering factors such as low-carbonisation tourism, this paper designs a design analysis and testing algorithm for assessing the AR digital design scheme. Around the consideration of low-carbon reform tourism ICH AR digital design process, combined with specific cases, through the

*Corresponding Author

analysis of AR digital design key technologies, using the restricted Boltzmann machine model [12] and heuristic optimization algorithms, designed a dragonfly algorithm [13] optimization restricted Boltzmann machine tourism ICH AR digital design test analysis algorithm.

This paper is composed of the following sections: Section II reviews the relevant research progress, reviews the existing AR and ICH digital research, and identifies gaps in technical details, performance analysis, and quantitative evaluation systems. In Section III: the research proposes an AR digitisation framework for Qinhuai Lantern Festival, which combines the dragonfly algorithm (DA) with the restricted Boltzmann machine (RBM) for optimisation. Key technologies such as AR interface design, data collection and model optimisation are analysed to improve the accuracy and efficiency of AR digitisation. In Section IV and Section V is the application part: through experiments comparing multiple algorithms, the DA-RBM model demonstrated superior performance, with a prediction accuracy of 93.5%, faster convergence and stronger testing capabilities. The paper concludes in Section VI by emphasizing the role of AR and intelligent algorithms in advancing cultural heritage protection and low-carbon tourism, demonstrating the effectiveness of the proposed DA-RBM model.

II. AR Digital Design Ideas

A. Background of the Study

In order to verify and analyse the effectiveness and superiority of the analysis algorithm of tourism intangible cultural heritage AR digitization design test analysis, this paper combines the case of intangible cultural heritage digitization of Qinhuai lanterns with the case of intangible cultural heritage digitization of Qinhuai lanterns by designing AR digitization scheme of intangible cultural heritage of intangible cultural heritage of Qinhuai lanterns, analyzing the key technologies, constructing the set of AR digitization design test indexes, and establishing the mapping relationship.

The Qinhuai Lantern Festival is one of China's traditional folk art forms, originating in the Qinhuai River Valley in Nanjing, with a long history and deep cultural heritage. It is not only known as "the first lantern festival in the world", but also famous for its grand scale and large number of participants. There are many types of lanterns, including lanterns, palace lanterns, and large lotus lanterns, many of which are used for children's games and adults' enjoyment, and have the nature of toys and handicrafts [14]. In 2008, Qinhuai lanterns were selected as part of the national intangible cultural heritage catalogue.

Qinhuai lanterns have the following characteristics [14]:

1) the shapes are mostly originated from nature;

2) the colours are characteristic of Jiangnan, mainly in the five basic tones of red, yellow, blue, green, and white;

3) they cover a variety of techniques such as paper-cutting, painting, carving, and papier-mâché.

The digital dissemination of Qinhuai lanterns is not only for the protection and inheritance of cultural heritage, but also for the diversification of culture to expand the means of dissemination, but also to inject new vitality into the traditional arts [15].

In order to improve the efficiency of digital design of Qinhuai lanterns and colours, this project obtains the digital design needs of Qinhuai lanterns and colours through questionnaires, field visits and user interviews. To summarize the digital design requirements of Qinhuai lanterns and colours, they include [16]:

- Improve the digital awareness of Qinhuai lanterns and colours AR;
- Diversify the digital functions of Qinhuai lanterns and colours AR;
- Enhance the digital stability of Qinhuai lanterns and colours AR;
- Improve the demand for digital social interaction and personalisation; and
- Increase the knowledge and fun, as display in Fig. 1.

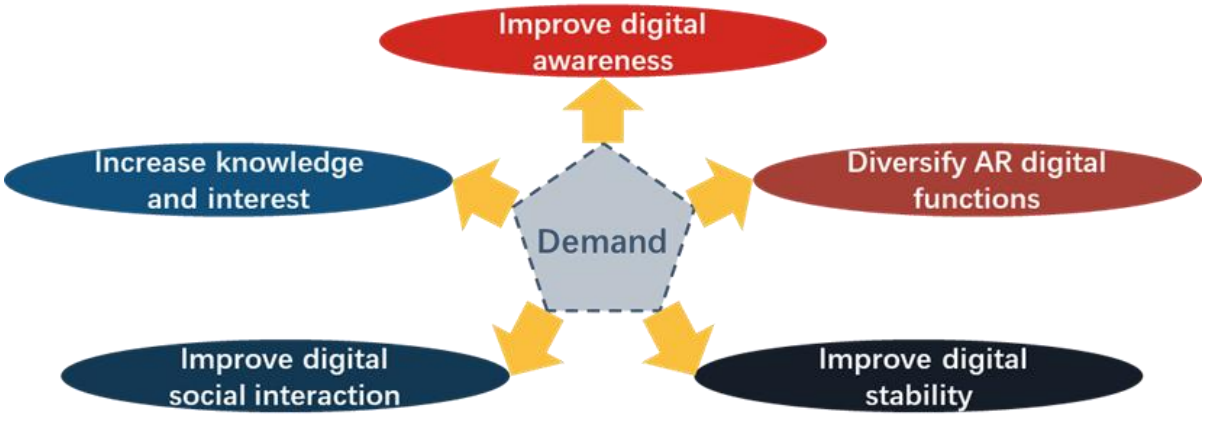


Fig. 1. Demand characterisation.

B. Design Programme

According to the digital design principles of real integrity, live scalability, and interactive experience [17], this paper designs a digital design and test and evaluation scheme for Qinhuai lantern AR that combines multiple technologies, as shown in Fig. 2.

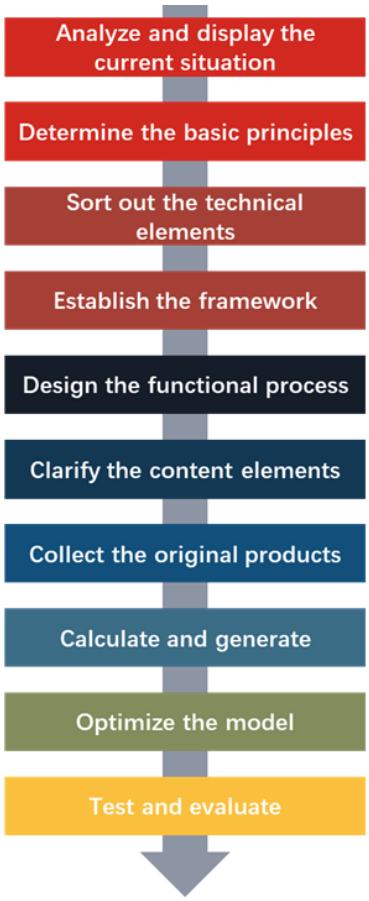


Fig. 2. General idea of the design.

From Fig. 2, the Qinhuai lantern AR digital design and testing and evaluation scheme is divided into three phases, such as pre-preparation, mid-phase, and post-phase, including the phases of analysing the current situation of the display, determining the basic principles, sorting out the technological elements, establishing the framework, designing the functional flow, clarifying the content elements, acquiring the original artefacts, generating the calculation, optimising the model, and testing and evaluating it [18].

In the Qinhuai lantern AR digital design and test and evaluation programme, based on the analysis of the current situation of the Qinhuai lantern display and its needs, the basic principles of the design are determined, and at the same time, the key technical elements of the digital display design are sorted out, and the digital display platform is structurally and functionally designed, the elements of the content are clarified, and the digital acquisition technology, storage technology, production technology, dissemination technology, and evaluation technology are used to complete the ICH. Digital display and effect analysis.

III. Analysis of Key Technologies for AR Digital Design

According to the flow of the design scheme, this section analyses the key technologies in terms of AR digital application interface design, data acquisition, digital model optimisation, digital design implementation and digital design testing, and the specific key technologies are shown in Fig. 3.

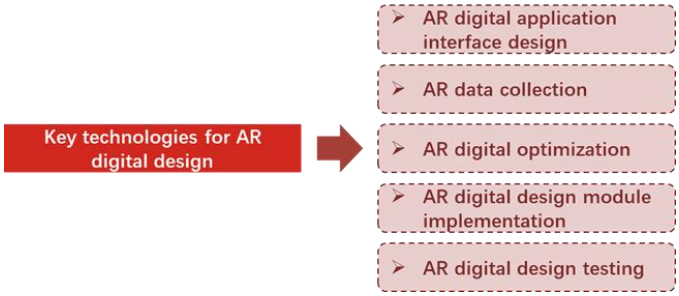


Fig. 3. Key technologies.

A. AR Digital Application Interface Design Techniques

In order to improve the satisfaction of the user experience, this paper divides the Qinhuai lantern AR application into three parts, such as discovery, AR display interface, and production techniques [19], and the specific information structure design is shown in Fig. 4. The discovery part is mainly used to provide users with information about the pavilions, activities, and works of the Qinhuai lanterns, and the display interface part is the entrance to the AR digital experience; the production techniques part is mainly used to provide users with the lantern production process.

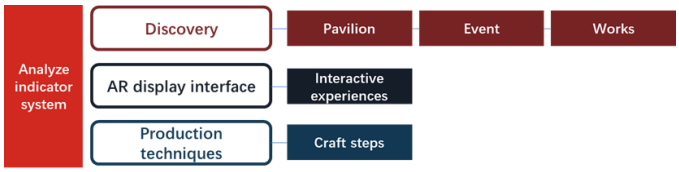


Fig. 4. AR digital information architecture design.

AR digital application interface design is divided into two phases, i.e., low-fidelity prototyping and high-fidelity prototyping [19]. Low-fidelity prototyping is generally at the initial stage of the design process and uses Figma design tools to quickly design concepts and interaction flows. High-fidelity prototype diagram design is based on low-fidelity prototype diagram design, with the help of Figma, illustrator and other design software, to carry out high-fidelity prototype diagram design and production. The high-fidelity prototype diagram design is generally designed from three aspects, such as discovery, AR display interface, and production techniques. For the AR digitisation of Qinhuai lanterns, the discovery page of this project is divided into three secondary pages, such as "Pavilion", "Activities" and "Works", as shown in Fig. 5; the AR display interface is mainly to serve the users with the AR display interface. The AR display interface mainly serves the AR interactive experience with users, as shown in Fig. 6; the production techniques are mainly selected from the perspective of universality and representativeness of the production steps of the lanterns and colours.

Fig. 5. Discovery page design.

Fig. 6. AR display interface.

B. AR Data Acquisition Techniques

The AR data acquisition technique is based on 3D scanning and modelling of Qinhuai lanterns through RealityCapture software (RC). The operation process based on RC software is shown in Fig. 7, including importing images, aligning images, calculating models, colouring, texturing, and exporting [20].

C. AR Digital Optimisation Techniques

In order to reduce the model size, reduce the rendering time, and maintain the original surface details and quality, the AR digital model optimisation method was used to process the model after AR data acquisition. The AR digital model optimisation method includes mesh simplification, material and texture optimisation, removal of hidden or redundant geometry, and merging of meshes and materials [21], as shown in Fig. 8.

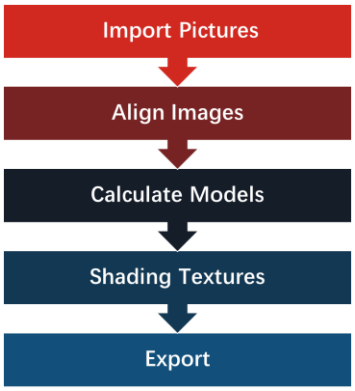


Fig. 7. RC software operation flow.

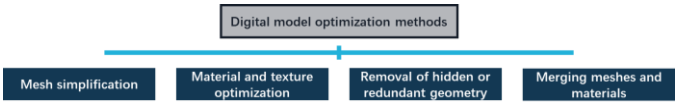


Fig. 8. AR digital optimisation method.

D. AR Digital Design Module Implementation Techniques

This project uses unity engine [22] for AR digital design module implementation.AR digital design module is implemented using SDK steps include importing Unity project, model shading, rendering model, target recognition and tracking.

E. AR Digital Design Testing Technology

AR digital design testing is mainly used for digital design performance analysis and effect evaluation [23].AR digital design testing technology tests and analyses the usability and stability of the AR digital design process by constructing and analysing the mapping relationship between AR digital design test index values and test evaluation levels, adopting certain fitting algorithms, and learning the training data, as shown in Fig. 9. This project uses an intelligent optimisation algorithm combined with a restricted Boltzmann machine.

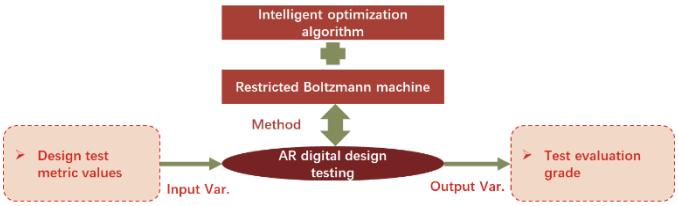


Fig. 9. AR digital design test methodology.

IV. AR Digital Design Testing

In order to analyse the AR digital quantification of Qinhuai lanterns and increase the accuracy of AR digital design test, this paper adopts the dragonfly algorithm to optimize the restricted Boltzmann mechanism to build the AR digital design test model.

A. AR Digital Design Test Models

1) AR digital design test metrics: In order to effectively analyse and assess the advantages and disadvantages of AR digital design solutions, based on the principles of systematic, objective, operable and other AR digital design test index selection, from the four aspects of the AR digital application interface A, acquisition B, model optimization C, design implementation D, to construct the AR digital design based on the discovery of A1, the AR display interface A2, the production techniques A3, aligning images B1, computing models B2, colouring B3, Texture B4, Mesh Simplification C1, Material and Texture Optimisation C2, Removal of Hidden or Redundant Geometry C3, Merging Mesh and Material C4, Model Colouring D1, Rendering Model D2, Target Recognition and Tracking D3, and other 14 test metrics sets [24], as shown in Fig. 10.

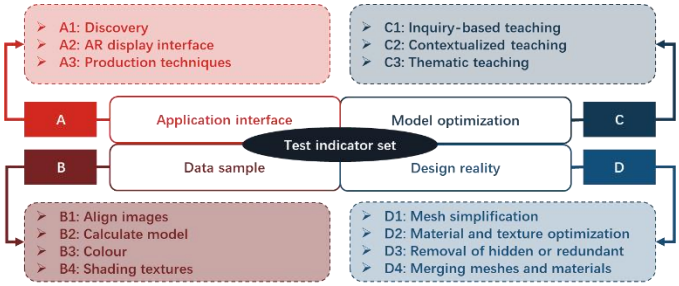


Fig. 10. Test indicator.

2) AR digital design test values and grades: In order to determine the quantitative AR digital design test assessment values and identify the test assessment levels, this paper divides the test assessment levels into six levels [24], and the specific relationship between each level and the corresponding test assessment value is shown in Table I.

TABLE I. CORRESPONDENCE BETWEEN TEST VALUE AND GRADE

No.	*Rank levels*	*Test scores*
1	Worse design	[0, 3]
2	Bad design	(3,6]
3	Fair design	(6,8]
4	Good design	(8, 10]
5	Better design	(10,12]
6	Great design	(12, 14]

3) Mapping relationships: For the Qinhuai lantern colours ICH, the AR digital design test model mainly constructs the mapping relationship between the AR digital design test indicators and the AR digital design test values as well as the grades, and the specific mapping calculation formula is as follows:

$$Y_{score} = f_{indicator-score}\left(X_{indicator}\right) \tag{1}$$

$$Y_{rank} = f_{ranklevel}\left(X_{indicator}\right) \tag{2}$$

Where $X_{indicator}$ denotes the AR digital design test metrics, $f_{indicator-score}$ denotes the mapping between the metrics and the test assessment values, Y_{score} denotes the AR digital design test values, $f_{ranklevel}$ denotes the mapping between the metrics and the test ratings, and Y_{rank} denotes the test ratings.

According to the principle of the AR digital design test model of Qinhuai lantern colour ICH, the mapping relationship between the AR digital design test indexes and the AR digital design test values as well as the grades is constructed as shown in Fig. 11.

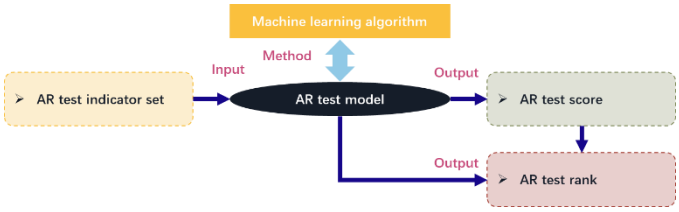


Fig. 11. AR digitisation test mapping relationship.

B. Intelligent Analysis Methods

Aiming at the problem of AR digital design testing of Qinhuai lanterns and ICH, this paper combines the constrained Boltzmann machine with the dragonfly algorithm, and proposes an intelligent analysis method for AR digital design testing based on the dragonfly algorithm to optimise the constrained Boltzmann machine.

1) Restricted boltzmann machines: Restricted Boltzmann Machine (RBM) [25] is an artificial neural network based on an energy model that consists of two layers of neurons: a visible layer and a hidden layer. The neurons between these two layers are fully connected, but there are no connections between neurons within the same layer. The RBM can be used for feature learning and generative model training, and is particularly adept at capturing high-level features of the data.

a) RBM probability distribution: RBM is a probability distribution model based on energy. Given the state vectors h and v, the current energy function of RBM is represented as follows:

$$E\left(v,h\right) = -a^T v - b^T h - h^T W v \tag{3}$$

where a and b denote the vector of bias coefficients and W denotes the weight matrix.

Define the probability distribution of the RBM in conjunction with the energy function:

$$P\left(v,h\right) = \frac{1}{Z} e^{-E(v,h)} \tag{4}$$

where Z is the normalisation factor.

$$Z = \sum_{v,h} e^{-E(v,h)} \tag{5}$$

According to the probability distribution, the conditional distribution is expressed as follows:

$$P\left(h|v\right) = \frac{P\left(h,v\right)}{P\left(v\right)} = \frac{1}{Z'} \prod_{j=1}^{n_h} \exp\left\{b_j^T h_j + h_j^T W_j; v\right\} \tag{6}$$

where Z' is the new normalisation factor.

$$\frac{1}{Z'} = \frac{1}{P\left(v\right)} \frac{1}{Z} \exp\left\{a^T v\right\} \tag{7}$$

The sigmoid activation function is used in the RBM from the visible layer to the hidden layer:

$$P\left(v_j=1|h\right)=sigmoid\left(a_j+W_{:,j}^{T}h\right) \tag{8}$$

b) RBM model loss function and optimisation: In order to solve for the parameters W , a and b , for m samples of the training set, the RBM generally uses a logarithmic loss function, i.e., it is expected to minimise the following equation:

$$L\left(W,a,b\right)=-\sum_{i=1}^{m}\ln\left(P\left(V^{(i)}\right)\right) \tag{9}$$

where $V^{(i)}$ denotes the ith particular training sample.

The gradient derivation results for the parameters W , a and b are as follows:

$$\frac{\partial\left(-\ln\left(P\left(V\right)\right)\right)}{\partial a_i}=\sum_{v}P\left(v\right)v_i-V_i \tag{10}$$

$$\frac{\partial\left(-\ln\left(P\left(V\right)\right)\right)}{\partial b_i}=\sum_{v}P\left(v\right)P\left(h_i=1|v\right)-P\left(h_i=1|V\right) \tag{11}$$

$$\frac{\partial\left(-\ln\left(P\left(V\right)\right)\right)}{\partial W_{ij}}=\sum_{v}P\left(v\right)P\left(h_i=1|v\right)v_j-P\left(h_i=1|V\right)V_j \tag{12}$$

RBM has a wide range of applications in a variety of machine learning tasks, including dimensionality reduction, feature learning, classification: for training classifiers, collaborative filtering, and generating models.

c) DA algorithm: Dragonfly Algorithm (DA) [26] is a novel intelligent optimisation algorithm proposed by Seyedali Mirjalili in 2016. The algorithm is inspired by the static and dynamic group behaviours of dragonflies in nature, in particular the collective intelligence they display in finding food and avoiding predators. The dragonfly algorithm is suitable for solving complex optimisation problems due to its high optimisation seeking capability and ease of implementation

The core of the dragonfly algorithm lies in modelling the group behaviour of dragonflies, including five behavioural styles such as separating, queuing, allying, searching for food and avoiding natural enemies. These behaviours are abstracted through mathematical models, forming the basic framework of the algorithm. In the algorithm, each individual dragonfly represents a potential solution, and the optimisation problem is explored and solved by simulating these behaviours.

- Separation behaviour: Neighbouring dragonflies are separated from each other and kept at a distance to avoid collisions:

$$S_i=-\sum_{j=1}^{N}\left(X-X_j\right) \tag{13}$$

where S_i denotes the value of the separation behaviour of the i^{th} dragonfly, and X_j denotes the individual position of the jth dragonfly.

- Queuing behaviour: Individual dragonflies maintain the same speed as other dragonflies in their neighbourhood by controlling their speed and direction, allowing populations to migrate in the same direction:

$$A_i=\frac{\sum_{j=1}^{N}\Delta X_j}{N} \tag{14}$$

Where A_i denotes the amount of queuing behaviour, ΔX denotes individual speed and N denotes the number of dragonflies in the neighbourhood.

- Allied behaviour: Individual dragonflies and neighbouring conspecifics converge towards the centre of the surrounding group:

$$C_i=\frac{\sum_{j=1}^{N}X_j}{N}-X \tag{15}$$

where A_i denotes the amount of aligned behaviour.

- Food-seeking behaviour: Individual dragonflies seek out and approach food in order to survive:

$$F_i=X^{+}-X \tag{16}$$

where F_i denotes the amount of foraging behaviour and X^{+} denotes the location of food.

- Avoidance of natural enemy behaviour: Individual dragonflies have instinctive vigilance and behaviour away from natural enemies:

$$E_i=X^{-}+X \tag{17}$$

Where E_i denotes the amount of enemy avoidance behaviour and X^{-} denotes the location of natural enemies.

When there are other dragonflies in the neighbourhood of an individual dragonfly, the step vector is calculated as follows:

$$\Delta X_{t+1}=\left(sS_i+aA_i+cC_i+fF_i+eE_i\right)+\omega\Delta X_t \tag{18}$$

Where t denotes the current iteration number, s , a , c , f and e denote the separation, formation, focusing, foraging and enemy avoidance behavioural weights respectively, and ω denotes the inertia weights.

The updated formula for the position of the next generation of dragonfly individuals is:

$$X_{t+1} = X_t + \Delta X_{t+1} \tag{19}$$

When there are no other dragonflies in the neighbourhood of a dragonfly individual, the current dragonfly is unable to update its own position through the information of other individuals in the neighbourhood. In order to ensure that the algorithm better explores the search space, the dragonfly individual position is calculated as follows using the Lévy flight random wandering strategy:

$$X_{t+1} = X_t + Levy(d) \times X_t \tag{20}$$

$$Levy(d) = 0.01 \times \frac{r_1 \times \sigma}{|r_2|^{\frac{1}{\beta}}} \tag{21}$$

$$\sigma = \left(\frac{\Gamma(1+\beta) \cdot \sin\frac{\pi\beta}{2}}{\Gamma\left(\frac{1+\beta}{2}\right) \cdot \beta \cdot 2^{\left(\frac{\beta-1}{2}\right)}} \right)^{\frac{1}{\beta}} \tag{22}$$

where $Levy(d)$ denotes the Levy flight strategy and $\Gamma(x) = (x-1)!$.

The formula for calculating the radius of the dragonfly neighbourhood is as follows:

$$r = \frac{ub - lb}{4} + 2(ub - lb)\frac{t}{t_{\max}} \tag{23}$$

Where ub and lb denote the upper and lower limits of the search space, t and $t_{\max}$ are the current iteration number and the maximum iteration number. The radius of the dragonfly neighbourhood increases with the number of iterations. When the number of iterations increases for a certain number of times, all dragonflies become individuals in the neighbourhood, updating their positions and step sizes through the information of all other dragonflies, which will eventually converge.

d) Constrained Boltzmann machine models incorporating DA algorithms: In order to increase the accuracy of the testing algorithm, this paper uses the dragonfly algorithm to optimise the RBM model, i.e., the dragonfly algorithm is used to find a set of RBM parameters to minimise the testing error.The decision variables of the DA-RBM algorithm are the parameters W , a , and b , and the RMSE value is used as the fitness evaluation function of the dragonfly algorithm:

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(\hat{y}_i - y_i)^2} \tag{24}$$

Where, x is the data sample, c_i denotes the i^{th} clustering centre and d denotes the dimension of the data sample.

The flow of the DA algorithm to optimise the RBM method is displayed in Fig.12 with the following steps:

- Step 1: Initialise parameters, including population size, maximum number of iterations, etc;
- Step 2: Based on the structural parameters of the RBM model, generate an initial population, i.e., dragonfly individuals randomly distributed in the search space;
- Step 3: Assess the fitness of each individual dragonfly, i.e. the quality of its problem solving;
- Step 4: Determine whether there are other individuals in the dragonfly's neighbourhood, if so, update the individual's position by using separation, queuing, alliance, food searching and natural enemy avoidance behaviours, otherwise update the individual's position by using the Levy flight strategy;
- Step 5: Repeat the above steps until the maximum number of iterations, output the optimal RBM model structure parameters, and construct the RBM model based on the DA algorithm.

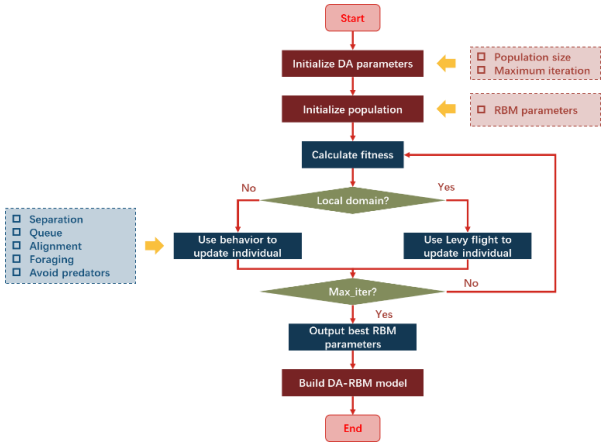


Fig. 12. Flowchart of DA-RBM algorithm.

e) Design test intelligent analysis model construction process: In order to construct an intelligent analysis model based on the AR digital design test of Qinhuai lanterns and colours, this paper adopts the DA-RBM model, by analysing the AR digital design test indexes and test values, taking the parameters of RBM W , a and b as the optimization variables, and taking the RMSE value as the fitness value function, and using DA algorithm optimization strategy to search for the optimal parameters of RBM W^* , a^* and b^* . The principle of the application of the DA-RBM model and the structure of the framework are shown in Fig. 13.

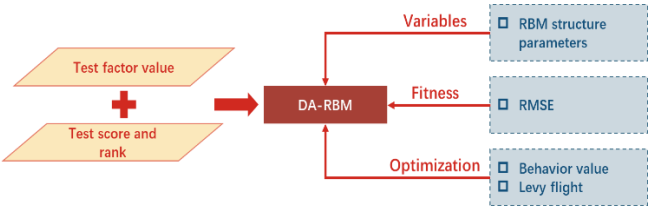


Fig. 13. DA-RBM application analysis.

According to the application of DA-RBM model in the AR digitisation problem of Qinhuai lantern colour low-carbon tourism ICH, the specific flow of the design test intelligent analysis method based on DA-RBM model is shown in Fig.14. As can be seen from Fig. 14, by analysing the key technology of AR digitization of Qinhuai lanterns and low-carbon tourism ICH, determining the AR digitization design test indexes, test scores, and grades, and after the data preprocessing technology, obtaining the training and testing sample set, combining with the intelligent optimization algorithm proposed in this paper to improve the machine learning algorithm, to achieve the construction and optimization of AR digitization design test model of AR digitization of ICH of low-carbon tourism of Qinhuai lanterns and colours.

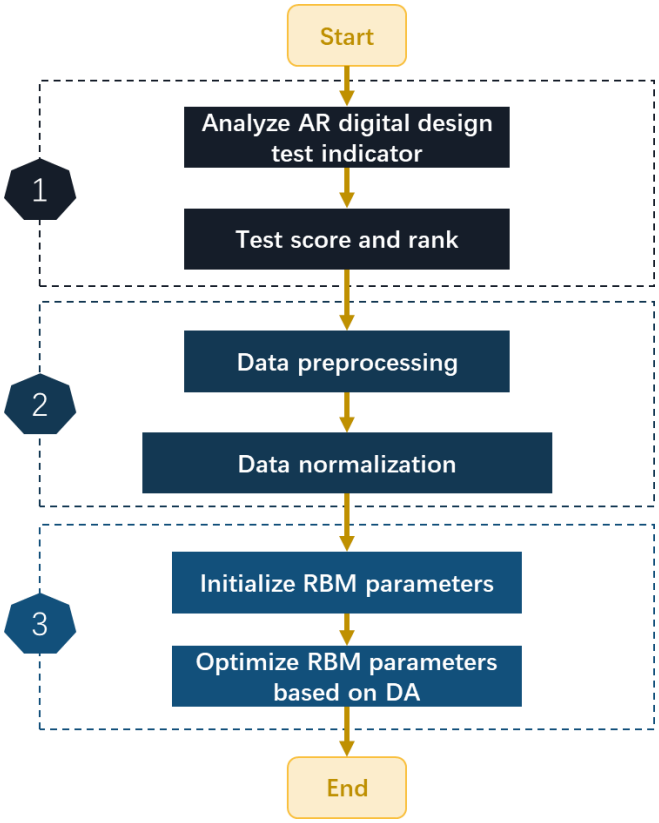


Fig. 14. Building intelligent analysis model for design test.

V. Analysis of Test Results

A. Test Environmental Setup

The experimental simulation environment is Windows 10, CPU is 2.80GHz, 8GB RAM. RealityCapture software is used for digitisation of Qinhuai lanterns, Vuforia SDK is used for functional implementation of AR, and Python 3.7 is used as the programming language for ICH AR digital design and testing algorithm.

B. Contrast Algorithm Parameter Settings

In order to verify the feasibility and efficiency of the ICH AR digital design and testing algorithms proposed in this paper, this paper takes the data collected during the AR digital design process of Qinhuai lanterns and colours as the analysis data, and six design testing algorithms are selected for comparison, as shown in Table II.

TABLE II. Comparison Algorithm Parameter Settings

Projects	*Parameter settings*
RBM	The number of hidden layer nodes is 150
PSO-RBM [27]	V_{max}=30, V_{min}=-30, r=0.5
GSA-RBM [28]	Alpha is 20, G0 is 100, Rnorm=2, Rpower=1
TLBO-RBM [29]	Tr=round(1+rand)
GWO-RBM [30]	a decreases linearly from 2 to 0
DA-RBM	Inertia weights are [0.4,0.9], separation, queuing, alliance, foraging, and enemy avoidance are 0.1, 0.1, 0.7, 1, and 1 respectively

As can be seen from Table II, the comparison algorithm parameters are mainly divided into two categories. For the RBM model parameters, the number of hidden layer nodes is set to 150; for the intelligent optimisation algorithm parameters, the maximum number of iterations is set to 100 and the population size is 50.

C. Analysis of AR Digital Realisation

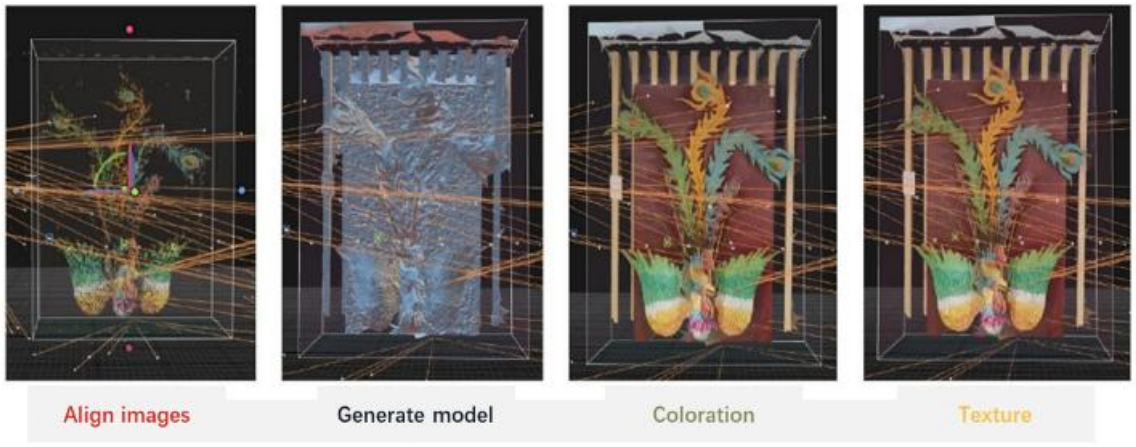
In order to verify the effectiveness of the AR digitisation design method, this paper adopts RealityCapture software and Vuforia SDK to implement the AR digitisation of Qinhuai lanterns and colours, and the specific analysis results are shown in Fig. 15 and Fig. 16.

From Fig. 15, the AR digital acquisition of Qinhuai lanterns is done through the process of aligning images, generating models, colouring and texturing to get the AR digital preliminary model.

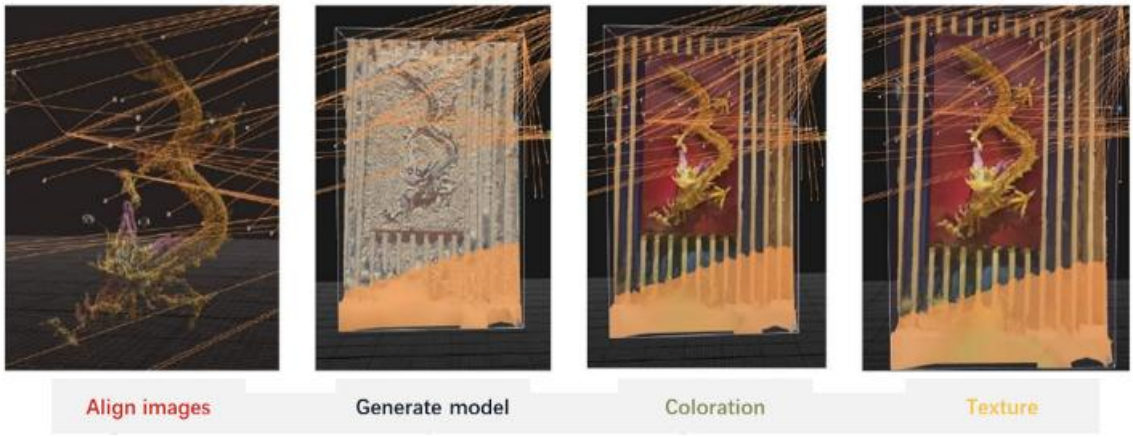
D. Test Performance Analysis

In order to validate the feasibility and effectiveness of the AR digitisation design test method based on the DA-RBM model, this paper compares the performance of the RBM, PSO-RBM, GSA-RBM, TLBO-RBM, GWO-RBM, and DA-RBM methods by using Qinhuai Colourful Lanterns AR digitisation process dataset.

From Fig. 16, we can see that the RBM optimisation based on the DA algorithm has a higher convergence accuracy than the other algorithms and has a better convergence speed than the other algorithms; the DA-RBM starts to converge at 20 iterations and converges to near 0.



(a) Intentions of the digital acquisition process for the Phoenix Lights model



(b) Intentions of the digital capture process for the dragon lantern model

Fig. 15. Digital acquisition process of Qinhuai lantern colours.

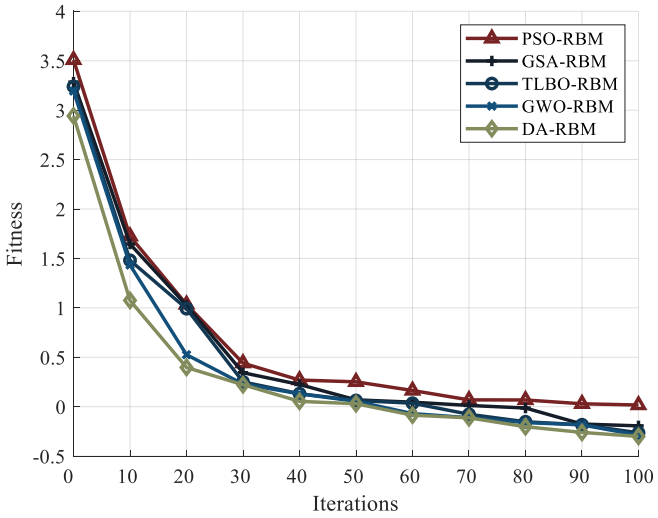


Fig. 16. AR digital design test method based on intelligent optimisation algorithm.

The test prediction scores for the RBM, PSO-RBM, GSA-RBM, TLBO-RBM, GWO-RBM, and DA-RBM methods, as well as the test rating results, are given in Figure 16. As can be seen from Fig. 17, the test prediction scores of the AR digital design test algorithm based on the DA-RBM model are closest to the true scores, and the test prediction errors are less than those of RBM, PSO-RBM, GSA-RBM, TLBO-RBM, and GWO-RBM.

The statistical results of RBM, PSO-RBM, GSA-RBM, TLBO-RBM, GWO-RBM, and DA-RBM method tests are given in Table III. In terms of RMSE, the prediction error of the DA-RBM-based AR digital design testing algorithm is the smallest, reaching 0.0178; in terms of accuracy, the test score prediction accuracy of the DA-RBM-based AR digital design testing algorithm is the smallest, reaching 93.5%; in terms of optimisation convergence time, the DA-RBM-based AR digital design testing algorithm has the shortest optimisation convergence time of is 4.40s; in terms of testing time, the DA-RBM-based AR digital design testing algorithm has the smallest test sample prediction time of 0.48s.

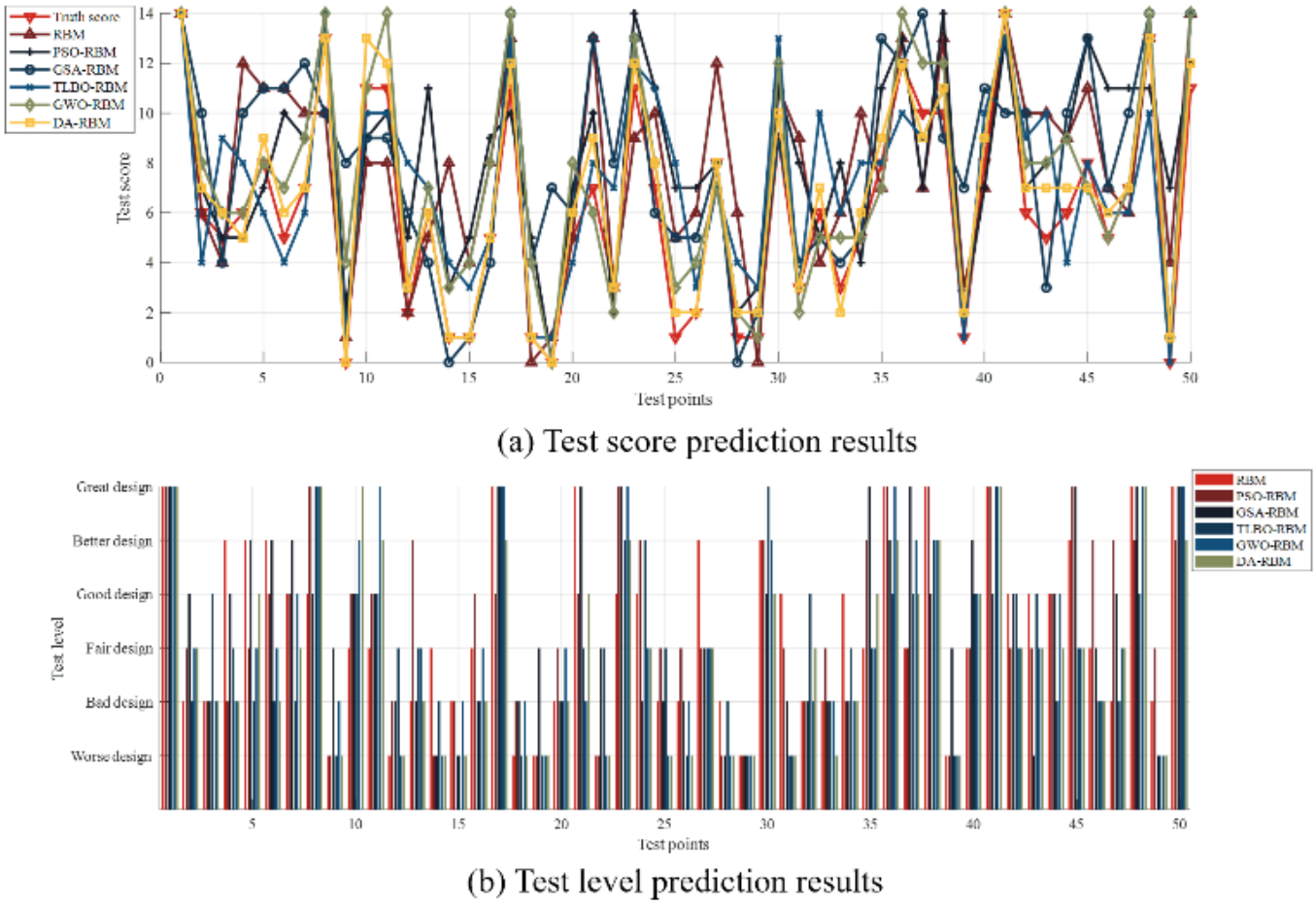


Fig. 17. Predicted performance results of test methods.

TABLE III. COMPARISON OF DESIGN TEST ALGORITHM PERFORMANCE

Arithmetic	*RMSE*	*Precision/%*	*Optimisation time/s*	*Test time/s*
RBM	0.0698	85.92	4.88	0.98
PSO-RBM	0.0549	88.76	5.83	0.87
GSA-RBM	0.0671	86.21	7.76	0.95
TLBO-RBM	0.0433	90.88	4.56	0.56
GWO-RBM	0.0368	90.35	5.45	0.59
DA-RBM	0.0178	93.50	4.40	0.48

VI. Conclusion

This study demonstrates that integrating augmented reality (AR) with advanced data-driven algorithms significantly enhances the digital design and testing process for intangible cultural heritage (ICH), particularly in the context of low-carbon tourism. By employing the Dragonfly Algorithm (DA) optimized Restricted Boltzmann Machine (RBM) model, we successfully improved the accuracy, efficiency, and robustness of AR digitization for the Qinhuai Lantern Festival.

The proposed AR digital design testing algorithm outperformed other methods, achieving a test score prediction accuracy of 93.5%. This superior performance highlights the model's capacity to process and analyze large datasets with greater speed and precision, making it an effective solution for assessing AR digital designs in the preservation and promotion of ICH.

By focusing on both cultural preservation and eco-friendly tourism, this method not only strengthens the protection of heritage but also contributes to sustainable tourism development. As AR technology continues to evolve, integrating intelligent optimization algorithms like DA will play a crucial role in shaping the future of digital cultural communication and heritage conservation.

References

[1] Yazli, N. C. Baka, E., Magnenat, T. N., Kaplanidi, D., Partarakis, N., Karuzaki, E.. Modelling craftspeople for cultural heritage: a case study[J]. Computer Animation and Virtual Worlds, 2022(3/4):33.

[2] Shouliang, L. Multidimensional Analysis on the Application of Digital Method in the Protection of Intangible Cultural Heritage[J].Design Research , 2014.

[3] Shi, G. W., Wang, Y. G., Liu, Y., Zheng, W. Application of augmented reality technology in digital preservation of cultural heritage[J]. Journal of System Simulation,2009,21(07):2090-2093+2097.

[4] Koutsabasis, P., Vosinakis, S. Kinesthetic interactions in museums: conveying cultural heritage by making use of ancient tools and (re-) constructing artworks[J].Virtual reality, 2018, 22(2):103–118. doi.org/10.1007/s10055-017-0325-0.

[5] Natalie, U. G. Participatory Research and Design in the Portal to Peru[J].Annals of Anthropological Practice, 2020, 44(1):119-125. doi.org/10.1111/napa.12131.

[6] Bai, J. S., Boo, J. C.Study on Museum Digital Exhibition Mode and Industrialisation of Intangible Cultural Heritage[J]. Communication convergence engineering, 2011, 9(2): 129-134. doi.org/10.6109/jicce.2011.9.2.129.

[7] Tu, S., Jiang, X. The "Tough Road" of Inheriting Intangible Cultural Heritage in the Age of Digital Media: Taking Dunhuang Art Academy as an Example[J]. British Journal of Philosophy, Sociology and History, 2022, 2 (2):01-04. doi.org/10.32996/pjpsh.2022.2.2.1.

[8] Gireesh, K. T. K. Sustainable preservation and accessibility to cultural heritage in India[J].Library Hi Tech News, 2023, 40(2):12-14. doi.org/10.1108/lhtn-02-2022-0029.

[9] Li, Y. J., He, Y. Trend, Mechanism and Path: Analysis of the Integrated Development of Henan Intangible Cultural Heritage and Tourist Industry with Digital Empowerment[J]. Chinese and Foreign Cultural Exchange: English Edition, 2023(4):12-15.

[10] Li, M., Wang, Y., Xu, Y. Q. Computing for Chinese Cultural Heritage[J]. Visual Informatics (English), 2022(001):006.

[11] Li, J. Grounded theory-based model of the influence of digital communication on handicraft intangible cultural heritage[J]. 2022, 10(1):1-12.

[12] Yu, X. Y., Li, N. A probabilistic prediction model for wind speed intervals based on restricted Boltzmann machines and rough sets[J]. Computer Application and Software,2023,40(03):157-166+240.

[13] Chen, Y. W., Wang, A. Q., Li, F. S. Application of population intelligence optimisation algorithm in medical field[J]. Chinese Journal of Medical Physics,2024,41(05):646-656.

[14] Jiang, M. L., Zu, Y. L., Jiang, S. Research on modern lamp design based on Qinhuai lantern culture[J]. Art Sea,2024,(03):74-78.

[15] Yang, Y. Protection and Innovation of Craft ICH Based on Digital Design--Taking Qinhuai Lantern Colours as an Example[J]. Art Education Research,2023,(20):47-49.

[16] Yang, F. Application of virtual reality technology in digital art design and preservation of intangible cultural heritage Qinhuai lantern colours[J]. Tiangong,2023,(28):40-42.

[17] Liu, Y. T., Tang, X. Y. Research on digital design of sex education for preschool children[J]. Design,2023,36(19):152-154.

[18] Skargren, F., Ambrosiani, K. G .The practitioners guide to a digital index: unearthing design-principles of an abstract artefact[J].Information Polity: The International Journal of Government and Democracy in the Information Age, 2022(1):27.

[19] Zhou, S. X. Shared information structure design of enrolment system based on digital campus[J]. Digital Technology and Application,2019,37(01):142-144.

[20] Peng, Z., Yin, Z. L., Chen, Z. Q. Research on optimisation method of intelligent substation operation auxiliary inspection system based on BIM+AR technology[J]. Electrotechnical Materials, 2023(2):4-6.

[21] Lu, K. H., Qiu, J., Liu, G. F. Research on life prediction method of electronic devices based on dynamic damage and optimised AR model[J]. Journal of Military Engineering, 2009(1):5.

[22] Su, A. Y. Research on the design and user experience of AR and voice interactive mobile game based on Unity engine[J]. Software, 2023, 44(12):76-80.

[23] Zhang, X., Xu J. Q., Wang, Y. H. Research on digital activation of Shanxi ancient theatre based on knowledge visualisation[J]. Art and Design: Theory Edition, 2022(2):77-79.

[24] Mirjalili, S. Dragonfly algorithm: a new meta-heuristic optimisation technique for solving single-objective, discrete, and multi-objective problems[J]. Neural Computing & Applications, 2016, 27(4): 1053-1073.

[25] Luo, H. Y. Research on bearing fault location technology for medical building construction based on restricted Boltzmann machine[J]. Automation Technology and Application, 2022, 41(5):14-17.

[26] Xie, D. L., Zhi, L. H., Zhou, K., Hu, F. Ultra-short-term wind speed combination prediction model based on VMD-ORELM-EC[J]. Journal of Hefei University of Technology(Natural Science Edition),2024,47(05):703-711.

[27] Wen, Y. B., Lei, J. Y. Optimal design of RBM network structure based on particle swarm algorithm[J]. Computer and Digital Engineering,2021,49(04):852-855.

[28] Puri, V., Chauhan, Y.K. Offline parameter estimation of a modified permanent magnet generator using GSA and GSA-PSO[J]. Soft Comput, 2022, 26, 6333–6345. https://doi.org/10.1007/s00500-021-06610-7.

[29] Cano-Ortega, A., Sanchez-Sutil F , Hernandez J C .Smart meter for residential electricity consumption with TLBO algorithm for LoRaWAN[J]. Electrical Engineering, 2023(4):2021-2040. https://doi.org/10.1007/s00202-023-01783-w.

[30] Pramanik, R., Pramanik P , Sarkar R .Breast cancer detection in thermograms using a hybrid of GA and GWO based deep feature selection method[J].Expert Systems with Application, 2023, 219, 119643 https://doi.org/10.1016/j.eswa.2023.11964.

Spatial Attention-Based Adaptive CNN Model for Differentiating Dementia with Lewy Bodies and Alzheimer's Disease

K Sravani, V RaviSankar
Department of Computer Science and Engineering, GITAM University, Hyderabad, India

***Abstract*—Differentiation of Alzheimer's Disease (AD) and Dementia with Lewy Bodies (DLB) utilizing brain perfusion Single Photon Emission Tomography (SPECT) is crucial and it might be difficult to distinguish between the two illnesses. The most recently discovered characteristic of DLB for a possible diagnosis is the Cingulate Island Sign (CIS). This work aims to differentiate DLB and AD by utilizing a deep learning model and this model is named AD-DLB-DNet. Initially, the required images are collected from the benchmark dataset. Further, the Spatial Attention-Based Adaptive Convolution Neural Network (SA-ACNN) is used to visualize the CIS features from the images where the attributes are tuned using Improved Random Function-based Birds Foraging Search (IRF-BFS). Further, CIS features attained from the SA-ACNN are used to accurately differentiate the DLB and AD. Finally, the Dilated Residual-Long Short-Term Memory (DR-LSTM) layer is proposed to accurately perform the AD and DLB differentiation for identifying the clinical characteristics of the DLB. The suggested model is used for differentiating between AD and DLB for taking effective therapeutic measures. Finally, the validation is performed to validate the effectiveness of the introduced system.**

Keywords*—Alzheimer's disease and dementia with lewy bodies differentiation; spatial attention-based adaptive convolution neural network; cingulate island sign; improved random function-based birds foraging search; dilated residual-long short-term memory

I. INTRODUCTION

Global healthcare systems are severely impacted by neurodegenerative dementias, particularly as the number of elderly people rises. The World Health Organization (WHO) reports that approximately 50 million individuals globally are affected by dementia [1]. AD is responsible for around 60% of these cases, making it the most common neurological disorder [2]. DLB, characterized by the accumulation of LB, is the second most prevalent type of neurodegenerative dementia, following AD and some cases are often misdiagnosed and overlooked [3], [4]. In addition to identifying and managing clinical aspects such as severe autonomic dysfunction, motor and mental symptoms, and dangerous antipsychotic sensitivity, an accurate and timely detection of DLB is crucial for ensuring appropriate care and treatment [5].

Predicting the disease's prognosis and organizing clinical trials also depend on a reliable diagnosis, but the significant clinical and cognitive similarities between AD and DLB may make the diagnostic procedure more difficult [6]. Additionally, a variety of clinical manifestations may result from the common presence of pathological variability in individuals containing DLB, particularly the presence of co-occurring AD pathology, such as tau tangles and amyloid beta (Aβ) plaques [7]. Compared with DLB patients exhibiting solely Lewy body pathology, those with Aβ pathology is linked to reduced life expectancy and a higher rate of cognitive impairment [8]. These findings highlight the clinical significance of detecting concomitant amyloid-beta (Aβ) pathology in patients with DLB. Functional neuroimaging, a commonly employed tool in the medical detection of dementia, it has also been integrated into the detection criteria for AD as well as DLB [9], [10]. Even seasoned neurologists find it difficult to diagnose certain conditions, and sometimes choosing the best course of action is also difficult. Therefore, to give more reliable clinical evaluations, doctors employ diagnostic techniques such as neurofunctional imaging [11], [12].

Recently deep learning techniques for medical image analysis are growing steadily, particularly in neurodegenerative illnesses [13]. This broad recognition stems from its capacity to automatically identify useful features and reduce the requirement for handcrafted feature extraction. Unlike typical machine learning approaches, it can learn intricate patterns in imaging information which is difficult for humans to perceive [14], [15]. Most deep learning models used in neurodegenerative illnesses primarily identify many stages of AD, ranging from no dementia to mild AD, utilizing 2D imaging scans. Nevertheless, these models are only useful for the AD diagnosis, which means they cannot distinguish the patterns between AD and DLB. Furthermore, it is challenging to confirm their robustness when non-AD dementias are present [16], [17]. The quantitative approach requires standardized methods for acquiring and interpreting structural scans and 18F-FDG-PET, which can be time-consuming. Interestingly, in differentiating DLB from AD, a straightforward visual evaluation of Cortical Involvement (CIS) as either present or absent proved to have higher diagnostic accuracy than the quantitative CIS ratio. Although visual evaluation of other imaging indicators and modalities is widely utilized and has shown to be a quick, accurate, and repeatable procedure in clinical practice, there are no standardized visual guidelines to assess the extent of CIS. In addition to increasing the diagnostic accuracy of DLB, the use of pertinent diagnostic data may be improved using a consistent approach for classifying and interpreting the presence of CIS. Additionally, it is simple to incorporate a visual grading system into clinical practice across sites. To

effectively identify the differences between DLB and AD, a new deep learning model is introduced. The following points highlight the contributions of the developed framework.

- To develop a deep learning model to differentiate DLB from AD by utilizing images from a benchmark dataset. This approach enhances diagnostic accuracy by allowing the model to learn subtle differences between DLB and AD from high-quality, standardized brain images.
- To employ SA-ACNN to visualize and extract CIS, enabling the model to focus on the most relevant regions in medical images, thereby improving feature accuracy and relevance.
- To reintroduce BFS as IRF-BFS to fine-tune and optimize the CIS features. The parameters of SA-ACNN such as hidden neuron count, steps per epoch count, and epoch size are optimized that maximize the accuracy.
- To integrate the Dilated Res-LSTM layer, enabling the model to accurately identify and differentiate clinical features of DLB and AD, which aids in early diagnosis and effective therapeutic intervention.

The structure of the newly introduced deep learning technique for differentiating DLB and AD is outlined as follows. Section II reviews the literature on DLB and AD differentiation models. Section III presents an adaptive deep-learning mechanism designed to enhance DLB and AD differentiation, utilizing an improved optimization algorithm to boost performance and accuracy. Section IV explains the proposed model for feature extraction. Section V introduces a novel approach for differentiating DLB and AD. Part VI provides the Experimental results. Section VII contains Comparative analysis detailed discussion. Finally, Section VIII concludes the study.

II. Literature Survey

A. Related Works

In 2023, Nakata et al. [18] assessed the brain imaging variance among MCI with Lewy Bodies (MCI-LB) as well as MCI due to AD (MCI-AD) by examining brain atrophy and brain perfusion patterns. The analysis focused on differences in regional brain changes in individuals with these two conditions. It was found that MCI-LB and MCI-AD exhibited distinct patterns of brain atrophy and blood flow abnormalities. These differences helped distinguish between the two types of MCI, highlighting the unique features associated with each condition.

In 2024, Karim et al. [19] have used graph theory and machine learning measures to forecast AD. Several machine learning models were developed for AD prediction using the OASIS and SALD datasets. The study identified key elements of functional connectivity and brain network structure in AD, noting a significant loss of connections between the thalamus and top 13 regions. These findings highlighted the potential of combining machine learning, graph theory for accurate AD diagnosis and for early prediction.

In 2024, Hasan and Wagler [20] have suggested CNN-GCN architecture which was produced by first implementing the CNNs and feeding it to the GCN classifier. To train and assess the suggested techniques the whole-brain images were used. They evaluated the effectiveness of the technique by presenting the findings from the best fold out of the five folds.

In 2022, Etminani et al. [21] have developed a 3D deep learning model that utilized PET scans with a specific radioactive tracer to forecast the final clinical diagnosis of DLB, AD, and other conditions. The performance of this model was compared to that of experienced nuclear medicine physicians. To visualize the regional metabolic changes, methods were employed to highlight the areas of interest.

In 2020, Gjerum et al. [22] implemented a strong visual CIS scale and assessed its ability to distinguish between AD and DLB. When compared to AD patients and controls, DLB patients' visual CIS scores were much greater. To sum up, the visual CIS scale was a clinically helpful tool for distinguishing AD from DLB. Aβ pathology in DLB patients may be connected to the severity of CIS.

In 2020, Kanetaka et al. [23] proposed prospective research comparing the CIS on Single Photon Emission Computed Tomography (SPECT) in individuals. The CIS score, calculated using eZIS software, is the ratio of the posterior cingulate gyrus (VOI-1) to areas of notably decreased regional cerebral blood perfusion (VOI-2). Due to insignificant RCBF decline in the PCG, diagnosing MCI with the CIS score is challenging.

In 2022, Lim et al. [24] suggested a multiclass categorization technique using 3D T1-weight brain MRI images. The ResNet-50 and VGG-16 convolutional bases were utilized as feature extractors. A novel densely connected classifier was put into place to do classification on top of the convolutional bases.

In 2017, McKeith et al. [25] made a clear distinction between clinical characteristics and diagnostic biomarkers and provided guidelines on the best ways to determine and interpret them. Here, the diagnostic role of laboratory, electrophysiologic, and neuroimaging tests has been expanded. Significant progress has been made in recognizing DLB.

B. Problem Statement

Millions of people worldwide are greatly affected by the serious disorder called AD. Behavioral abnormalities and memory loss are the symptoms associated with AD. The structural changes in the brain are the main cognitive dysfunction caused by AD. To initiate the treatment approaches, dementia and AD must be detected at an earlier stage but the traditional model faces various issues, and it is listed in Table I.

Traditional strategies do not have the capability to diagnose the CIS from the images, so they fail to differentiate among the AD and DLB.

The functional connectivity of the brain is not detected by this model and this model is so invasive and costly making it unsuitable for the early diagnosis process.

TABLE I. FEATURES AND CHALLENGES OF EXISTING DLB AND AD DIFFERENTIATION MODEL

Author [citation]	Methodology	Features	Challenges
Nakata et al. [17]	RCBF	The symptoms of mild cognitive impairment are effectively detected by this model. It is used for the early AD detection.	This model fails to identify the signs of dementia.
Karim et al. [18]	SVM	This model accurately defines the structure of the brain network.	The characteristics of the brain network are not analyzed by this model.
Hasan and Wagler [19]	CNN-GCN	The initial symptoms of AD are diagnosed.	The imbalanced dataset cannot handle.
Etminani et al. [20]	3D deep learning	The systems robustness is high. The proposed model is applied in the clinical setting for making the effective decision.	The transparency of the system is low.
Gjerum et al. [21]	robust visual rating scale	The presence of dementia is effectively detected by this model. The degree of the CIS is determined by this model.	It lacks in the pathological information. The memory cohort is not analyzed by this model.
Kanetaka et al. [22]	DLB	The volume of the CIS is measured using this model.	The symptoms of the disease cannot be diagnosed.
Lim et al. [23]	CNN	It is used for executing the multi-classification of images. It uses dense connections for accurately classifying the AD in the humans.	This model does not evaluate the low-dimensional feature scores
McKeith et al. [24]	Optimal AD detection methods	It is employed to support the medical decision-making process. It is used to provide adequate medical support to the patients	The behavioral abnormalities are not detected by this model

The prior systems are ineffective for preventing the progression of DLB in individuals as they are not effectively determining the synchronization of the brain regions.

The prior approaches are unsuitable for discriminating against DLB patterns from AD patterns, so it is quite difficult to automatically detect the presence of DLB and AD from humans.

III. IMAGING CLASSIFICATION OF DEMENTIA WITH LEWY BODIES AND ALZHEIMER'S DISEASE USING DEEP LEARNING NETWORK

A. Proposed DLB and AD Imaging Classification Model: Description

Models for differentiating DLB and AD typically depend on manual feature extraction, a process that can be time-intensive and susceptible to human error. The accuracy of these models is limited due to the overlap in clinical symptoms between the two conditions. Traditional methods may not effectively capture complex patterns in neuroimaging data, leading to misdiagnosis. Additionally, these models often lack the ability to integrate and analyze multiple types of data simultaneously. Consequently, a deep learning model was developed to enhance differentiation by automatically learning and identifying intricate patterns in imaging and clinical data, improving diagnostic accuracy and efficiency. However, using deep learning for differentiating DLB and AD may face challenges including limited availability of labeled data, high variability in brain scans due to individual differences, difficulty in differentiating subtle disease patterns and so on. Thus, it is necessary to develop novel DLB and AD differentiating systems with the support of enhanced deep learning mechanism.

This study aims to differentiate DLB from AD using an advanced deep learning model. The process begins with collecting the necessary brain images from a benchmark dataset. Initially, the brain images are fed into the SA-ACNN to visualize critical image features, where SA-ACNN is developed by integrating spatial attention layer into the CNN architecture along with the network parameter optimization. For performing this optimization, an efficient heuristic algorithm named BFS is reintroduced as EBFS. Here, the parameters such as steps per epoch count, hidden neuron count, and epoch size in SA-ACNN are optimized to maximize the accuracy. The SA-ACNN's spatial attention mechanism focuses on the most relevant parts of the images, improving CIS feature extraction. Moreover, the EBFS fine-tunes this process for supporting more precise differentiation. These extracted CIS features are crucial for distinguishing between DLB and AD. The CIS involves the preservation of the Posterior Cingulate Cortex (PCC) in DLB, while hypoperfusion is typically seen in this region during the early stages of AD. The presence of the CIS has gained attention as a key differentiator, reflecting AD-related pathology that influences clinical symptoms in DLB. Notably, CIS is most prominent during the mild dementia stage and tends to decline as DLB advances. This makes CIS particularly useful for distinguishing DLB from AD, especially in the early stages, though exceptions like posterior cortical atrophy may complicate this distinction. The CIS features are then fed into the DR-LSTM for classifying DLB and AD images. This DR-LSTM combines the benefits of dilated convolutions (captures multi-scale context), and the strengths of Res-LSTM (effectively handles sequential data and long-term dependencies). Moreover, it helps in capturing detailed temporal patterns and clinical features, enhancing diagnostic capability. Thus, the DR-LSTM is expected to significantly improve the differential diagnosis of DLB and AD, enabling more effective therapeutic measures. Fig. 1 presents the pictorial presentation of the proposed DLB and AD imaging classification model.

B. Brain Image Dataset for Model Analysis

The developed framework employs brain images to differentiate DLB from AD. Table II provides a description of the dataset, which consists of images gathered from online

resources. In this context, the term $Im\,g_c^{Br}$ represents the brain images, here $c = 1,2,\ldots,C$ and C indicates the image count.

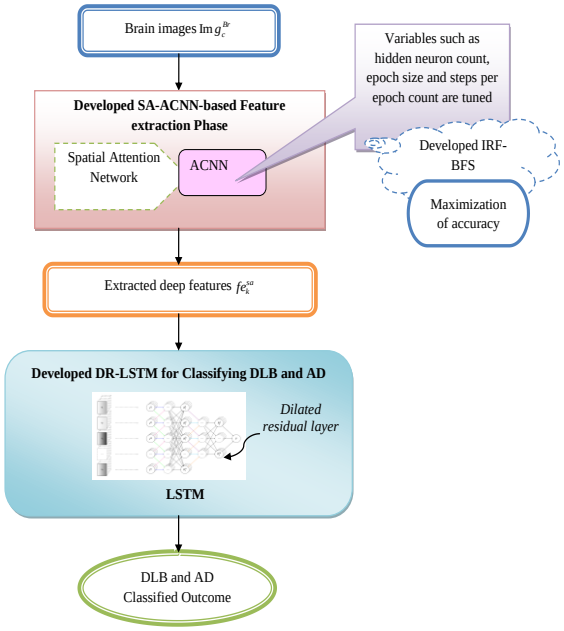


Fig. 1. Pictorial representation of the proposed DLB and AD imaging classification system.

TABLE II. DESCRIPTION OF THE INPUT IMAGE DATASET

Dataset name	Dataset link	Dataset description
Dataset1 (FDATA ADNI DATASET)	https://www.kaggle.com/datasets/ahmedashrafahmed/fdata-adni-dataset.	This dataset consists of 33,984 records, with 6,000 records selected for use. These 6,000 records are divided into training and testing sets, with 4,500 used for training and 1,500 for testing. The dataset is categorized into four classes, each with 1,500 records: CN (Cognitively Normal), LMCI (Late Mild Cognitive Impairment), AD, and EMCI (Early Mild Cognitive Impairment), where LMCI and EMCI comes under the category of DLB.

IV. SPATIAL ATTENTION BASED ADAPTIVE CONVOLUTION NEURAL NETWORK FOR CINGULATE ISLAND SIGN FEATURE EXTRACTION

A. Convolution Neural Network

The convolutional layer is a crucial part of feature extraction in a Convolutional Neural Network (CNN) [26], which employs certain hidden layers. CNN contains more than two hidden layers, and these layers interpret the image as a tensor, automatically extracting features and performing eventual categorization from input data. The usual CNN layers are as follows:

Input layer: The width, length, and number of channels, or their transformations, constitute the input tensor, which determines the size of the input layer, which contains the image information.

Convolutional layer: Transformation layers are those that imply the warping process from the preceding layer. This layer gathers the training outcome's weights or parameters. Usually smaller in width and length than the input layer, the output of this layer is a tensor known as a feature map, with a depth dimension. This layer aids in storing the training weight, which is represented by Eq. (1).

$$E_{wx} = \left(k1 * k2 * E_{PZ} * E_{QW} + E_{MF}\right) \quad (1)$$

Here, the kernel size is represented with the terms $k1$ and $k2$.

Pooling layer-This layer helps in reducing the size of the previous layer to identify the important features from the input tensor. The output dimension is determined by the kernel size; for instance, with a kernel size of two, the output dimension is divided. Fig. 2 illustrates the pictorial representation of a CNN for feature extraction process.

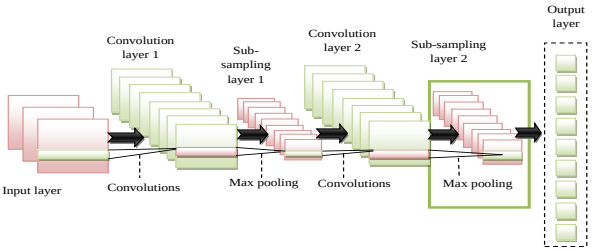


Fig. 2. Diagrammatic representation of CNN for feature extraction process.

B. Developed SA-ACNN-Based CIS Feature Extraction

Initially, the input brain images $Im\,g_c^{Br}$ are given for the feature extraction phase. The SA-ACNN utilizes spatial attention mechanisms to focus on the most relevant areas of medical images, such as lesions or other distinguishing patterns, enhancing the accuracy of feature extraction. This ability to prioritize important image regions allows the network to more effectively differentiate between DLB and AD. To further refine this process, IRF-BFS is applied to optimize the attributes of SA-ACNN. The parameters such as steps per epoch count, hidden neuron count, and epoch size in SA-ACNN are optimized to maximize the accuracy. Thus, it ensures that only the most relevant features are extracted for accurate diagnosis, with SA-ACNN focusing on key image areas. Finally, the CIS features are extracted which means neuro imaging features seen on DLB. The extracted features are represented with the term fe_k^{sa}.

CIS-based feature processing: In the process of differentiating DLB and AD, the model integrates Grad-CAM (Gradient-weighted Class Activation Mapping) to offer a deeper understanding of how the deep learning network reaches its diagnostic conclusions. Grad-CAM is a powerful visualization tool that helps to highlight which areas of the input image are most influential in the model's final predictions. Grad-CAM is specifically employed to locate and emphasize the CIS. This feature enables both clinicians and researchers to visually track the model's focus during the DLB–AD classification process. For images of DLB patients, Grad-CAM frequently highlights the CIS, demonstrating that the model places significant emphasis on this feature to differentiate DLB from AD. As the model continues to learn, its focus on the CIS becomes more pronounced and localized. When the CIS appears more prominently, the model assigns higher confidence to the DLB diagnosis, while images with

lower CIS ratios, suggesting the presence of AD pathology, lead to reduced confidence in DLB classification. The description of the spatial attention mechanism is provided below.

The spatial attention [27], [28] refers to a process that focuses on areas of input (usually images or sequences) that are most pertinent to the task at hand. When the model is creating predictions or extracting features, it uses this type of attention mechanism to help it prioritize geographic locations in the incoming images. Eq. (2) and Eq. (3) illustrate how the spatial attention mechanism is calculated.

$$F = N_f . \sigma \left(\left(v_i^{(y-1)} R_1 \right) R_2 (R_3 v_p^{(y-1)})^U + m_f \right) \quad (2)$$

$$F'_{p,l} = \frac{exp(F_{p,l})}{\sum_{l=1}^{Z} exp(F_{p,l})} \quad (3)$$

Here, the input of the y^{th} spatial block is signified with the term $v_i^{(y-1)}$ and the channels in the input images are specified with the term B_{y-1}. The learnable parameters are represented with the terms N_f and m_f and the term σ is utilized as the activation function. The element $F'_{p,l}$ in F signifies the semantic correlation strength between node p and node l. The diagrammatic representation of the developed SA-ACNN-based CIS Feature Extraction is presented in Fig. 3.

C. Parameter Optimization with IRF-BFS

The BFS algorithm is selected in this model as it offers several advantages, including efficient global search capabilities, simplicity, adaptability to various optimization problems, parallelism, and a good balance between exploration and exploitation. However, BFS also has some drawbacks, such as limited exploration of the search space, premature convergence and sensitivity to parameter tuning. These disadvantages can be addressed by using the IRF-BFS, which avoids randomness in the search space by updating the random variables to make accurate solution for the search process.

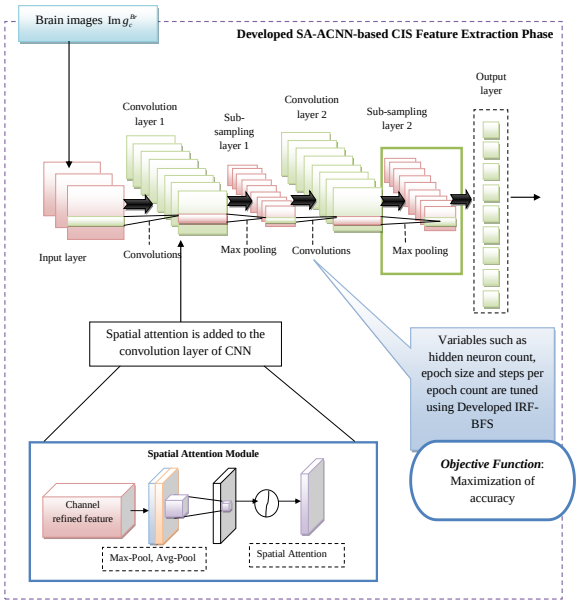


Fig. 3. Graphical representation of the implemented SA-ACNN-based CIS feature extraction.

IRF-BFS overcome premature convergence by adding diversity to the search, improving exploration, and minimizing the risk of getting stuck in local optima. Additionally, IRF-BFS reduces the dependency on manually tuning parameters by adapting them during the search. In this improved IRF-BFS approach, the random variable f is upgraded utilizing Eq. (4).

$$f = \frac{c_f}{\frac{a}{b}} \quad (4)$$

$$a = \frac{b_f}{c_f + w_f} \quad (5)$$

$$b = \frac{c_f + w_f}{b_f} \quad (6)$$

Here, the current fitness value is signified with the term c_f, the mean fitness value is signified with the term m_f, the worst fitness value is specified with the term w_f, and the best fitness value is represented with the term b_f. The pseudocode of the IRF-BFS is given in Algorithm 1.

Algorithm 1: Developed IRF-BFS

Set the values for the parameters: population Z and maximum iteration Max_{ite}

While $(ite \leq Max_{ite})$ do

- Update random variable $\boldsymbol{f}$ that is computed in Eq. (4)
- Perform flying search behavior
  - Estimate the new location of the bird in flying search region
- Perform Territorial behavior
  - Determine the new territory bird's position
  - Estimate the fitness function
  - Determine the new incursion birds' position
- If the position of the leading bird is superior to that of all other birds
  - Execute the role change mechanism
- End if
  - Examine the border
  - Estimate the fitness function
  - Upgrade M_p with M_p^{ite+1}

$$ite = ite + 1$$

End while

Output

Output

D. Objective Function of IRF-BFS-SA-ACNN Model

In the developed IRF-BFS-SA-ACNN-based feature extraction model, IRF-BFS is applied to optimize the attributes of SA-ACNN. The parameters such as steps per epoch count, hidden neuron count, and epoch size are optimized to maximize accuracy. The objective function of the developed IRF-BFS-SA-ACNN is expressed mathematically in Eq. (7).

$$Obj = \underset{\left\{Hdd_i^{cnn}, Eoo_j^{cnn}, Spe_k^{cnn}\right\}}{argmax} (acc) \quad (7)$$

Here, the term accindicates the accuracy, the term Spe_k^{cnn} denotes the steps per epoch count in the CNN with range of [500,1000], the term Eoo_j^{cnn} represents the epoch size with the range of [5,50] and the term Hdd_i^{cnn}signifies the hidden neuron count in the CNN with the range of [5,255].

The accuracy acc is a measure of how often a method accurately forecasts an outcome and it is derived in Eq. (8).

$$acc = \frac{(ka+en)}{(ka+en+s\square+ik)} \quad (8)$$

Here, the termkadenotes the true positive, en indicates the true negative, sh denotes the false positive, ik denotes the false negative values.

V. Dilated Residual-Long Short Term Memory for Differentiating Dementia with Lewy Bodies and Alzheimer's Disease

A. Long Short Term Memory

A memory element could replace each hidden element in the LSTM [29] . Each memory element is made up of different parts including input, output, forget, and internal states. The operation of the input and reset gate is achieved by Eq. (9) and Eq. (10).

$$m_d = \sigma\left(V_{bg} i_d + V_{lz} z_{d-1} + b_b\right) \quad (9)$$

$$j_d = \sigma\left(V_{ji} i_d + V_{zj} z_{d-1} + b_j\right) \quad (10)$$

The operation of the cell state and output gate is expressed by Eq. (11) and Eq. (12).

$$g_d = j_d \Theta g_{d-1} + l_d \Theta k\left(V_{gi} i_d + V_{gz} z_{d-1} + b_g\right) \quad (11)$$

$$t_d = \sigma\left(V_{wg} i_k + V_{wz} z_{k-1} + b_w\right) \quad (12)$$

The hidden state and memory state is attained in Eq. (13) and Eq. (14).

$$z_d = w_d \Theta p(g_d) \quad (13)$$

$$z_a = X_{zn} n_a + b_z \quad (14)$$

Here, the terms g_{d-1},$j_d$denotes the internal state, a forget gate, the terms m_d and t_d denotes an input gate, an output gate, the term V represents the weight matrix, the term σ represents the logistic sigmoid function, element-wise multiplication is denoted by the symbol Θ, z represents cell result activation point, the term $b$indicates bias, the term i and c indicates the input point and output point, and the terms $k$and $p$indicates the $tanh$activation operations, accordingly.

B. Dilated Residual-LSTM for Classifying DLB and AD

The extracted CIS features fe_k^{sa} are provided into the classification phase. The developed DR-LSTM helps to perform the differentiation process of DLB and AD by accurately identifying specific clinical features. It combines the strengths of dilated convolutions and residual layer networks. Dilated convolutions expand the receptive field, enabling the model to capture multi-scale contextual data, which is crucial for identifying subtle, long-range dependencies in disease progression. This combination of dilated convolution and residual layers improves the ability of the model to represent and track the evolving clinical features of DLB and AD, improving diagnostic accuracy. This integrated approach leads to a more robust and precise diagnostic process, ultimately aiding in the early and effective therapeutic intervention for these neurodegenerative diseases.

A residual block [30] is a fundamental component of residual network architecture. It is developed to mitigate the vanishing gradient issue and make training deep networks more feasible. The residual block can be mathematically indicated in Eq. (15).

$$i = H\left(v, \{R_p\}\right) + v \quad (15)$$

Here, the term $v$signifies the input of the residual block; the term $H\left(v, \{R_p\}\right)$indicates the residual mapping.

Dilated convolutions [32] are commonly used in models where capturing long-range dependencies is crucial and they allow the model to maintain high resolution while increasing the receptive field. This method is particularly advantageous in scenarios where traditional convolutions would result in excessive computation or loss of resolution due to down sampling. The function for a convolution can be computed using Eq. (16).

$$i(u) = \sum_{p=0}^{a} v(u + p.g).r(p) \quad (16)$$

Here, the term $r(p)$represents the weight at the index$p$, the term $g$represents the dilation factor, and $a$indicates the filter dimension. Finally, the classified outcome is obtained for identifying the DLB and AD classes. Fig. 4 represents the pictorial representation of developed DR-LSTM for Classifying DLB and AD.

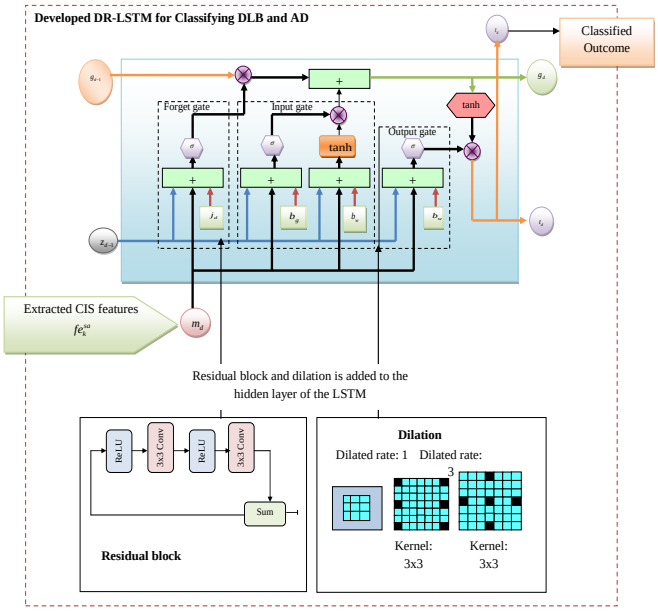


Fig. 4. Diagrammatic illustration of the developed DR-LSTM for classifying DLB and AD.

VI. Experimental Results

A. Resultant Feature Images by Varying Iteration

The developed DLB and AD Imaging Classification system was implemented utilizing Python. While developing the network, a maximum iteration of 50, a chromosome length of 3, and populations of 10, was considered. Fig. 5 provides the resultant feature images of the developed SA-ACNN by varying iteration.

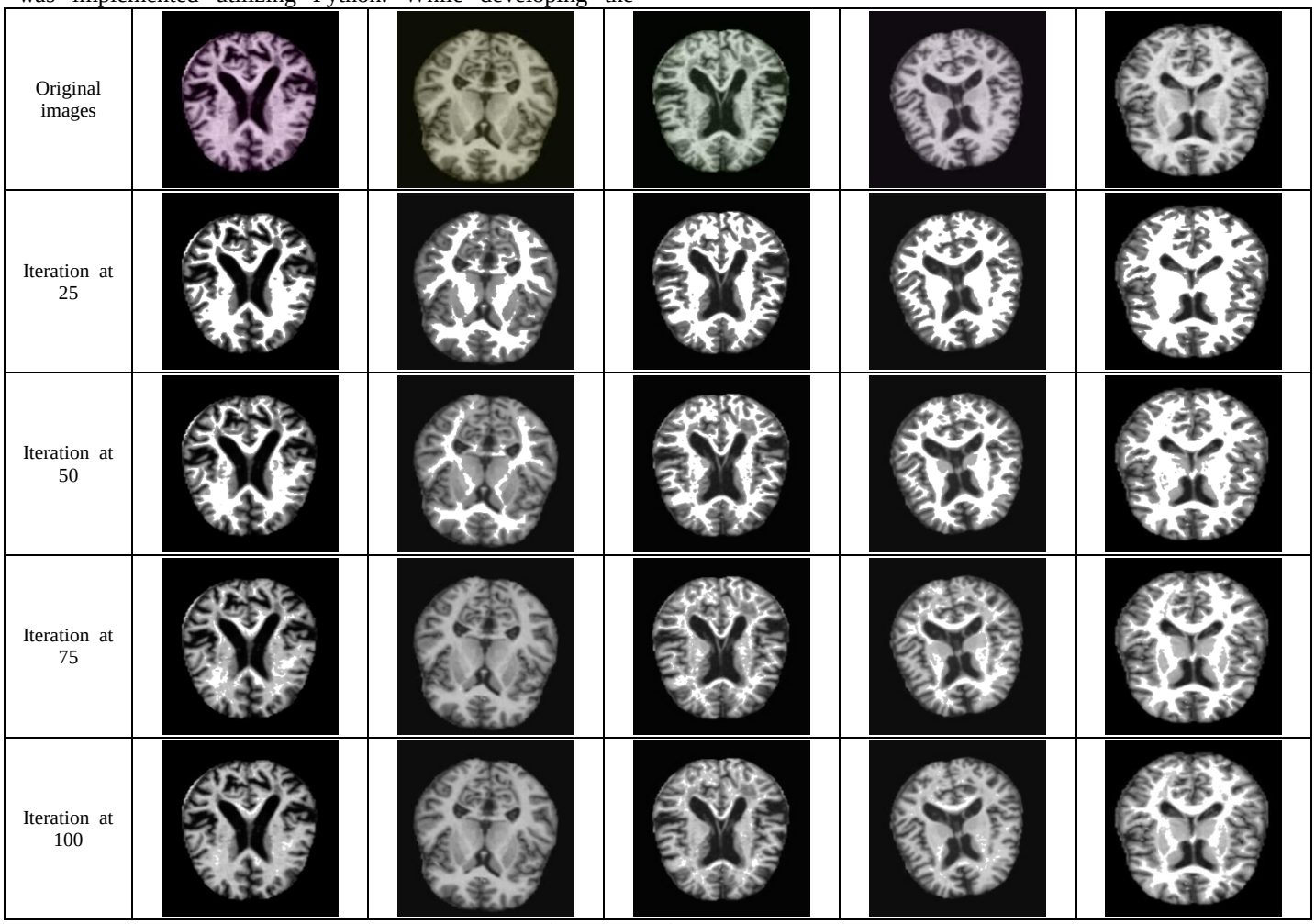


Fig. 5. Resultant feature images from the developed SA-ACNN.

B. Training and Testing Progress

The training progress graph is used to visualize the performance of a machine learning model during training and testing periods. This graph is a plot of the testing accuracy and testing loss over 1000 epochs. Fig. 6 and Fig. 7 provide the training progress graphs for the developed method. In Fig. 8 and Fig. 9, the testing accuracy increased rapidly, indicating that the model is learning and improving. However, after around 100 epochs, the accuracy fluctuates and stabilizes with some noise. The fluctuations suggest that the model's performance varies slightly with each epoch. In testing, the loss minimizes rapidly, showing that the model is learning to make better predictions. After around 100 epochs, the decrease in loss slows down and eventually stabilizes, showing minor reductions. The constant loss model suggests that a point has been reached where further training will not significantly improve performance.

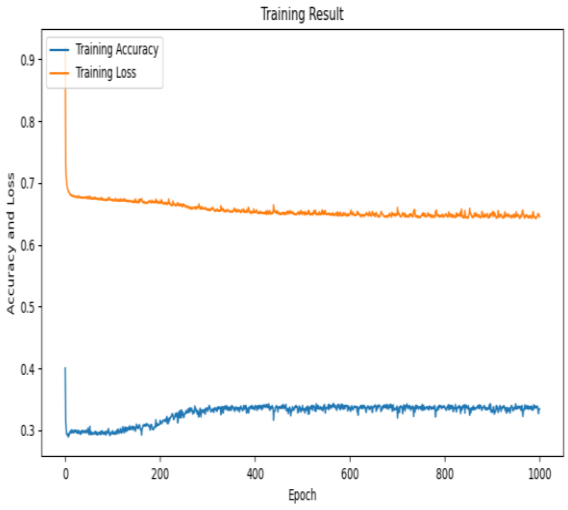


Fig. 6. Training accuracy of CNN model.

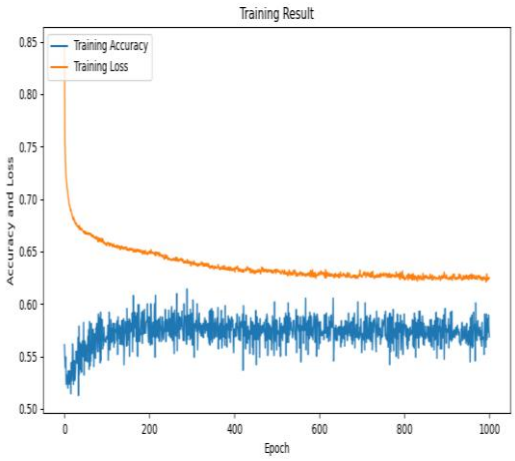


Fig. 7. Training accuracy of LSTM model.

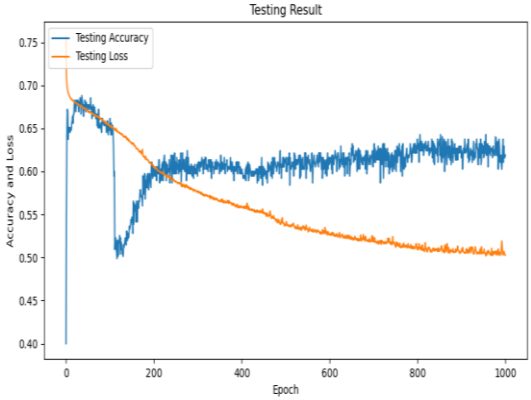


Fig. 8. Testing accuracy of CNN model.

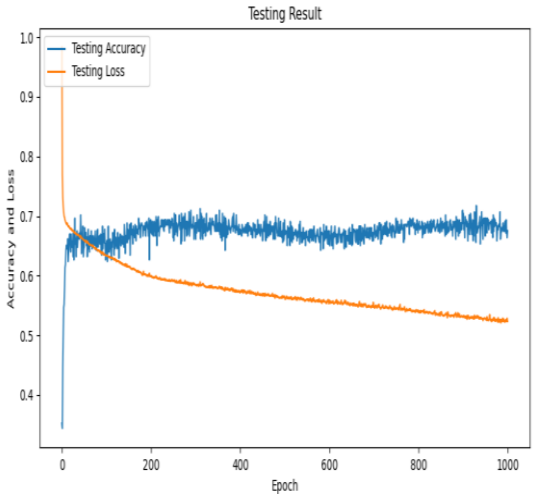


Fig. 9. Testing accuracy of CNN model.

C. Correlation Analysis

This statistical method assesses the strength and direction of the relationship between two continuous variables. Here, the black line represents a linear regression model, fitting the data points, which indicates a trend that supports this positive relationship. This type of analysis provides evidence that the features extracted by the SA-ACNN (i.e., the CIS ratio) are relevant for distinguishing between DLB and AD. It supports the effectiveness of the model in capturing clinically meaningful data that can be linked to the disease's severity or progression. Fig. 10 represents the correlation analysis of the developed SA-ACNN-based feature extraction model between AD and DLB classes.

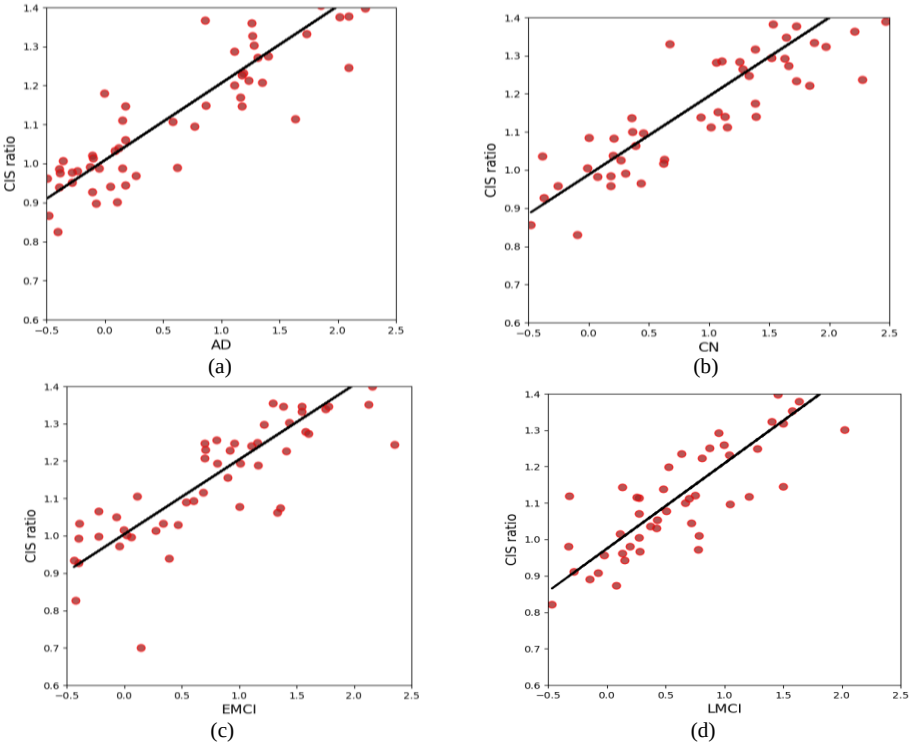


Fig. 10. Correlation analysis of the developed SA-ACNN-based Feature Extraction Model regarding different classes (a) AD, (b) CN, (c) EMCI, and (d) LMCI.

VII. Comparative Analysis and Discussion

A. Batch Size-Based Performance Analysis of Proposed Classification Model

The batch size-based performance analysis evaluates how different models perform on a given task by testing them on various batch sizes and evaluating metrics such as accuracy, MCC, CSI, FPR, FDR, precision. In Fig. 11(a), the accuracy of the developed AD-DLB-DNet framework outperforms RAN,

SVM, CNN-GCN, and CNN by 11.76%, 6.74%, 13.09%, and 3.26% in batch size-4, correspondingly. Thus, it is noted that the introduced DR-LSTM provides better performance than other classification techniques. Fig. 11 provides performance analysis by varying batch sizes.

B. ROC Analysis

Receiver Operating Characteristic (ROC), which is a metric, utilized to estimate the execution of a classification model, such as the DLB and AD differentiation model. The techniques compared, including RAN, SVM, CNN-GCN, CNN, and AD-DLB-DNet, all exhibit powerful performances. However, AD-DLB-DNet appears to have a slight edge over the others, making it a potentially more reliable choice for differentiating AD and DLB in clinical settings. Fig. 12 offers the ROC graph analysis of the proposed network.

C. Convergence Analysis

Convergence analysis refers to the study of how well the model's performance improves as the number of training iterations or epochs increases. The effectiveness of the implemented framework was estimated by comparing with several heuristic algorithms like Dwarf Mongoose Optimization (DMO) [31], Sparrow Search Algorithm (SSA) [32], Dingo Optimization Algorithm (DOA) [33], Birds Foraging Search (BFS) [34], and classifiers like RAN [35] , SVM, CNN-GCN, and CNN.

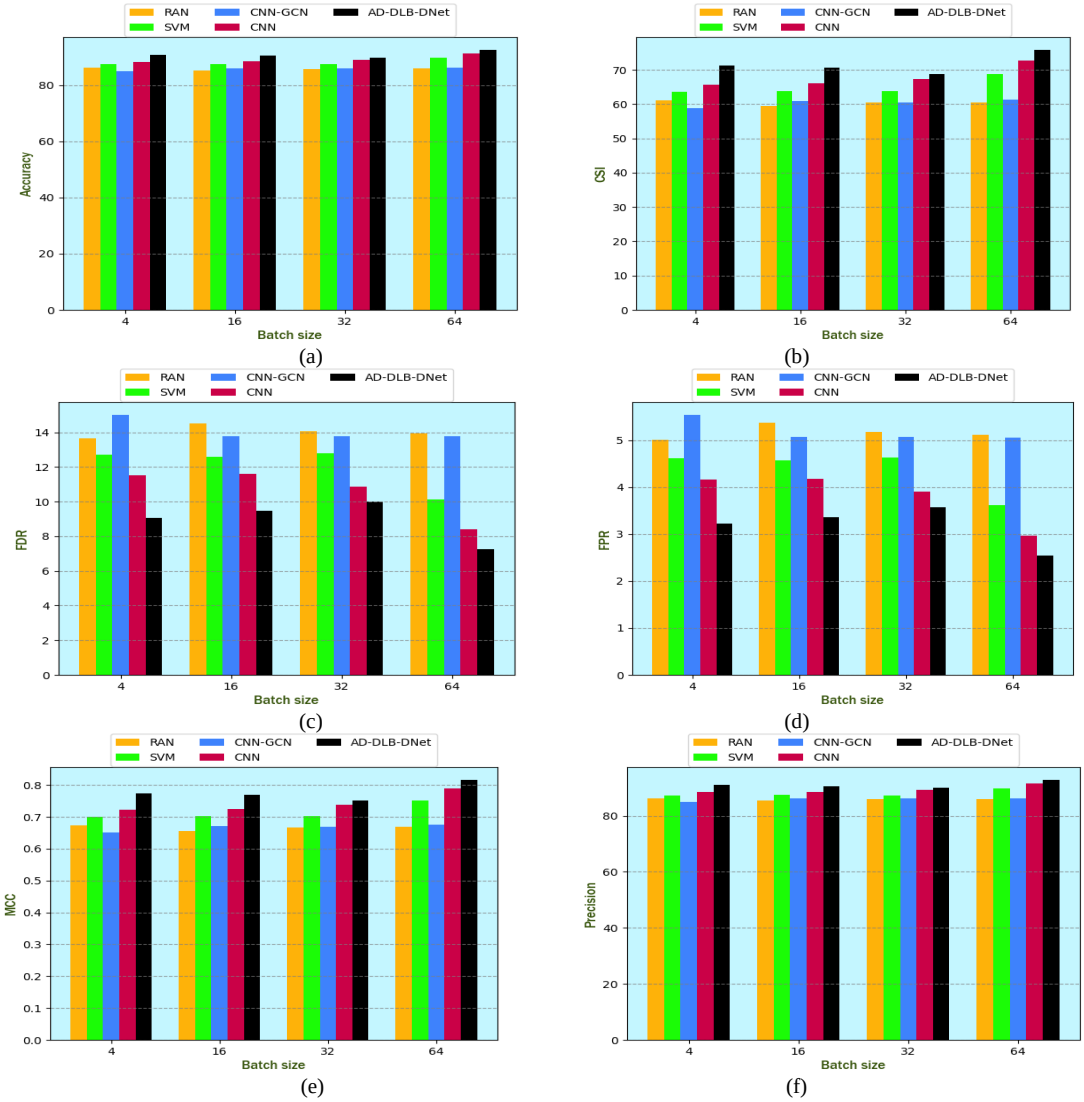


Fig. 11. Batchsize-based performance analysis of the developed method regarding (a) Accuracy, (b) CSI, (c) FDR, (d) FPR, (e) MCC, (f) Precision.

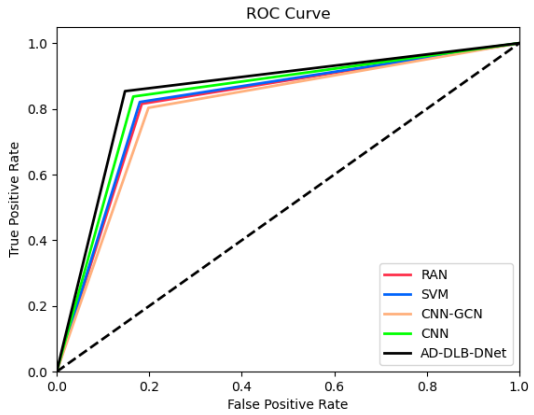


Fig. 12. ROC analysis of the developed technique.

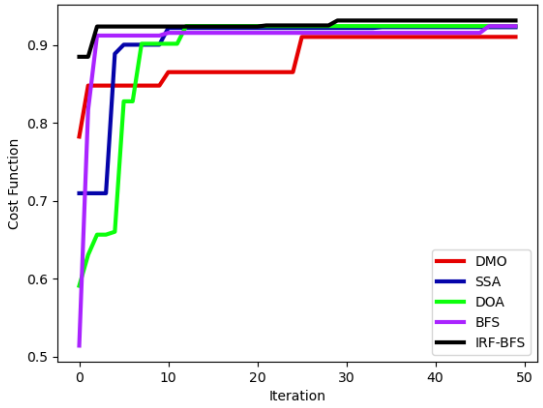


Fig. 13. Convergence analysis of the proposed technique.

At 10th iteration, the developed IRF-BFS approach performs better than the existing algorithms like DMO, SSA, DOA, and BFS by 13.63%, 11.11%, 12.35%, and 8.69%. By performing convergence analysis, researchers can develop a robust and accurate model for differentiating DLB and AD, ultimately improving diagnostic accuracy and patient outcomes. Fig. 13 provides the convergence analysis of the developed framework.

D. Confusion Matrix of the Classification Model

The confusion matrix is a commonly used performance measurement tool in classification issues. It compares a model's classified classes with the actual ground truth labels, to assess how well the system performs. Moreover, it helps to identify areas where the model performs well and where improvements are needed. Consequently, it shows how well the model can differentiate between AD and DLB, particularly by analyzing how often the model misclassifies one disease as another. Fig. 14 provides the confusion matrix of the developed model.

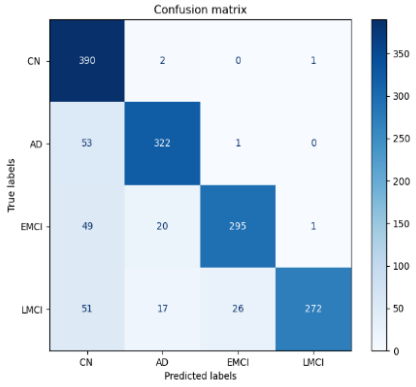


Fig. 14. Confusion matrix of the proposed classification model.

E. Comparative Analysis with K-Fold Cross Vadildation

Comparative Analysis of the Proposed Classification model is always needed to analyze the effectiveness of the developed model. K-fold cross-validation analysis refers to a method that is utilized to assess the effectiveness of a model, especially in cases where you want to ensure your method generalizes well with new information. K-fold analysis is used to evaluate effectively how the model distinguishes between the two diseases. In the below Table III, it is clearly shown that the developed AD-DLB-DNet system, in terms of accuracy, is better in performance than the existing methods such as RAN, SVM, CNN-GCN, and CNN by 4.27%, 2.85%, 3.26%, and 1.988%, respectively. AD-DLB-DNet consistently outperforms the other models with the highest accuracy of 92.104%, specificity of 97.215, F1 Score of 85.361, MCC value 0.804, and CSI value 74.461, while also maintaining the lowest FPR of 2.785 and FDR of 7.917 at K-Fold-5. This suggests that AD-DLB-DNet is the most robust model for differentiating between DLB and AD. Table III shows the Comparative K-fold analysis of the developed framework.

TABLE III. COMPARATIVE K-FOLD ANALYSIS OF THE PROPOSED METHOD WITH EXISTING METHODS

TERMS	RAN [30]	SVM [2]	CNN-GCN [3]	CNN [7]	AD-DLB-DNet
K-Fold-1					
Accuracy	87.750	88.958	88.604	90.604	91.500
Specificity	95.664	96.013	95.910	96.680	96.969
FPR	4.336	3.987	4.090	3.320	3.031
FDR	11.917	11.083	11.333	9.333	8.583
F1	78.238	80.105	79.551	82.832	84.320
MCC	0.707	0.733	0.725	0.770	0.790
CSI	64.255	66.813	66.046	70.695	72.890
K-Fold-2					

Accuracy	84.708	87.688	85.750	89.146	91.833
Specificity	94.338	95.494	94.724	96.134	97.121
FPR	5.662	4.506	5.276	3.866	2.879
FDR	15.250	12.417	14.333	10.750	8.167
F1	73.483	78.054	75.036	80.436	84.900
MCC	0.641	0.704	0.663	0.737	0.798
CSI	58.081	64.007	60.047	67.274	73.762
K-Fold-3					
Accuracy	85.896	87.021	86.563	88.583	90.708
Specificity	94.790	95.450	95.087	95.826	96.740
FPR	5.210	4.550	4.913	4.174	3.260
FDR	14.167	12.417	13.417	11.583	9.167
F1	75.265	77.138	76.313	79.476	83.016
MCC	0.666	0.692	0.680	0.724	0.772
CSI	60.340	62.784	61.698	65.942	70.964
K-Fold-4					
Accuracy	85.500	85.813	83.771	89.229	89.229
Specificity	94.732	94.701	94.009	96.139	96.167
FPR	5.268	5.299	5.991	3.861	3.833
FDR	14.250	14.417	16.000	10.750	10.667
F1	74.728	75.101	72.129	80.557	80.571
MCC	0.659	0.664	0.622	0.739	0.739
CSI	59.652	60.129	56.407	67.443	67.464
K-Fold-5					
Accuracy	86.500	88.667	87.854	90.500	92.104
Specificity	95.082	95.969	95.643	96.647	97.215
FPR	4.918	4.031	4.357	3.353	2.785
FDR	13.417	11.167	12.000	9.417	7.917
F1	76.229	79.671	78.367	82.662	85.361
MCC	0.679	0.727	0.709	0.767	0.804
CSI	61.589	66.211	64.430	70.447	74.461

VIII. CONCLUSION

This study aimed to differentiate DLB from AD using a deep learning model. The process began with collecting the necessary images from a benchmark dataset. The images were fed into the SA-ACNN for CIS feature extraction. Using Grad-CAM, the deep learning model not only provides accurate predictions for distinguishing DLB from AD but also offers a clear visual representation of the CIS as an essential feature for DLB diagnosis. This visualization technique enhances the model's interpretability, fostering greater trust in its decision-making process. To perform differentiation, a DR-LSTM was proposed, which effectively identified clinical features. This comprehensive model aimed to enhance the differential diagnosis of DLB as well as AD, facilitating more effective therapeutic measures. Finally, validation steps were performed to confirm the efficacy of the method that ensures its reliability in clinical settings. The accuracy of the developed AD-DLB-DNet framework is more effective than RAN, SVM, CNN-GCN, and CNN by 8.41%, 4.72%, 7.09%, and 3.01%, respectively at the k-fold value to be 2. The suggested model is used for the differentiation of DLB and AD for taking effective therapeutic measures. The present study faces limitations such as data constraints, generalization challenges, and the need for extensive clinical validation to ensure reliability and ethical compliance. In future work, several avenues can be explored to improve the differentiation of DLB and AD utilizing deep learning models. Expanding the dataset to include more diverse images from various demographics and medical conditions can improve model robustness.

REFERENCES

[1] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection," Brain Inform, vol. 11, no. 1, p. 10, Dec. 2024, doi: 10.1186/s40708-024-00222-1.

[2] B. Lei et al., "Hybrid federated learning with brain-region attention network for multi-center Alzheimer's disease detection," Pattern

Recognit, vol. 153, p. 110423, Sep. 2024, doi: 10.1016/j.patcog.2024.110423.

[3] N. Pradhan, S. Sagar, and A. S. Singh, "Analysis of MRI image data for Alzheimer disease detection using deep learning techniques," Multimed Tools Appl, vol. 83, no. 6, pp. 17729–17752, Jul. 2023, doi: 10.1007/s11042-023-16256-2.

[4] S. M. Mahim et al., "Unlocking the Potential of XAI for Improved Alzheimer's Disease Detection and Classification Using a ViT-GRU Model," IEEE Access, vol. 12, pp. 8390–8412, 2024, doi: 10.1109/ACCESS.2024.3351809.

[5] M. Trinh, R. Shahbaba, C. Stark, and Y. Ren, "Alzheimer's disease detection using data fusion with a deep supervised encoder," Frontiers in Dementia, vol. 3, Feb. 2024, doi: 10.3389/frdem.2024.1332928.

[6] D. M. O'Shea et al., "Practical use of DAT SPECT imaging in diagnosing dementia with Lewy bodies: a US perspective of current guidelines and future directions," Front Neurol, vol. 15, Apr. 2024, doi: 10.3389/fneur.2024.1395413.

[7] M. J. Plastini et al., "Multiple biomarkers improve diagnostic accuracy across Lewy body and Alzheimer's disease spectra," Ann Clin Transl Neurol, vol. 11, no. 5, pp. 1197–1210, May 2024, doi: 10.1002/acn3.52034.

[8] N. S. Sjaelland, M. H. Gramkow, S. G. Hasselbalch, and K. S. Frederiksen, "Digital Biomarkers for the Assessment of Non-Cognitive Symptoms in Patients with Dementia with Lewy Bodies: A Systematic Review," Journal of Alzheimer's Disease, vol. 100, no. 2, pp. 431–451, Jul. 2024, doi: 10.3233/JAD-240327.

[9] J. Levin et al., "α - Synuclein seed amplification assay detects Lewy body co - pathology in autosomal dominant Alzheimer's disease late in the disease course and dependent on Lewy pathology burden," Alzheimer's & Dementia, vol. 20, no. 6, pp. 4351‐4365, Jun. 2024, doi: 10.1002/alz.13818.

[10] K. Sravani and V. RaviSankar, "Intelligent Differentiation Framework for Lewy Body Dementia and Alzheimer's disease using Adaptive Multi-Cascaded ResNet–Autoencoder–LSTM Network," Int J Image Graph, Apr. 2024, doi: 10.1142/S0219467825500664.

[11] H. Sohrabnavi, M. Mohammadimasoudi, and H. Hajghassem, "Early detection of Alzheimer's disease by measuring amyloid beta-42 concentration in human serum based on liquid crystals," Sens Actuators B Chem, vol. 401, p. 134966, Feb. 2024, doi: 10.1016/j.snb.2023.134966.

[12] Y. Zeng, Z. Huang, Y. Liu, and T. Xu, "Printed Biosensors for the Detection of Alzheimer's Disease Based on Blood Biomarkers," J Anal Test, vol. 8, no. 2, pp. 133–142, Jun. 2024, doi: 10.1007/s41664-023-00277-9.

[13] M. J. Armstrong, D. J. Irwin, J. B. Leverenz, N. Gamez, A. Taylor, and J. E. Galvin, "Biomarker Use for Dementia With Lewy Body Diagnosis," Alzheimer Dis Assoc Disord, vol. 35, no. 1, pp. 55–61, Jan. 2021, doi: 10.1097/WAD.0000000000000414.

[14] S. Siuly, Ö. F. Alçin, H. Wang, Y. Li et al., "Exploring Rhythms and Channels-Based EEG Biomarkers for Early Detection of Alzheimer's Disease," IEEE Trans Emerg Top Comput Intell, vol. 8, no. 2, pp. 1609–1623, Apr. 2024, doi: 10.1109/TETCI.2024.3353610.

[15] B. TaghiBeyglou and F. Rudzicz, "Context is not key: Detecting Alzheimer's disease with both classical and transformer-based neural language models," Natural Language Processing Journal, vol. 6, p. 100046, Mar. 2024, doi: 10.1016/j.nlp.2023.100046.

[16] I. Bazarbekov, A. Razaque, M. Ipalakova et al., "A review of artificial intelligence methods for Alzheimer's disease diagnosis: Insights from neuroimaging to sensor data analysis," Biomed Signal Process Control, vol. 92, p. 106023, Jun. 2024, doi: 10.1016/j.bspc.2024.106023.

[17] J. Therriault et al., "Comparison of immunoassay- with mass spectrometry-derived p-tau quantification for the detection of Alzheimer's disease pathology," Mol Neurodegener, vol. 19, no. 1, p. 2, Jan. 2024, doi: 10.1186/s13024-023-00689-2.

[18] T. Nakata et al., "Differential diagnosis of MCI with Lewy bodies and MCI due to Alzheimer's disease by visual assessment of occipital hypoperfusion on SPECT images," Jpn J Radiol, vol. 42, no. 3, pp. 308–318, Mar. 2024, doi: 10.1007/s11604-023-01501-3.

[19] S. M. S. Karim, M. S. Fahad, and R. S. Rathore, "Identifying discriminative features of brain network for prediction of Alzheimer's disease using graph theory and machine learning," Front Neuroinform, vol. 18, Jun. 2024, doi: 10.3389/fninf.2024.1384720.

[20] M. E. Hasan and A. Wagler, "New Convolutional Neural Network and Graph Convolutional Network-Based Architecture for AI Applications in Alzheimer's Disease and Dementia-Stage Classification," AI, vol. 5, no. 1, pp. 342–363, Feb. 2024, doi: 10.3390/ai5010017.

[21] K. Etminani et al., "A 3D deep learning model to predict the diagnosis of dementia with Lewy bodies, Alzheimer's disease, and mild cognitive impairment using brain 18F-FDG PET," Eur J Nucl Med Mol Imaging, vol. 49, no. 2, pp. 563–584, Jan. 2022, doi: 10.1007/s00259-021-05483-0.

[22] L. Gjerum et al., "A visual rating scale for cingulate island sign on 18F-FDG-PET to differentiate dementia with Lewy bodies and Alzheimer's disease," J Neurol Sci, vol. 410, p. 116645, Mar. 2020, doi: 10.1016/j.jns.2019.116645.

[23] H. Kanetaka et al., "Differentiating Mild Cognitive Impairment, Alzheimer's Disease, and Dementia With Lewy Bodies Using Cingulate Island Sign on Perfusion IMP-SPECT," Front Neurol, vol. 11, Nov. 2020, doi: 10.3389/fneur.2020.568438.

[24] B. Y. Lim et al., "Deep Learning Model for Prediction of Progressive Mild Cognitive Impairment to Alzheimer's Disease Using Structural MRI," Front Aging Neurosci, vol. 14, Jun. 2022, doi: 10.3389/fnagi.2022.876202.

[25] I. G. McKeith et al., "Diagnosis and management of dementia with Lewy bodies," Neurology, vol. 89, no. 1, pp. 88–100, Jul. 2017, doi: 10.1212/WNL.0000000000004058.

[26] S. Basheera and M. Satya Sai Ram, "A novel CNN based Alzheimer's disease classification using hybrid enhanced ICA segmented gray matter of MRI," Computerized Medical Imaging and Graphics, vol. 81, p. 101713, Apr. 2020, doi: 10.1016/j.compmedimag.2020.101713.

[27] C. Li, H. Zhang, Z. Wang, Y. Wu, and F. Yang, "Spatial-Temporal Attention Mechanism and Graph Convolutional Networks for Destination Prediction," Front Neurorobot, vol. 16, Jul. 2022, doi: 10.3389/fnbot.2022.925210.

[28] H. Wang et al., "A Residual LSTM and Seq2Seq Neural Network Based on GPT for Chinese Rice-Related Question and Answer System," Agriculture, vol. 12, no. 6, p. 813, Jun. 2022, doi: 10.3390/agriculture12060813.

[29] C. Chen, X. Lin, and G. Terejanu, "An Approximate Bayesian Long Short- Term Memory Algorithm for Outlier Detection," in 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, Aug. 2018, pp. 201–206. doi: 10.1109/ICPR.2018.8545695.

[30] C. Tian, X. Zhu, Z. Hu, and J. Ma, "Deep spatial-temporal networks for crowd flows prediction by dilated convolutions and region-shifting attention mechanism," Applied Intelligence, vol. 50, no. 10, pp. 3057–3070, Oct. 2020, doi: 10.1007/s10489-020-01698-0.

[31] J. O. Agushaka, A. E. Ezugwu, and L. Abualigah, "Dwarf Mongoose Optimization Algorithm," Comput Methods Appl Mech Eng, vol. 391, p. 114570, Mar. 2022, doi: 10.1016/j.cma.2022.114570.

[32] J. Xue and B. Shen, "A novel swarm intelligence optimization approach: sparrow search algorithm," Systems Science & Control Engineering, vol. 8, no. 1, pp. 22–34, Jan. 2020, doi: 10.1080/21642583.2019.1708830.

[33] J. H. Almazán-Covarrubias, H. Peraza-Vázquez, A. F. Peña-Delgado, and P. M. García-Vite, "An Improved Dingo Optimization Algorithm Applied to SHE-PWM Modulation Strategy," Applied Sciences, vol. 12, no. 3, p. 992, Jan. 2022, doi: 10.3390/app12030992.

[34] Z. Zhang, C. Huang, K. Dong, and H. Huang, "Birds foraging search: a novel population-based algorithm for global optimization," Memet Comput, vol. 11, no. 3, pp. 221–250, Sep. 2019, doi: 10.1007/s12293-019-00286-1.

[35] A. Behera, Z. Wharton, Y. Liu, M. Ghahremani et al., "Regional Attention Network (RAN) for Head Pose and Fine-Grained Gesture Recognition," IEEE Trans Affect Comput, vol. 14, no. 1, pp. 549–562, Jan. 2023, doi: 10.1109/TAFFC.2020.3031841.

Energy-Balance-Based Out-of-Distribution Detection of Skin Lesions

Jiahui Sun[1], Guan Yang[2*], Yishuo Chen[3], Hongyan Wu[4], Xiaoming Liu[5]
School of Computer Science, Zhongyuan University of Technology, Zhengzhou, Henan 450007, China[1, 2]
School of Artificial Intelligence, Zhongyuan University of Technology, Zhengzhou, Henan 450007, China[1, 2, 5]
Henan Key Laboratory on Public Opinion Intelligent Analysis, Zhengzhou, Henan 450007, China[1, 2, 5]
School of Textiles, Zhongyuan University of Technology, Zhengzhou, Henan 450007, China[3, 4]

***Abstract*—Skin lesion detection plays a crucial role in the diagnosis and treatment of skin diseases. Due to the wide variety of skin lesion types, especially when dealing with unknown or rare lesions, models tend to exhibit overconfidence. Out-of-distribution (OOD) detection techniques are capable of identifying lesion types that were not present in the training data, thereby enhancing the model's robustness and diagnostic reliability. However, the issue of class imbalance makes it difficult for models to effectively learn the features of minority class lesions. To address this challenge, a Balanced Energy Regularization Loss is proposed in this paper, aimed at mitigating the class imbalance problem in OOD detection. This method applies stronger regularization to majority class samples, promoting the model's learning of minority class samples, which significantly improves model performance. Experimental results demonstrate that the Balanced Energy Regularization Loss effectively enhances the model's robustness and accuracy in OOD detection tasks, providing a viable solution to the class imbalance issue in skin lesion detection.**

***Keywords*—Balanced energy regularization loss; skin lesions; out-of-distribution detection; convolutional neural networks**

I. INTRODUCTION

Early detection and regular monitoring of skin lesions, one of the most common diseases in daily life, is of great importance; this not only helps to improve cure rates and develop accurate treatment plans but also effectively reduces mortality [1]. Especially for melanoma, the most lethal form of skin cancer, this importance is particularly pronounced in study [2]. When using convolutional neural networks (CNNs) for skin lesion detection, only data with known specific distributions are exposed during training; however, due to the characteristics of skin lesions, including the diversity and complexity of their presentation [3], the actual distribution of data in the clinical setting is often uncertain. This poses a challenge with limited training data, which is often insufficient to fully cover the variety of skin lesions encountered. In addition, the distributions of the training data and the actual clinical data may differ significantly, further complicating the task. The predictive performance of the model is greatly reduced when faced with the significant challenges posed by different data distributions [4]. By employing an effective out-of-distribution (OOD) detection method, the model can identify new data that are different from the distribution of the training data. This detection mechanism enables the model to perform the special treatment or directly reject these OOD data for prediction, avoiding making wrong judgments on uncertain data, thus significantly improving the robustness and safety of the model [4] [5]. Effective OOD detection not only enhances the generalization ability of the model but also improves the reliability and robustness of the system in practical applications [6] [7]. Therefore, an effective tool in the diagnosis and management of skin lesions will be methods that can accurately detect OOD images of skin lesions.

In recent years, CNNs have made significant advances in the use of medical image data for disease diagnosis and analysis [8], and these models demonstrate performance comparable to that of professional physicians in the identification and classification of a wide range of common skin lesions [9], especially in binary and multiclassification tasks [10]. Specifically, CNNs can accurately differentiate between malignant melanoma, basal cell carcinoma, or other types of skin lesions [11]. However, OOD detection remains a challenging problem when confronted with skin lesions of unknown characteristics. Furthermore, for assessing the performance of models in different datasets, cross-dataset validation has not yet been widely applied to the OOD detection of skin lesions, which is crucial to ensure the generalizability and validity of the models given the potential differences in the data information in different datasets.

CNNs achieve great success in natural language processing and image recognition tasks, mainly due to their excellent feature learning, large-scale data processing, and generalization capabilities [12] [13]. Although the use of techniques such as self-attention mechanisms [14], regularization [15], and transfer learning [16] can significantly improve the performance of a model, their performance in OOD detection may still be unsatisfactory. This is because the features of the OOD samples are significantly different from the features of the distribution of the training data, which causes the model to be prone to prediction errors when dealing with these samples, thus reducing the robustness and reliability of the model. Therefore, a method called Balanced Energy Regularization Loss (BERL) [17] is applied to CNNs for OOD detection of skin lesions; The model is named energy-balance based OOD detection (EBOD).

In skin lesion image datasets, the class distribution is often highly imbalanced, with some lesion categories having a large number of samples, while others have relatively few. Traditional OOD detection methods struggle to effectively handle the disparity between majority and minority categories under such imbalanced distributions. The BERL method addresses this issue by introducing regularization based on the prior class

*Corresponding Author

probabilities, which enhances the regularization of majority class samples, thereby allowing the model to focus more on minority class samples and improving the detection accuracy for these categories. This method effectively mitigates the detection bias caused by class imbalance and optimizes the overall OOD detection performance. During the training process, the prior probability of each class is computed, allowing the model to adjust for the class distribution imbalance, thereby improving detection accuracy. Particularly in high-dimensional image datasets, the BERL enhances the model's robustness to unseen skin lesions, providing strong support for OOD detection of skin lesions. The successful application of this method not only generates significant impacts in skin lesion detection but also provides a new perspective for OOD detection in other medical imaging tasks. Through this technology, medical imaging systems are enabled to more accurately identify emerging disease types or lesions, thereby enhancing the intelligence and precision of medical diagnosis.

The remainder of this paper is structured as follows: Section II reviews the research advancements in the relevant field; Section III provides a detailed description of the proposed model's architecture, data acquisition and preparation process, as well as the evaluation metrics; Section IV discusses the model evaluation, computational cost, and presents the experimental results, accompanied by a comparative analysis with existing methods; Section V summarizes the main contributions of this paper, clarifies the motivation and potential advantages of the proposed method, and discusses the limitations of the research; Section VI outlines the directions for future research.

II. Related Work

During the past few years, a variety of methods based on CNNs have emerged in the field of OOD detection. These methods are not only innovative in theory but also show excellent performance in practical applications. These methods can be broadly classified into the following groups: output score-based methods, generative model-based methods, adversarial training-based methods, and feature space-based methods, according to their basic principles and application characteristics.

A. Methodology Based on Output Scores

Output score-based methods rely heavily on the output probability distribution of the classifier for the detection of OOD samples. This type of approach works by analyzing the confidence level of the classifier as it processes the samples, and samples with a low confidence level are considered to be possible OOD samples. Hendrycks and Gimpel [5] propose this method, which is widely used due to its simplicity and low computational cost. However, in some cases, such as skin lesion OOD detection, certain OOD samples may have higher softmax values, resulting in detection errors.

B. Generative Modeling-Based Methods

Generative models detect OOD samples by learning the latent distribution of the data, and the variation autoencoder (VAE) method proposed by An and Cho [18] is a typical example. The VAE detects samples that do not match the distribution of the training data by reconstructing the data. This method is particularly suitable for OOD detection of medical image data and can identify rare or unseen lesion types.

C. Adversarial Training-Based Methods

The adversarial training-based approach utilizes Generative Adversarial Networks (GANs) for OOD detection. The method proposed by Schlegl, Seeböck, and Waldstein [19] generates samples that are similar to normal data employing GANs and identifies OOD samples utilizing reconstruction errors. This approach significantly improves the sensitivity of the model to OOD samples and is well-suited for application in complex clinical settings.

D. Feature Space-Based Methods

Feature space-based methods include the Mahalanobis distance-based method introduced by Lee et al. [20] and the One-Class Support Vector Machine (One-Class SVM) method developed by Schölkopf et al. [21]. The former identifies OOD samples by calculating the Mahalanobis distance of the input data in the feature space, which is suitable for feature extraction of high-dimensional data, but requires careful tuning of the distance metric and high computational cost. The latter separates most of the training data by constructing hyperplanes or decision boundaries to distinguish between normal and abnormal data, and is suitable for initial screening for OOD detection of skin lesions, but may not perform well on large and complex datasets.

In addition, several studies explore ways to enhance OOD detection by integrating multiple models. Xu et al. [22] present a deep integrated learning approach to enhance the accuracy and robustness of detection by combining the predictions of multiple deep neural network models. Dai et al. [23] designed a multimodal detection method utilizing multiple data sources (e.g., images, text, and clinical data) to improve the detection performance, and the combination of patient history, symptom descriptions, and image data in OOD detection of skin lesions can significantly improve the accuracy of OOD detection.

III. Methodology

This section describes the three modules of the experiment in this study: data acquisition, model architecture, and model evaluation. The first section outlines the methodological steps adopted by the researchers in the data collection and analysis process, which is the core part of the study. The next subsections explain the specific steps of the study in terms of model architecture design and model evaluation, respectively. The experimental workflow shown in Fig. 1 provides a clear overview of the experimental process.

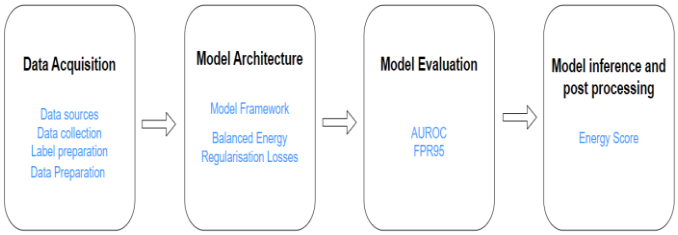


Fig. 1. Experiment module diagram.

A. Data Acquisition

1) Data sources: The medical datasets used in this study are obtained from open-source databases, including but not limited to the International Skin Imaging Collaboration (ISIC), the Harvard University database (https://data.harvard.edu.dataverse), the Portuguese Pedro Hispano Hospital Dermoscopy Image Database (https://www.fc.up.pt/addi/project.html), and Stanford AIMI Shared Database (https://stanfordaimi.azurewebsites.net). These databases provide rich and diverse images of skin lesions for this study, ensuring broad applicability and reliability for model training and evaluation. To validate the external generalization ability of the model, several datasets of non-skin lesions are also obtained from the Kaggle platform (www.kaggle.com) for OOD detection. In terms of data use, this study strictly follows ethical principles to ensure that patient privacy is adequately protected, data security is effectively guaranteed, and patients' rights are respected. At the same time, this study actively promotes data sharing and open scientific research and assumes corresponding responsibilities and obligations.

2) Data collection: In this study, several open-source skin lesion datasets are used. These include images of different types of skin lesions as well as images of normal skin from different locations with manually created or corrected annotation information. From the ISIC2018, ISIC2019, ISIC2020, HAM10000 [24], PH2 [25], DDI [26], Dermnet [27], UMCG [28], and PAD-UFES-20 [29] datasets, several datasets containing multiple lesion types are screened. The prevalence of various skin lesions varies due to differences in the number of images of various lesion types in different datasets. In addition, abdominal MRI, brain tumor, kidney stone, and Places365 datasets are obtained from the Kaggle platform. These are used as auxiliary datasets [30] along with the skin lesion datasets for model training. Table I provides a summary of the fundamental characteristics of the datasets employed in this study. To ensure the breadth of the training datasets, 13 datasets containing four diseases and one non-disease are used during model training. For OOD data, three other disease datasets and two object detection datasets are used in this study: the colon adenocarcinoma dataset, the gastrointestinal disease dataset, the cataract dataset, the Street View House Numbers (SVHN) dataset [31], and the Cifar10 dataset [32].

TABLE I. BASIC CHARACTERISTICS OF EACH SKIN LESION DATASET

Dataset	Count of Types	Number of Photos	Source
ISIC2018	7	11720	ISIC
ISIC2019	8	25331	ISIC
ISIC2020	5	33126	ISIC
HAM10000	7	10015	Harvard Dataset
PH2	2	200	PH2 Dataset
Dermnet	5	579	Kaggle
PAD-UFES-20	5	2298	GitHub
UMCG	2	170	MED-NODE Dataset
DDI	6	656	Stanfordaimi AIMI

3) Label preparation: In the many open-source datasets on skin lesions, many different types of skin lesions are usually covered. However, the types of skin lesions that are recorded in the different datasets are not the same. It is not possible to train directly with these raw datasets as the model needs to be trained by lesion type for classification when performing training. For example, the ISIC2019 database contains eight different types of skin lesions. The HAM10000 database [24] contains seven types, and the lesion types differ between the two. To solve this problem, we use a strategy that requires a two-stage labeling process. Firstly, skin lesion types are classified according to the label files in each data set. When the same lesion is encountered but in different locations, it is combined into the same lesion type, and the original label file is reordered to generate a new label file based on the lesion type. Subsequently, the images in each of the datasets are then retrieved and organized according to these new label files. By using this method, each data set is divided into sub-datasets that contain multiple types of lesions. Finally, all datasets are regrouped and fused by skin lesion type to better support model training. Images of each type of skin lesion are shown in Fig. 2.

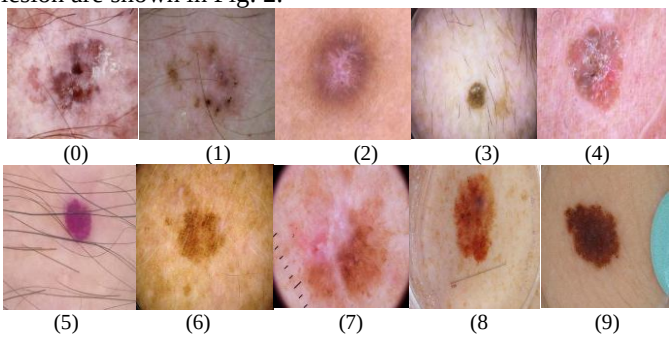


Fig. 2. Images of each type of skin lesion. 0: actinic keratosis; 1: basal cell carcinoma; 2: dermatofibroma; 3: seborrheic keratosis; 4: benign keratosis; 5: vascular lesions; 6: freckles; 7: squamous cell carcinoma; 8: melanoma; 9: melanocytic nevus.

4) Data preparation: All datasets used in this study show significant heterogeneity [33], involving differences in lesion characteristics, lesion sites, and recording devices. To reduce the impact of the differences between these datasets on the model, appropriate corrective measures are taken to standardize the datasets to ensure their compatibility with the model. When determining the size of the input image, the standard 224 × 224 size is selected. If the dimensions of the input image exceed 224 × 224, the image undergoes a process of cropping and scaling, referred to as center crop scaling [34], in order to align with the specifications of the model. The rationale behind the selection of this size is that image dimensions fluctuate across the datasets, rendering a uniform input size conducive to the model's capacity to discern pivotal characteristics. This uniformity simplifies data manipulation, reduces the complexity of computational operations, accelerates the convergence process, enhances the model's capacity to generalize, and facilitates the improvement of the model's performance. In both steps of training and detecting the model,

the images are cropped and the newly generated data obtained from cropping is used for training and detection.

B. Model Architecture

1) Model framework: The current model, similar to other models for OOD detection, uses neural networks to design appropriate OOD detection scores [4] [5] [35]. However, unlike most previous OOD detection methods that focus on designing OOD scores or introducing multiple outlier samples to retrain the model [36] [37], this study delves into the obstacle factors in OOD detection from the perspective of class imbalance in the auxiliary datasets [38]. To address the imbalance problem, a BERL is used, with different regularizations for each category of auxiliary data, to achieve reliable uncertainty estimates.

The training model employed in this study is primarily ResNet18 [39]. In comparison with other neural network models, ResNet18 demonstrates notable advantages in terms of feature extraction and generalizability. ResNet18 is a relatively deep network architecture that is capable of learning more abstract and complex feature representations through multi-level convolutional operations and feature extraction. This allows ResNet18 to extract more information from skin lesion data. In addition, ResNet18 is connected in a way that helps mitigate the problem of vanishing gradients and allows the network to be deeper. In the field of skin lesion OOD detection, residual connectivity and CNNs are both useful in enhancing the efficiency of feature capturing in images, which in turn leads to improved performance and greater model generalization. Although ResNet18 itself is not specifically designed for the detection of OOD, it can be appropriately adapted and enhanced to make it applicable to the detection of OOD [40]. ResNet18, with its deep convolutional architecture, is capable of effectively capturing latent patterns within skin lesion images, including both the intricate details and the overall morphology of the lesions. These patterns aid the model in distinguishing between in-distribution (known) and OOD (unknown) samples. During training, the residual connections in ResNet18 effectively alleviate the issue of overfitting, particularly when dealing with complex textures or lesion structures. This enables the extraction of robust and discriminative features, thereby enhancing the performance of OOD detection. Specifically, ResNet18 learns multi-scale features through convolution operations at different layers, from fine details to global representations. At lower layers, the network is capable of identifying minute textures and details on the skin surface, while at higher layers, it can capture the broader shapes of larger skin lesion areas. This hierarchical feature learning enables ResNet18 to accurately identify unknown lesion types in OOD detection tasks and effectively avoid misclassifying them as known types. The restrained ResNet18 model extracts feature from the image and feeds those features into a separate classifier to map the features of the image into a specific space and use the classification boundaries in that space to distinguish between known and unknown data. Finally, an energy function is introduced into the model for the calculation of OOD scores. Since the energy function does not require labeling information between known and unknown data, OOD detection can be performed without unknown data labels. In addition, energy functions usually have a good generalization ability to deal with different types of unknown data and to establish reasonable boundaries between the known data and the unknown data. The model structure is schematically shown in Fig. 3.

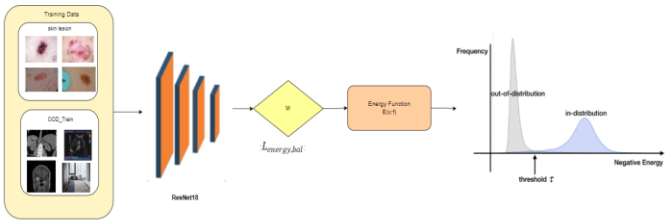


Fig. 3. Schematic diagram of the balanced energy regularization loss OOD detection model.

2) Balanced energy regularization loss: When regularizing auxiliary data, due to the imbalance of its class distribution, this may result in the model not being able to effectively learn information about the data of a few classes, thus affecting the model's ability to generalize to OOD data. In order to solve this problem, a variable M is introduced to measure whether the sample of the auxiliary data belongs to the majority class or the minority class [17]. In addition, the prior probabilities of the distributions of the auxiliary data were used to determine which class was to be categorized as a minority class. When this model is used to make inferences on the OOD auxiliary data, a statistical value N_i can be obtained that indicates the number of samples that are classified in class i. Then, the prior probability of the OOD distribution can be calculated using the following formula:

$$P(y=i \mid o)=\frac{N_i}{N_1+N_2+\cdots+N_K}. \tag{1}$$

For a neural network classifier f, the a posteriori probability that the input image x belongs to a class i is acquired by performing a softmax operation on the production of f, which is,

$$P(y=i \mid \mathbf{x},o)=\frac{e^{f_i(\mathbf{x})}}{\Sigma_{j=1}^{K} e^{f_j(\mathbf{x})}}. \tag{2}$$

As the posterior probability that $\mathbf{x}$ belongs to the class i increases, the probability that $\mathbf{x}$ belongs to the class i increases accordingly. Similarly, if the prior probability of category i is higher, then the probability of category i becoming the majority category will increase. Thus, as the probability that $\mathbf{x}$ belongs to the majority class i increases, the product of $\mathrm{P}(y=i \mid o)$ and $\mathrm{P}(y=i \mid \mathbf{x},o)$ must increase. Based on this result, the metric M for measuring the probability that $\mathbf{x}$ belongs to the majority class is defined as follows:

$$M=\Sigma_{j=1}^{K} P(y=j \mid \mathbf{x},o)P(y=j \mid o). \tag{3}$$

Moreover, a hyperparameter λ is introduced to model an additional generalized prior probability, which is used to

regulate the degree of prior difference between categories. Ultimately, the generalized form M_λ is as follows:

$$M_\lambda = \Sigma_{j=1}^{K} P(y=j \mid \mathbf{x}, o) P_\lambda(y=j \mid o). \quad (4)$$

Where $P_\lambda(y=i \mid o) = L^1 norm\{ P^\lambda(y=i \mid o) \}$. In order to ensure numerical reliability, the λ -th power operation was performed on the a priori probability $\mathrm{P}(y=i \mid o)$, and the L1-normalization was applied [41]. When $\lambda = 0$, the uniform prior probability model is established and M_λ becomes a constant value $\frac{1}{K}$, resulting in equal regularization strength across samples of different classes during the regularization process. When $\lambda > 0$, M_λ amplifies the regularization strength for the majority class while reducing it for the minority class, thus directing the model's focus more towards enhancing the OOD detection capability of majority class samples. In contrast, when $\lambda < 0$, the inverse distribution of the prior probability will be modeled, and the regularization strength for the minority class samples is increased, allowing the model to better adapt to OOD samples from the minority class. The optimal value λ varies across different OOD datasets. Generally, a larger value λ indicates a greater prior difference between classes, which is more suitable for datasets dominated by the majority class. Conversely, smaller or negative values λ are better suited for scenarios where the class distribution is more balanced or where the minority classes are of greater importance. With λ growth, the prior probability gap among classes grows. According to the M_λ component, the BERL is given by:

$$\begin{aligned} L_{\text{energy,bal}} &= L_{in,hinge} + L_{out,bal} \\ &= E_{(\mathbf{x}_{in},y)\sim D_{in}^{train}} \left[\left(\max\left(0, E(\mathbf{x}) - m_{in}\right) \right)^2 \right] \\ &+ E_{\mathbf{x}\sim D_{in}^{train}} \left[\left(\max\left(0, E(\mathbf{x}) - m_{out} - \alpha M_\lambda\right) \right)^2 M_\lambda \right] \end{aligned} \quad (5)$$

where $E(x; f) = -T \cdot \log(\Sigma_{j=1}^{K} e^{f_j(\mathbf{x})/T})$. In this formula, $f_j(x)$ denotes the logit output of the model for the input sample x in class j, while T representing the temperature parameter, which is employed to smooth the logits distribution, By performing an exponentially weighted summation of the logits across all classes, the energy value for each sample is computed. Lower energy values are generally associated with OOD samples, whereas higher energy values are typically indicative of in-distribution samples. The energy loss $L_{\text{energy,bal}}$ of the model is the sum of both $L_{in,hinge}$ and $L_{out,bal}$.

The key characteristic of the energy function lies in its ability to compute an energy value for each sample that is associated with its corresponding class. For in-distribution samples, the energy values are typically low because the predictions for these samples are usually accurate and consistent with the distribution of the training data. In contrast, the energy values for OOD samples are generally higher, as they do not conform to the distribution of the training data, thereby reflecting the greater uncertainty the model has regarding these samples. For instance, consider a pre-trained model designed to classify cats, dogs, and birds. When an image of a car is input, the model's logit output tends to be more dispersed, resulting in a higher energy value, which indicates that the model has lower confidence in classifying this sample. Conversely, when an image of a cat is input, the model's logit output is more concentrated, leading to a lower energy value, thus demonstrating the model's higher confidence in classifying this sample.

C. Model Evaluation

The key metrics that researchers typically focus on when performing model evaluations include Recall, the area under the receiver operating characteristic curve (AUROC), and the false positive rate at 95% true positive rate (FPR95). Recall is used to measure the proportion of OOD samples that are correctly detected by the model out of all actual OOD samples.

1) AUROC: AUROC, on the other hand, represents the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR), calculated as the area under the ROC curve, which can synthesize the performance of the model under different classification thresholds. The value of AUROC ranges from 0.5 to 1. The closer the value is to 1, the better the ability of the model to discriminate.

2) FPR95: On the other hand, FPR95 focuses on the false alarm situation under high recall conditions, specifically calculating the false positive rate while maintaining a 95% true positive rate. FPR95, as a performance evaluation index, reflects the ability of the model to control the false alarm rate under the premise of guaranteeing a high detection rate in practical applications, and the smaller its value is, the lower the false alarm rate of the model is, thus proving the better performance of OOD detection.

Therefore, in this study, AUROC and FPR95 are used as the core metrics to assess the reliability of the model in OOD image detection of skin lesions. The FPR95 evaluates the ability to achieve high recall while controlling for false positives, while the AUROC provides an overall performance evaluation showing the average performance of the model across all thresholds. Recall as a base metric plays an important role in the calculation of AUROC and FPR95 as it has a direct impact on the results and performance analysis of these two metrics. The formulas for these indicators are shown below:

$$\mathrm{Re}\,call = TPR = \frac{TP}{TP + FN} \quad (6)$$

$$FPR = \frac{FP}{FP + TN} \quad (7)$$

$$FPR95 = FPR\left(TPR = 0.95\right) \quad (8)$$

$$AUROC = \int_0^1 TPR(x)\, dx \ \left(x = FPR\right) \quad (9)$$

Where TP denotes true positive, TN denotes true negative, FP denotes false positive and FN denotes false negative.

D. Model Inference and Post Processing

Although the model is trained using only the skin lesion datasets, it is equally capable of handling datasets from other diseases. The output generated by the model is an energy score that reflects the difference between the input sample data distribution and the known sample data distribution. Since the model is designed for OOD detection of skin lesions, its output can be interpreted as a measure of the model's classification accuracy in distinguishing between known and unknown samples, i.e., the model's ability to determine whether or not a sample is an OOD. The level of the energy score then indicates how confident the model is in classifying the input samples. Therefore, a post-processing step was used so that when the energy score is low, the model has a higher confidence that the input samples belong to known data, and conversely, when the energy score is high, the model has a higher confidence that the input samples belong to unknown data [42]. The course of the training and testing steps is summarized in Fig. 4.

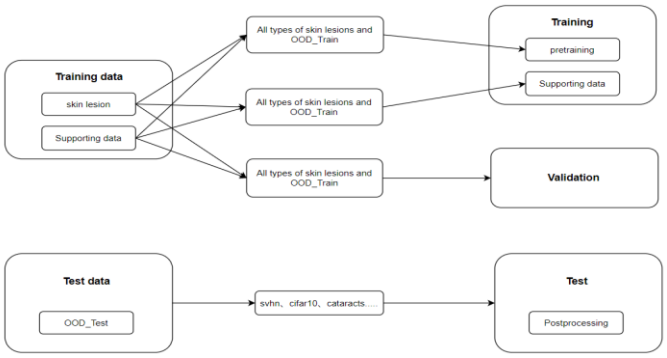


Fig. 4. Overview of training and testing procedures.

In the context of OOD detection for skin lesions, the model may encounter challenging scenarios such as noise, occlusion, and small-sized OOD samples. When confronted with noise and small-sized OOD samples, the application of energy regularization loss results in greater regularization being imposed on these samples, while less regularization is applied to other samples, thereby maintaining the model's ability to effectively detect OOD samples. In the case of occlusion, the integration of an attention mechanism enables the model to adaptively focus on the key regions of interest during training. Furthermore, by adjusting the class distribution of auxiliary data and weighting each sample according to the prior probability of its respective class, the model can effectively control the regularization strength across different classes, ensuring fairness in the data distribution.

IV. Results

A. Performance Evaluation

In order to evaluate the performance of the models, this experiment presents and analyzes the detection results of five models on OOD datasets from multiple domains, both medical and non-medical; the models used in this analysis are MSP [5], OE [43], OECC [44], Energy OOD [45], and EBOD. The FPR95 and AUROC metrics for the five models on each of the OOD data sets are shown in Table II. The experimental results show a slight difference in the performance of EBOD on OOD data in non-medical domains. However, the difference is not significant compared to its performance on OOD data in medical domains. This means that although there may be some differences in the performance of the EBOD model on OOD data from different domains, over the performance is relatively stable and has strong generalization capabilities. Compared to the other four models, the EBOD model shows significant performance gains on most OOD datasets. A sample of the OOD data used in the study is shown in Fig. 5, where colon_aca represents colon adenocarcinoma images, stomach represents gastrointestinal disorders, and cataracts represent cataract images.

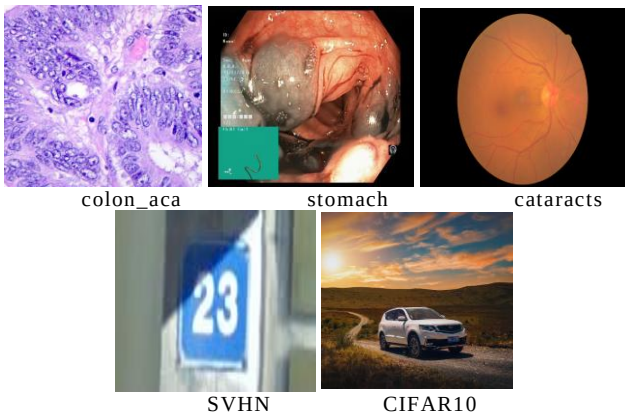


Fig. 5. Sample graph of OOD data.

In exploring the factors that enhance the performance of the model, the significant difference between the EBOD model compared to the WideResNet-based Energy OOD model [45] is the use of BERL for the auxiliary datasets. This innovative approach plays a key role in the optimization process and has a profound impact on the OOD detection performance of the model. A comparison of the FPR95 and AUROC metrics for the Energy OOD and EBOD models reveals that the incorporation of BERL is a pivotal factor in enhancing the OOD detection capabilities of the models. The EBOD model introduces BERL, which serves to balance the energy distribution of the auxiliary datasets during the training process. This significantly improves the model's generalization ability on different datasets. Table II details the results of the comparison of the models on different datasets. The FPR95 metrics for multiple OOD datasets are significantly lower when only the energy model is used and no BERL is introduced.

TABLE II. Evaluation Results of the Model on the OOD Datasets

	MSP	OE	OECC	Energy OOD	EBOD (Ours)
colon_aca	6.28	5.43	4.26	3.61	1.07
stomach	4.76	2.57	0.79	1.21	0.41
cataracts	3.55	1.93	2.87	2.08	0.49
SVHN	43.11	35.41	29.68	36.39	14.03
CIFAR10	45.26	38.71	32.52	29.01	5.85

(a) FPR95

	MSP	OE	OECC	Energy OOD	EBOD (Ours)
colon_aca	88.53	92.81	95.83	95.02	99.36
stomach	89.37	93.46	94.29	92.38	99.85
cataracts	88.61	95.36	99.57	97.54	99.76
SVHN	86.83	90.19	90.68	91.49	95.83
CIFAR10	88.15	91.72	92.64	92.78	98.53

(b) AUROC

B. Calculate Cost

In the practical deployment of OOD detection for skin lesions, the computational cost is a crucial factor that determines model selection and deployment efficiency. To ensure the effectiveness and scalability of OOD detection methods in real-world applications, it is essential to optimize training and inference times, as well as reduce memory consumption, thereby minimizing computational overhead. Through the implementation of effective optimization strategies, the usability of the model in resource-constrained environments can be enhanced while maintaining its accuracy, thus meeting the demands of practical applications. Regarding training time, the BERL function proposed in this paper, compared to traditional methods, mitigates overfitting on minority class samples and excessive training on all class samples by precisely adjusting the regularization strength for each sample. This accelerates model convergence, thus reducing training time. In terms of inference time, the introduction of the M-value to quantify the likelihood of each sample belonging to a specific class, coupled with the adjustment of the loss function based on this value, reduces the computational complexity required for each sample during inference, thereby optimizing inference speed. Table III lists the calculated costs of different skin lesion OOD detection techniques.

TABLE III. CALCULATED COST OF DIFFERENT SKIN LESION OOD DETECTION TECHNIQUES

	Training time (Sheets/ms)	Reasoning time (Sheets/ms)	MEM
La-OOD [46]	4.57	1.46	6.87%
Bayesian [47]	5.13	1.51	7.63%
Ours	3.28	1.32	6.50%

C. Ablation Study

To validate the effectiveness of the proposed method, two ablation experiments were conducted. First, the BERL was introduced into DenseNet [48] for OOD detection of skin lesions. Subsequently, the BERL was removed from the EBOD model (i.e., EBOD-), and the same experiment was repeated. The experimental results demonstrate that the EBOD model exhibits superior performance in the OOD detection of various types of skin lesions, with the specific results summarized in Table IV.

TABLE IV. RESULTS OF ABLATION EXPERIMENTS

	DenseNet	EBOD-	EBOD (Ours)
colon_aca	2.61	4.39	1.07
stomach	1.82	2.07	0.41
cataracts	2.69	1.85	0.49
SVHN	34.50	42.37	14.03
CIFAR10	27.14	16.29	5.85

(a) FPR95

	DenseNet	EBOD-	EBOD (Ours)
colon_aca	92.21	95.77	99.36
stomach	93.09	96.18	99.85
cataracts	95.61	94.80	99.76
SVHN	90.04	89.45	95.83
CIFAR10	91.62	93.37	98.53

(b) AUROC

In contrast to previous studies, few studies synthesize multiple datasets on skin lesions by lesion type and further validate them using OOD datasets from a variety of different domains. Therefore, the present study is based on this innovation and improvement. Firstly, this study employs a methodology analogous to that employed in previous studies, whereby the training and test sets are rationalized in order to ensure a balanced data distribution and representative samples. Furthermore, this study introduces auxiliary datasets, which are employed to facilitate the model's ability to characterize a more expansive range of data, thereby enhancing its performance in the context of novel and previously unseen data. This enhanced generalization capacity enables the model to more effectively adapt to diverse application contexts and data distributions, and remains stable in the presence of noise, outliers, and other disturbances, thereby enhancing the model's resilience.

V. DISCUSSION

This study introduces BERL into CNNs, aiming to accurately detect OOD data in skin lesions. EBOD achieves state-of-the-art performance in cross-database evaluations and demonstrates a high degree of accuracy, even under a wide range of special conditions. In comparing the model performance under different training methods, the introduction of energy balance regularization [17] plays an important role in improving the excellent performance of the model. In addition, EBOD has potential applications in other clinical OOD detection situations.

CNNs produce good results in several areas, such as natural language processing and image recognition [12][13]. In addition, its application is extended to the medical field, bringing great convenience to medical research and clinical practice [49][50]. CNNs have significant potential to improve the accuracy of medical image data analysis, which may have far-reaching implications in the field of medical image diagnosis [51]. The performance of EBOD may be overestimated due to the exclusion of certain types of skin lesions (e.g., common nevi) from the dataset, but the model demonstrates excellent performance in high noise and strong motion environments. Interpreting skin lesions in practical clinical applications is a challenging and complex task. Therefore utilizing this model can potentially reduce the misdiagnosis rate of skin lesions in clinical practice and enhance patient care.

It is well known that the application of energy regularization techniques plays a crucial role in CNNs. Despite the demonstrated efficacy of energy regularization techniques in many other fields, there remains a paucity of research investigating their application to the domain of medical image processing. Inspired by the obstacles in OOD detection caused by a class imbalance in auxiliary datasets and internal mechanisms of the model, this study employs BERL to enhance the performance of CNNs in OOD detection. The introduction of BERL lays the foundation for applying CNNs to OOD

detection of medical images. If OOD detection is performed on multiple categories of medical images in the future, it is possible to achieve more efficient performance. Skin lesions are complex in shape and type, and especially under actual clinical conditions, OOD detection of skin lesions is more important than identification to minimize panic in the minds of patients and the rate of misdiagnosis.

In the case of ancillary data, especially in real-world scenarios, there is often an imbalance in the distribution of classes in the ancillary OOD data, e.g. there is a significant difference in the amount of data in the two classes. It can be difficult to effectively capture the diversity of the auxiliary samples of the OOD data using the traditional methods of cross-entropy loss and regularized loss. The model tends to learn the features of samples from a numerically larger number of categories while ignoring the features of samples from a numerically smaller number of categories, which reduces the model's ability to generalize when dealing with samples from unknown distributions. To overcome this problem, this study uses BERL to apply higher regularization to the data of the more numerous categories in the auxiliary data to ensure that the features of the data of the less numerous categories are also adequately extracted. This approach improves the robustness of the model and makes its performance more stable in the face of a variety of unknown distribution samples.

The evaluation metrics are determined based on the information from each OOD data after it has been tested by the model, which is critical to understanding and measuring the overall performance of the model. Previous research has shown that many detection models tend to produce overly confident prediction results when confronted with OOD data, resulting in less reliable detection of this OOD data. If the skin lesion features are similar in each training dataset, the model may suffer from overfitting, which weakens its generalization ability and leads to poor performance in OOD detection. As open-source datasets continue to proliferate in the medical field, researchers are able to utilize datasets with more comprehensive and diverse lesion characteristics, providing a valuable resource for OOD detection research. The current study involves skin lesion information from multiple datasets, which helps to ensure diversity in the training dataset and significantly improves the stability and robustness of the model. In the OOD detection model, the diversity of training data samples has a particularly significant impact on the detection results. The diversity of the training data allows the model to better recognize data with unknown distributions, demonstrating the potential value and reliability of the model in real-world applications.

Despite the progress made in this study in the field of OOD detection of skin lesions, however, there are still some limitations that need to be further explored and addressed. First, the limited size of the dataset used may not adequately represent the diversity of skin lesions in actual clinical settings. Therefore, when confronted with certain rare or emerging skin lesion types, the generalization ability and detection accuracy of the model may be insufficient. Moreover, in the OOD detection of skin lesions, the similarity between images presents a complex and challenging issue. In particular, the high similarity between certain non-skin lesion images and skin lesion images often leads to misclassifications in OOD detection. For example, varicose veins may cause the appearance of red or purple net-like patches on the skin surface, accompanied by localized swelling, which is prone to be misjudged as hemangiomas or purpuric skin lesions. Similarly, in some cases, lymphadenitis may lead to the formation of pustule-like or erythematous areas on the skin surface, especially when skin changes caused by enlarged lymph nodes closely resemble those of skin lesions, leading to incorrect identification as ulcers or nodules. Given the visual similarity between these non-skin lesions and actual lesions, effectively distinguishing these similar types of lesions has become a significant challenge in the OOD detection task for skin lesions. Finally, factors such as noise and image quality differences that may be encountered in practical applications are not yet fully considered in this study. Therefore, future research should focus on expanding the size of the dataset, improving the generalization ability of the model, exploring more different models, and testing them under conditions closer to clinical application scenarios to further validate and improve the performance of the model.

VI. Conclusion and Future Research

The current study is testing the accuracy of depth models for OOD detection of skin lesions, specifically for the detection of unknown distribution skin lesion images from known distribution skin lesion images. The results of the study show that the models tested exhibit almost similar performance. However, the best-performing model was EBOD, which significantly outperformed the other models in both the AUROC and FPR95 metrics. Future research should focus on the creation of larger databases and the expansion of the variety of skin lesions used to train the models, which could help the models learn a wider range of features and allow them to better understand the differences between known distribution data and unknown data, thus improving their performance in OOD detection. This means that the model not only performs well on training data but also maintains high accuracy on unknown distribution data. As more and more research is devoted to OOD detection modeling, researchers believe that more advanced algorithms will emerge to improve the accuracy and stability of the models and make them perform better in the face of different types of OOD data. This study provides a basis for further research on OOD detection of skin lesions using depth modeling and demonstrates its great potential for medical applications.

References

[1] A. Brunssen, A. Waldmann, N. Eisemann, and A. Katalinic, "Impact of skin cancer screening and secondary prevention campaigns on skin cancer incidence and mortality: a systematic review," Journal of the American Academy of Dermatology, vol. 76, no. 1, pp. 129-139, 2017.

[2] C. Di Raimondo, F. Lozzi, P. P. Di Domenico, E. Campione, and L. Bianchi, "The diagnosis and management of cutaneous metastases from melanoma," International Journal of Molecular Sciences, vol. 24, no. 19, p. 14535, 2023.

[3] A. Courtney, D. J. Lopez, A. J. Lowe, Z. Holmes, and J. C. Su, "Burden of disease and unmet needs in the diagnosis and management of atopic dermatitis in diverse skin types in Australia," Journal of Clinical Medicine, vol. 12, no. 11, p. 3812, 2023.

[4] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," In Proc. 6th Int. Conf. Learning Representations (ICLR), 2018.

[5] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," International Conference on Learning Representations, 2022.

[6] M. Yi et al., "Improved OOD generalization via adversarial training and pretraining," International Conference on Machine Learning, PMLR, pp. 11987-11997, Jul. 2021.

[7] J. Liu et al., "Towards out-of-distribution generalization: A survey," arXiv preprint arXiv:2108.13624, vol. 1, no. 1, p. 1, 2021.

[8] C. Zhang et al., "Enhancing lung cancer diagnosis with data fusion and mobile edge computing using DenseNet and CNN," Journal of Cloud Computing, vol. 13, no. 1, p. 91, 2024.

[9] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115-118, 2017.

[10] L. Hoang, S. H. Lee, E. J. Lee, and K. R. Kwon, "Multiclass skin lesion classification using a novel lightweight deep learning framework for smart healthcare," Applied Sciences, vol. 12, no. 5, p. 2677, 2022.

[11] D. Moturi, R. K. Surapaneni, and V. S. G. Avanigadda, "Developing an efficient method for melanoma detection using CNN techniques," Journal of the Egyptian National Cancer Institute, vol. 36, no. 1, p. 6, 2024.

[12] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," Artificial Intelligence Review, vol. 53, pp. 5455-5516, 2020.

[13] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 12, pp. 6999-7019, 2021.

[14] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3286-3295.

[15] R. Moradi, R. Berangi, and B. Minaei, "A survey of regularization strategies for deep models," Artificial Intelligence Review, vol. 53, no. 6, pp. 3947-3986, 2020.

[16] S. A. Munoz, J. Park, C. M. Stewart, A. M. Martin, and J. D. Hedengren, "Deep transfer learning for approximate model predictive control," Processes, vol. 11, no. 1, p. 197, 2023.

[17] H. Choi, H. Jeong, and J. Y. Choi, "Balanced energy regularization loss for out-of-distribution detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15691-15700.

[18] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," Special Lecture on IE, vol. 2, no. 1, pp. 1-18, 2015.

[19] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," International Conference on Information Processing in Medical Imaging, Cham: Springer International Publishing, pp. 146-157, May 2017.

[20] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," Advances in Neural Information Processing Systems, vol. 31, pp. 7167-7177, 2018.

[21] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," Neural Computation, vol. 13, no. 7, pp. 1443-1471, 2001.

[22] C. Xu, F. Yu, Z. Xu, N. Inkawhich, and X. Chen, "Out-of-Distribution Detection via Deep Multi-Comprehension Ensemble," arXiv preprint arXiv:2403.16260, vol. 1, no. 1, p. 1, 2024.

[23] Y. Dai, H. Lang, K. Zeng, F. Huang, and Y. Li, "Exploring Large Language Models for Multi-Modal Out-of-Distribution Detection," Findings of the Association for Computational Linguistics*: EMNLP 2023*, pp. 5292-5305, Dec. 2023.

[24] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific Data, vol. 5, no. 1, pp. 1-9, 2018.

[25] E. M. Senan, M. E. Jadhav, and A. Kadam, "Classification of PH2 images for early detection of skin diseases," In 2021 6th international conference for convergence in technology (I2CT), pp. 1-7, Apr. 2021.

[26] R. Daneshjou et al., "Disparities in dermatology AI performance on a diverse, curated clinical image set," Science Advances, vol. 8, no. 31, p. eabq6147, 2022.

[27] A. Aboulmira, H. Hrimech, and M. Lachgar, "Comparative Study of Multiple CNN Models for Classification of 23 Skin Diseases," International Journal of Online & Biomedical Engineering, vol. 18, no. 11, 2022.

[28] A. A. D. Alsaeed, "On the development of a skin cancer computer aided diagnosis system using support vector machine," Biosci. Biotechnol. Res. Commun., vol. 12, pp. 297-308, 2019.

[29] A. G. Pacheco et al., "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," Data in Brief, vol. 32, p. 106221, 2020.

[30] J. Mitros and B. Mac Namee, "On the importance of regularisation and auxiliary information in OOD detection," International Conference on Neural Information Processing, Cham: Springer International Publishing, pp. 361-368, Dec. 2021.

[31] P. Sermanet, S. Chintala, and Y. LeCun, "Convolutional neural networks applied to house numbers digit classification," Proceedings of the 21st international conference on pattern recognition (ICPR2012), pp. 3288-3291, Nov. 2012.

[32] V. Thakkar, S. Tewary, and C. Chakraborty, "Batch normalization in convolutional neural networks—A comparative study with CIFAR-10 data," 2018 fifth international conference on emerging applications of information technology (EAIT), pp. 1-5, Jan. 2018.

[33] L. Peng, G. Wang, and C. Zou, "Measuring, testing, and identifying heterogeneity of large parallel datasets," Statistica Sinica, vol. 33, no. 4, pp. 2787-2808, 2023.

[34] T. N. Minh, M. Sinn, H. T. Lam, and M. Wistuba, "Automated image data preprocessing with deep reinforcement learning," arXiv preprint arXiv:1806.05886, vol. 1, no. 1, p. 1, 2018.

[35] S. Thulasidasan, G. Chennupati, J. Bilmes, T. Bhattacharya, and S. Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 13911-13922, Dec. 2019.

[36] Y. Zhu et al., "Boosting out-of-distribution detection with typical features," Advances in Neural Information Processing Systems, vol. 35, pp. 20758-20769, 2022.

[37] T. Wei, B. L. Wang, and M. L. Zhang, "EAT: Towards Long-Tailed Out-of-Distribution Detection," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 14, pp. 15787-15795, Mar. 2024.

[38] S. M. Xie et al., "In-n-out: Pre-training and self-training using auxiliary information for out-of-distribution robustness," arXiv preprint arXiv:2012.04550, vol. 1, no. 1, p. 1, 2020.

[39] A. Ullah, H. Elahi, Z. Sun, A. Khatoon, and I. Ahmad, "Comparative analysis of AlexNet, ResNet18 and SqueezeNet with diverse modification and arduous implementation," Arabian Journal for Science and Engineering, vol. 47, no. 2, pp. 2397-2417, 2022.

[40] S. Regmi et al., "ReweightOOD: Loss Reweighting for Distance-based OOD Detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 131-141, 2024.

[41] S. Wu et al., "L1-norm batch normalization for efficient training of deep neural networks," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 7, pp. 2043-2051, 2018.

[42] T. DeVries and G. W. Taylor, "Learning confidence for out-of-distribution detection in neural networks," arXiv preprint arXiv:1802.04865, vol. 1, no. 1, p. 1, 2018.

[43] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," arXiv preprint arXiv:1812.04606, vol. 1, no. 1, p. 1, 2018.

[44] A. A. Papadopoulos, M. R. Rajati, N. Shaikh, and J. Wang, "Outlier exposure with confidence control for out-of-distribution detection," Neurocomputing, vol. 441, pp. 138-150, 2021.

[45] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," Advances in Neural Information Processing Systems, vol. 33, pp. 21464-21475, 2020.

[46] H. Wang, C. Zhao, X. Zhao, and F. Chen, "Layer adaptive deep neural networks for out-of-distribution detection," Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 526-538, May 2022.

[47] A. T. Nguyen, F. Lu, G. L. Munoz, E. Raff, C. Nicholas, and J. Holt, "Out of distribution data detection using dropout bayesian neural networks," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 7, pp. 7877-7885, June 2022.

[48] Y. Zhu and S. Newsam, "Densenet for dense flow," 2017 IEEE International Conference on Image Processing (ICIP), pp. 790-794, Sep. 2017.

[49] A. Nogales, A. J. Garcia-Tejedor, D. Monge, J. S. Vara, and C. Anton, "A survey of deep learning models in medical therapeutic areas," Artificial Intelligence in Medicine, vol. 112, p. 102020, 2021.

[50] C. Bhatt, I. Kumar, V. Vijayakumar, K. U. Singh, and A. Kumar, "The state of the art of deep learning models in medical science and their challenges," Multimedia Systems, vol. 27, no. 4, pp. 599-613, 2021.

[51] S. S. Kshatri and D. Singh, "Convolutional neural network in medical image analysis: a review," Archives of Computational Methods in Engineering, vol. 30, no. 4, pp. 2793-2810, 2023.

Using Fuzzy Matter-Element Extension Method to Cultural Tourism Resources Data Mining and Evaluation

Fei Liu
School of Humanities and Tourism, Zhejiang Institute of Economics and Trade, Hangzhou, 310018 Zhejiang, China

***Abstract*—This study explores the mining and evaluation of cultural and tourism resources based on fuzzy matter-element extension in the context of cultural and tourism integration. Through fieldwork and analysis of cultural and tourism resources, it is found that the fuzzy matter-element extension theory can be effectively applied to the mining and evaluation of cultural and tourism resources in the context of cultural and tourism integration. The application of integration of cultural and tourism resources has a significant driving effect in tourism development, which can effectively enhance the tourist experience and improve the visibility and attractiveness. Meanwhile, through field research and data analysis, this study also puts forward relevant improvement suggestions for the characteristics and actual situation of the research object, aiming at further optimising the development mode, realising the organic integration of culture and tourism resources, and promoting the prosperity and development of the local cultural industry. Overall, this study has certain theoretical and practical significance for promoting the integrated development of culture and tourism and the sustainable development of tourism.**

***Keywords*—Cultural and tourism integration; fuzzy object meta-theory; development; organic integration**

I. INTRODUCTION

Culture and tourism integration is the current hot topic of culture and tourism development. With the development of economic globalisation and cultural diversity, the integration of culture and tourism has become an important means of promoting local economic development and cultural inheritance [1-2]. The integration of culture and tourism is not only a combination of culture and tourism, but also a brandy new mode of development and way of thinking, which organically combines cultural and tourism resources to create unique cultural and tourism products, and to enhance the soft power of local economy and culture of [3].

In the context of the integration of culture and tourism, the excavation and evaluation of cultural and tourism resources have become the focus of attention, and the research on the excavation and evaluation of cultural and tourism resources is becoming a hot topic in both the academic and practical fields [4]. Firstly, with the integration and development of the culture and tourism industry, people have begun to pay attention to how to effectively excavate and utilise cultural resources, which includes examining the cultural characteristics of different regions, historical relics, traditional handicrafts, etc., as well as how to combine these resources with tourism products to enhance the tourism experience and attractiveness [5]. Secondly, much attention has also been paid to the evaluation research of cultural and tourism resources, which includes not only measuring and assessing the actual value of cultural and tourism resources [6], but also understanding tourists' perceptions and experiences of cultural and tourism resources, so as to better improve and enhance the utilisation efficiency and attractiveness of the resources [7]. In addition, the sustainable development of cultural and tourism resources is also a popular direction of current research [8], and many researchers have begun to pay attention to how to achieve sustainable development of resources while exploring and utilising cultural and tourism resources, so as to protect the cultural heritage and ecological environment, and to ensure the long-term and stable utilisation of the resources [9], [10].

It can be seen that in the current context of the integration of culture and tourism, the research on the excavation and evaluation of cultural and tourism resources has important practical and theoretical significance [11]. However, at present, the excavation of cultural and tourism resources mainly relies on the experience of experts, and lacks a systematic and scientific evaluation model [12], a practicable evaluation model can promote the integrated development of the culture and tourism industry, and through the excavation and evaluation of cultural and tourism resources, it can create tourism products with characteristics and attraction, and promote the coordinated development of the culture and tourism industry. As an emerging evaluation method, the fuzzy material element theory can solve this problem well. The fuzzy material element theory is a new type of theory combining fuzzy mathematical theory and material element theory, which can effectively deal with the fuzzy and uncertainty information, and provides new ideas and methods for the evaluation of cultural and tourism resources [13]. Therefore, this paper takes the current development of cultural tourism in a certain place as an example, through the fuzzy matter-element extension theory can quantitatively evaluate the cultural and tourism resources, and creatively solves the uncertainty and fuzziness problems in the evaluation of cultural and tourism resources. Through the fuzzy object element theory, a set of scientific evaluation system of cultural and tourism resources can be established, which can provide powerful support for the integrated development of culture and tourism.

II. System Building in the Context of Cultural and Tourism Integration

A. Concept of Cultural and Tourism Integration

The integration of culture and tourism, as a new trend in the current tourism development, is a product of the in-depth integration of culture and tourism fields [14]. It is not only an organic combination of culture and tourism resources, but also a new development model and concept. The integration of culture and tourism is embodied in practice as the intermingling of culture and tourism patterns, which can not only promote the inheritance and promotion of local culture, but also inject new vitality and elements into the development of tourism [15]. Therefore, the integration of culture and tourism is not only the integration of a single field, but also a new mode of co-operation between industries to complement and promote each other.

B. Indicator System Construction

The system should have a wide coverage and strong applicability, and needs to be able to comprehensively reflect the characteristics and quality of cultural and tourism resources. The system of valuation indicators constructed should comprehensively consider all aspects of cultural and tourism resources, including cultural connotation, historical value, tourism attraction, sustainability and other aspects [16]. Finally, this paper considers that the evaluation indexes should be operable to ensure that relevant data can be collected and processed effectively. Based on this, this paper finally constructs the indicator system as shown in Table I.

It is worth mentioning that in the process of constructing the indicator system, this paper learnt that many scholars have better requirements for the evaluation of cultural resources [19]. For the indicator of cultural relics protection, it is interpreted in the definition as assessing the protection of cultural relics in a specific geographical area, including the number, completeness and degree of protection of cultural heritage resources such as historical monuments, cultural relics and ancient buildings, etc. However, scholars believe that the historical, artistic and scientific values of these cultural relics in the assessment are important factors that should be taken into account comprehensively in the assessment; for the indicator of cultural industry, although scholars believe it is For the indicator of cultural industries, although scholars believe that it is to assess the development of cultural industries in a specific region, including the scale, vitality and innovation capacity of cultural industries such as cultural creative industries, cultural and artistic performances, cultural exhibitions, etc., some other scholars still believe that this indicator should take into account the contribution of the relevant industries to the local economic and social development; for the indicator of cultural traditions, many scholars believe that whether the inheritance of these traditions and culture has been effectively protected and disseminated should also be an important factor in the assessment [20]. For the indicator of cultural traditions, many scholars believe that the effective protection and dissemination of the traditional culture should also be an important part of the assessment; for the indicator of cultural activities, the promotion of these cultural activities to the local cultural industry and tourism should not be neglected in the assessment process; and for the last indicator of cultural resources, the assessment of this indicator needs to include the impact of cultural education on the cultural literacy and cultural identity of the local residents.

TABLE I. Evaluation Index System based on the Concept of Cultural Tourism Integration

Normative layer	***Indicator layer***	***Interpretation of indicators***
Cultures resource (such as manpower or tourism)[17]	Heritage conservation	To assess the status of heritage conservation in a given geographical area, including the number, integrity and degree of protection of cultural heritage resources such as historical monuments, artefacts and ancient buildings.
	Cultural industry	The development of cultural industries in the region, including the scale, vitality and innovation capacity of cultural industries such as cultural and creative industries, cultural and artistic performances, and cultural exhibitions.
	Cultural tradition	To assess the transmission and development of cultural traditions in a given geographical area, including the degree of transmission and contemporary value of traditional cultural resources such as traditional festivals, folk culture and intangible cultural heritage.
	Cultural activity	To assess the richness and impact of cultural activities in a given geographical area, including the number, scale and impact of cultural activities such as cultural exhibitions, performances and cultural exchange activities.
	Cultural education	To assess the level and coverage of cultural education in a given geographical area, including the quantity and quality of cultural and educational resources such as cultural and educational institutions, cultural and educational programmes, and public cultural services.
Journey resource (such as manpower or tourism)[18]	Natural landscape	To assess the richness and attractiveness of the natural landscape of a given territory, including the quantity and quality of natural scenery resources such as mountain scenery, lakes and rivers, forests and grasslands, as well as their attractiveness to tourists.
	Cultural landscape	To assess the richness and attractiveness of the cultural landscape of a given territory, including the quantity and quality of cultural heritage resources such as historical monuments, traditional architecture, religious sites, etc., and their attractiveness to tourists.
	Tourist facility	To assess the degree of sophistication of tourism facilities in a given geographical area, including the quantity and quality of tourism infrastructure such as hotels, restaurants, transport, guide services, etc., as well as the degree of accessibility to tourists.
	Tourism activity	To assess the diversity and attractiveness of tourism activities in a given geographical area, including the richness and characteristics of tourism activities such as mountain hiking, water sports, cultural experiences, festivals and events, as well as their attractiveness to tourists.
	Tourism Services	To assess the quality and level of tourism services in a given geographical area, including the quality of services provided by tour guides, tourism information and counselling, tourism safety and security, as well as the level of satisfaction with tourists.

C. Coefficient of Variation Method

The coefficient of variation method is a statistically based method for the objective calculation of weights and is suitable for assessing situations where there are different degrees of variability between indicators. The coefficient of variation is the ratio of the standard deviation of an indicator to its mean value, which reflects the relative degree of dispersion of the indicator data. The larger the coefficient of variation, the greater the degree of volatility of the indicator, and vice versa [21],[22]. When calculating the weights, the coefficient of variation can be used to measure the importance of each indicator in the overall evaluation. In the evaluation index system of cultural and tourism resources based on fuzzy object elements in the context of cultural and tourism integration, first of all, taking into account that the indicators cover a wide range so that the objective weight assignment method can reduce the variability brought by the collection of data, and the data often involves different degrees of variability between the indicators, so the coefficient of variation method can be a very good solution to this problem. The calculation steps are shown below.

Data standardization:

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^{\ 2}}} \tag{1}$$

Where: r_{ij} - Normalised data matrix elements;

x_{ij} -Data matrix elements;

m -Indicator Data Matrix Calculation Sample.

Calculation of the coefficient of variation

The new data matrix $R = \left(r_{ij}\right)_{m\times n}$ can be formed after the first step of processing.

Calculate the mean value of the indicator:

$$A_j = \frac{1}{n}\sum_{i=1}^{m} r_{ij} \tag{2}$$

Where: A_j - Mean value of the indicator;

n -Number of data samples.

Calculate the standard deviation of the indicator:

$$S_j = \sqrt{\frac{1}{n}\sum_{i=1}^{m}\left(r_{ij} - A\right)^2} \tag{3}$$

Where: S_j Indicator standard deviation.

Calculate the coefficient of variation:

$$V_j = \frac{S_j}{A_j} \tag{4}$$

Where: V_j - Standard deviation of the indicator.

(3) Calculation of weights

$$\omega_{ij} = \frac{V_j}{\sum_{j=1}^{n} V_j} \tag{5}$$

D. Fuzzy Object Element Modelling

Under the background of cultural and tourism integration, cultural and tourism resource evaluation is very important because it can help relevant departments and enterprises better understand and utilise the resources and promote the development of cultural and tourism industries. And the fuzzy object element model is a mathematical model that can deal with uncertainty and ambiguity information, which can effectively deal with all kinds of fuzzy and uncertain information and data involved in the evaluation process of cultural and tourism resources, specifically, the fuzzy object element model transforms the uncertainty information into the affiliation function, so that it can better describe and deal with all kinds of uncertainty [23]. The calculation process is as follows.

1) Constructing fuzzy object elements: Let the given thing be M, v is its feature, C has the measure value, using the ordered triad R= (M, C, v), R is called the crop element, when the measure value v has the fuzzy nature, R is the fuzzy object element, and μ(x) is the degree of affiliation of the thing M to the corresponding measure value x of its feature C is recorded as:

$$R = \begin{bmatrix} & M \\ C & \mu_{(x)} \end{bmatrix} \tag{6}$$

2) Fuzzy object element construction: Constructing fuzzy object elements based on the normalisation in the weighting results of this paper.

$$R_{mn} = \begin{bmatrix} & M_1' & M_2' & \cdots & M_n' \\ C_1 & x_{11}' & x_{12}' & \cdots & x_{1n}' \\ C_2 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots' & \vdots' & \vdots & \vdots' \\ C_m & x_{m1} & x_{m2} & \vdots & x_{mn} \end{bmatrix} \tag{7}$$

Where: x_{ij}' - Indicator normalised value.

3) Constructing standard objects

$$R_0 = \begin{bmatrix} & M_0 \\ C_1 & x_{10} \\ C_2 & x_{20} \\ \vdots & \\ C_{\mathrm{m}} & x_{m0} \end{bmatrix} \tag{8}$$

Where: x_{i0}' - optimal value of the indicator, normalised to 1.

4) Constructing fuzzy object elements for the difference between the standard object and the evaluation object

Calculate the standard object and evaluation object difference$\Delta_{ij}=\left[x_{i0}-x_{ij}'\right]^2$, based on which the object element matrix is constructed:

$$R_{mn}=\begin{bmatrix} & M_1, & M_2, & \cdots & M_n, \\ C_1 & \Delta x_{11}, & \Delta x_{12}, & \cdots & \Delta x_{1n}, \\ C_2 & \Delta x_{21} & \Delta x_{22} & \cdots & \Delta x_{2n} \\ \vdots & \vdots, & \vdots, & \vdots & \vdots, \\ C_m & \Delta x_{m1} & \Delta x_{m2} & \vdots & \Delta x_{mn} \end{bmatrix} \quad (9)$$

Where: Δ_{ij} - calculated difference

5) Euclid closeness: The larger the value of Euclidean closeness the closer it is to the optimum. The formula for calculating the Euclidean closeness is as follows:

$$\rho H=1-\sqrt{\sum_{i=1}^{m}\omega_i\Delta_{ij}} \quad (10)$$

Where: ρH - Euclid closeness

III. Cultural Tourism Resources

A. Analysis of the Study Area

The study area belongs to the key cultural and tourism project of Jiangsu Province, the national 3A level tourist attraction - Shazhou, Zhangjiagang City. Through the field investigation of the exhibition halls, cultural relics, traditional handicrafts, etc. in the study area, the cultural elements and historical background displayed in the study area were understood, and after in-depth exchanges with the managers and staff, the development history, cultural activities, and future development planning in the study area were understood, based on which the state of the indicators needed to be examined in the study area was quantified by the data. Through the investigation and analysis, it was found that the study area possesses rich cultural resources, including traditional handicrafts, historical relics, folk culture and other elements, among which, many traditional handicrafts skills are preserved, such as ceramics, weaving, etc. which represent the unique local cultural traditions. Many precious historical relics are also displayed, which reflect the important position and cultural heritage of a certain place in history. In addition, local folk culture and traditional festivals, etc. are shown to tourists through various cultural activities and exhibitions, enhancing their understanding and awareness of local culture.

B. Coefficient of Variation Method for Calculating Weights

The relevant values and weights of the coefficient of variation method calculated according to equations (1-5) are shown in Table II.

C. Analysis of Weighting Results

Based on the results of the weight calculation, the weight values are sorted and then plotted in a bar chart as shown below.

According to the results of the chart, cultural resources, cultural relics protection, cultural industry, cultural activities accounted for 13.82%, 16.80%, 11.79%, according to the survey and field consulting analysis of cultural relics protection is an important part of the cultural resources, which carries the history, traditions and cultural memories of a region or a country, and cultural relics protection not only protects the historical heritage, but also passes on the cultural heritage. Tradition, it is of great significance to enhance the cultural soft power of a region or a country. Meanwhile, with the development of culture and tourism integration, cultural industry has become an important force to support the development of tourism, which generally includes cultural creative industry, cultural tourism industry, etc., and they play an important role in the economic development and job creation of a region or a country. Corresponding cultural activities are an important way to enrich cultural life and promote cultural exchanges, and rich and diversified cultural activities can attract tourists, enhance the cultural brand image of the region or country, and then promote the development of cultural tourism. Therefore, the three indicators of heritage protection, cultural industry and cultural activities are in the study, and their weights are relatively high.

TABLE II. Coefficient of Variation Method

Indicator layer	Average value	Variance (statistics)	Coefficient of variation	Weights
Heritage conservation	0.2580	0.0093	0.0361	0.1382
Cultural industry	0.2580	0.0113	0.0439	0.1680
Cultural tradition	0.2582	0.0033	0.0127	0.0486
Cultural activity	0.2581	0.0080	0.0308	0.1179
Cultural education	0.2582	0.0028	0.0110	0.0422
Natural landscape	0.2582	0.0040	0.0154	0.0589
Cultural landscape	0.2582	0.0026	0.0101	0.0386
Tourist facility	0.2580	0.0102	0.0395	0.1509
Tourism activity	0.2579	0.0129	0.0500	0.1914
Tourism services	0.2582	0.0031	0.0118	0.0453

The same analysis shows that for tourism resources, the weight of tourism facilities and tourism activities is relatively large, accounting for 19.14% and 15.09% respectively, while the landscape indicator instead accounts for a moderate proportion, which indicates that the landscape, whether it is a natural landscape or a cultural landscape, is a necessity for any tourism region, which explains why the importance of the proportion of the performance of the general. Under the background of cultural and tourism integration, the reasons for the relatively high weighting of the two criteria of tourism facilities and tourism activities in tourism resources are analysed as follows: in tourism activities, tourism facilities are the infrastructure to support tourists' experience of tourism services, and high-quality tourism facilities can provide comfortable and convenient tourism environments to satisfy tourists' needs for tourism experience. For example, wine, scenic facilities, transport facilities, etc., their quality and service level directly affects the tour's satisfaction and tourism experience. Tourism activities are an important factor in attracting tourists, and colourful tourism activities can enhance the attractiveness and competitiveness of tourist destinations. For example, cultural festivals, experiential activities, theme performances, etc., which can enrich tourists' travel experience, increase their stay time and consumption, and play an important role in promoting the tourism economy of a region or country. In summary, tourism facilities and tourism activities are important indicators for evaluating tourism resources in the context of cultural and tourism integration, and they are directly related to tourists' tourism experience and the attractiveness of the destination, so their weights are relatively high. It is of great significance for excavating and evaluating cultural and tourism resources.

D. Fuzzy Object Element Modelling Analysis

The application of fuzzy object-element model in the study area is carried out according to Equation (6-10), and the data collected from managers, tourists, local residents, and government personnel are now listed in Table III, and the results of single-indicator calculations and overall calculations are listed in Table IV.

TABLE III. DATA FOR THE STUDY AREA

Indicator layer	***Data 1***	***Data 2***	***Data 3***	***.....***	***Data 14***	***Data 15***
Heritage conservation	85	85	84	.....	85	85
Cultural industry	86	85	83	.....	90	91
Cultural tradition	89	90	89	.....	91	90
Cultural activity	60	59	59	.....	56	56
Cultural education	99	98	97	.....	97	96
Natural landscape	98	97	96	.....	96	95
Cultural landscape	99	98	100	.....	97	100
Tourist facility	82	81	79	.....	85	87
Tourism activity	78	77	75	.....	76	85
Tourism Services	98	97	99	.....	96	99

TABLE IV. EVALUATION RESULTS

Basic indicators	***Euclid approximation***
Heritage conservation	0.9178
Cultural industry	0.8950
Cultural tradition	0.8899
Cultural activity	0.8544
Cultural education	0.8915
Natural landscape	0.8831
Cultural landscape	0.9034
Tourist facility	0.8789
Tourism activity	0.9090
Tourism Services	0.8923
Normative level indicators	***Euclid approximation***
Cultural resource	0.8897
Tourism resource	0.8933
Status of development of cultural and tourism resources	0.8915

E. Analysis of the Evaluation Results of the Fuzzy Material Element Model

Among the indicators belonging to the whole cultural resources, the Euclid approximation of cultural relics protection is 0.9178, the Euclid approximation of cultural industry is 0.8950, and the Euclid approximation of cultural activities is only 0.8544, which can be seen that the overall Euclid approximation of the cultural resources is 0.8897. The indicator with the lowest evaluative value is the cultural activities, which is considered to be influenced by the individual's subjective feeling, and this subjectivity and personalisation may cause the evaluative value of the same cultural activity to be relatively low. Considering that the evaluation of cultural activities is often influenced by the subjective feelings of individuals, the evaluation value of the same cultural activity by different people may differ greatly, and this subjectivity and individualisation leads to the relatively low evaluation value of cultural activities. However, analysing the data of the single indicator, it can be seen that although there are differences in the evaluation of this indicator by the collection object, the results show that it is generally not high, thus indicating that the cultural activities in the study area need to be improved in a targeted manner, and that the study area needs to enhance the sustainable development and innovation of cultural activities, and to improve the activities. The study area needs to strengthen the sustainable development and innovation of cultural activities to improve the attractiveness and influence of the activities, so as to enhance the appraisal value of cultural activities; at the same time, government departments should encourage the organisation of diversified cultural activities, including traditional culture, contemporary culture, folk culture, etc. so as to promote the inclusiveness of cultural activities, thus enhancing the appraisal value.

Among the indicators belonging to tourism resources, the evaluation value is relatively close, and the Euclid closeness of natural landscape, cultural landscape, tourism facilities, tourism activities, and tourism services are 0.8831, 0.9034, 0.9090, 0.8789, and 0.8923, respectively, of which the closeness of the tourism facilities in the study area is the lowest at 0.8789, which means that compared to the rest of the indicators, there is more room for improvement. The study area has a long cultural history, so attracting tourists by exploring history and culture, folk customs, traditional crafts and rich cultural elements can increase the evaluation value. At the same time, the tell-tale development of society has prompted tourists to be more and more positive about the demand for facilities and the sense of experience, so the study area can attract tourists by creating richer and more varied tourism experience projects, such as cultural performances, interactive experiences, and themed activities. In addition, consider improving the service level of tourism facilities, including guided tours, scenic area management, convenient facilities, etc., to improve visitor satisfaction, and combining contemporary technological means, such as virtual reality, augmented reality and other technologies, to bring novel experiences to tourists, to attract more tourists, and to promote the development of cultural and tourism integration.

After analysing the indicators belonging to cultural resources and tourism resources respectively, we are now analysing the indicators with relatively high degree of Euclid closeness, which are cultural relics protection, cultural landscape and tourism activities. Tourism activities in the general environment of regionalisation is not obvious, depending on the delicacy of tourism activities around the world, and cultural landscape, heritage protection of the two indicators are closely related, but also the highlight of the research in this paper, in this paper in the subsection of the indicator system of cultural heritage protection of this indicator needs to assess the situation of regional heritage protection, including the number of historical monuments, cultural relics, ancient buildings and other cultural heritage resources, the completeness and the degree of protection, and at the same time, the historical value, artistic value and scientific value of these The historical value, artistic value and scientific value of cultural relics are all important factors for assessment. The local historical monuments, ancient buildings and cultural relics also form the cultural landscape of tourism, which shows that there is a certain difference between heritage conservation and cultural landscape but the connection is close. First of all, heritage protection and cultural landscape are both for the protection and inheritance of historical heritage, promote traditional culture, enhance national cohesion and cultural self-confidence. Secondly, the success of heritage protection determines whether it can become an important part of the cultural landscape, and the formation of the cultural landscape cannot be separated from the protection and use of cultural relics, therefore, heritage protection and cultural landscape are often intertwined and mutually supportive in practice, and jointly promote the sustainable development of the integration of culture and tourism.

IV. Conclusion and Outlook

This paper based on the fuzzy object element theory on the study area for field research and case analysis, we found that the applicability of fuzzy object element method is good, and for the evaluation of the results of the realisation of the clear, through the evaluation results of the model can promote the integration of the development of culture and tourism industry to enhance the effect of the local cultural heritage and promotion, and to achieve the sustainable development of the tourism industry. The results of this paper are now combined with the current stage of China's tourism development for specific elaboration.

A. Development Orientation Analysis

The positioning of efficient development is to organically integrate cultural and tourism resources to create a comprehensive scenic spot integrating cultural heritage, tourism experience, leisure and entertainment. The different attractions then become as an important carrier for the integration of culture and tourism, with the important role of inheriting and displaying local culture, promoting tourism development, and enhancing the image of the city.

Secondly, in the context of the integration of culture and tourism, local tourist attractions and even Netflix attractions need to be committed to tapping and showcasing local traditional cultural resources, promoting local cultural characteristics, attracting tourists to come and experience them, and promoting the development of the local tourism industry. At

the same time, they also need to focus on innovation and provide tourists with colourful cultural experiences through various cultural activities, exhibitions and performances, so that they can feel the charm of local culture while playing.

Cultural landscapes are pivotal in the study of this paper, so cultural attractions can be used as a tourist destination with high cultural value, and constantly explore its development potential. Its status is not only a tourist attraction, but also a comprehensive place with education and cultural inheritance functions. By tapping into its cultural and tourism resources, it can better play its role in attracting more tourists to visit and experience, thus promoting the prosperity of the local tourism industry.

Overall, in future research, field studies and questionnaires can be combined to gain an in-depth understanding of the needs of tourists and the development direction of the cultural park, and to promote the prosperity of the local tourism industry.

B. Analysis of the Development Path of Cultural and Tourism Integration

This paper analyses the relationship between cultural landscape and cultural relics protection in the fuzzy object element result analysis, and it can be found that the development path of cultural and tourism integration needs to be achieved through scientific planning and careful design. Planning is the foundation of culture and tourism integration, and it needs to start from the perspective of overall development, taking into account the differences and complementarities of culture and tourism resources. In the development process, the planning should focus on cultural heritage, tourism experience and industrial development, and through the coordination of all kinds of resources and elements, create cultural and tourism integration products with local characteristics and competitiveness. Design is a key link in the implementation of cultural and tourism integration, and needs to focus on innovation and inclusiveness. Design work should incorporate local cultural elements and take into account the needs and experiences of tourists, so as to make tourist attractions attractive destinations, and at the same time, it should take into account environmental protection and cultural heritage, so as to create an integrated space for cultural and tourism integration.

C. Shared and Sustainable Development

Tourism is dependent on local buildings and residents, so it is necessary to encourage local communities to participate in the organisation of tourism activities and promote the shared development of communities and tourism, such as the development of tourism hospitality in the form of agro-entertainment and B&Bs, so as to increase the sources of income of local residents. Meanwhile, measures to protect the local cultural heritage and natural environment will be amicably bound to the residents of the tourism area, so as to strengthen the protection and inheritance of the cultural heritage, promote eco-friendly tourism activities, and reduce the negative impact on the environment. In addition to focusing on the dissemination and sharing of culture, tourism practitioners are encouraged to provide personalised and customised cultural tourism products and services to meet the needs of different groups of tourists, such as customised cultural and creative handicrafts and customised travel routes with cultural themes.

This paper hopes that through the above suggestions for upgrading and transformation, the attractiveness and competitiveness of cultural and tourism resources can be further enhanced, the integrated development of culture and tourism can be promoted, and the goal of cultural heritage and sustainable tourism development can be realised.

REFERENCES

[1] Canavan, B. Tourism culture: nexus, characteristics, context and sustainability[J]. Tourism management, 2016, 53: 229-243.

[2] Tang, C., Liu, Y., Wan, Z., Liang, W. Evaluation system and influencing paths for the integration of culture and tourism in traditional villages[J]. Journal of Geographical Sciences, 2023,33(12), 2489-2510.

[3] Ma, Y., Chen, Y. The Inspiration of the Fusion of Chinese and Western Cultures for the Development of Macau City[J]. Journal of Sociology and Ethnology, 2023,5(11), 162-166.

[4] Jiang, T. Q., Dong, P. H., Yang, J. D. Progress and Insights in Research on Chinese Cultural Landscape[J]. Tourism Travel Guide 2024, (02), 91-109.

[5] Pal, S., Swarnali, B. Apitourism in agritourism: a fusion of greenery, apiculture & tourism in the valley of Jampui hills of North East India[J]. Plant Archives, 2023,1 (23): 51-55.

[6] Guo, S. Z., Yao, J., Wen, L. Urban tourism competitiveness evaluation system and its application: comparison and analysis of regression and classification methods[J]. Procedia Computer Science, 2019, (162): 429-437.

[7] Kalnitska, M. Assessment of the development state of organisational and cultural resources of international tourism business[J]. Euclid Journal of Management Issues, 2018, 26 (4): 71-81.

[8] Zou, Y., Meng, F., Bi, J., Zhang, Q. Evaluating sustainability of cultural festival tourism: From the perspective of ecological niche[J]. Journal of Hospitality and Tourism Management, 2021, 48:191-199.

[9] Joun, H. J., Hany, K. Productivity evaluation of tourism and culture for sustainable economic development: analysing South Korea 's metropolitan regions[J]. Sustainability, 2020, 12(7): 2912.

[10] Meng, Q., Wang, C., Xu, T., Pi, H., Wei, Y. Evaluation of the sustainable development of traditional ethnic village tourist destinations: a case study of Jiaju Tibetan village in Danba county, China[J]. China Land, 2022,11(7), 1008.

[11] Luo, Q. Y. Research on the development strategy of red cultural tourism resources in Yunnan Province under the background of culture and tourism integration[J]. Tourism Overview, 2022,(22):167-169.

[12] Ma, Z. Z. M., Cai, Y., Pan, J. Y. Application of AHP in the evaluation of tourism resources in ethnic regions: case study of Xichang, Liangshan Yi autonomous prefecture, China [C]. IOP Conference Series: Earth and Environmental Science, 2019, 358(3): 032018. DOI 10.1088/1755-1315/358/3/032018

[13] Xiong, W., Dong, Z. C., Lu, J. Q. Ma, J. Y.,Wu, S.J., Li, C. C. River health assessment of the Gansu section of the Weihe River Basin based on a combined-assignment fuzzy elemental topological model[J]. Advances in Water Resources and Hydropower Science and Technology,2023,43(04):9-14+30.

[14] Tang, X. Z., Xie, N. M. Research on the evaluation of tourism development potential of tea intangible cultural heritage based on grey clustering[J]. Grey Systems: Theory and Application , 2019,9 (3): 295-304.

[15] Saiken, A., Zhaoping, Y., Mazbaev, O., Duissembayev, A., Izdenbaev, B., Nassanbekova, S. Ethnic cultural tourism resources evaluation and development: analysis of Kazakh cultural tourism resources[J]. Journal of Environmental Management and Tourism, 2017, 2 (18): 467.

[16] Zheng, Q. Q., Chen, Q. H., Kong, D. Y. Performance evaluation of the development of eco-cultural tourism in Fujian Province based on the method of fuzzy comprehensive evaluation[J]. Frontiers in Environmental Science, 2022, (10): 1022349.

[17] Bing, H., Zhou, X. Q., Lu, X. X., Tao, R., Zhang, A. P. The Construction and Empirical Analysis of the Evaluation System of Urban Cultural Tourism Competitiveness--Taking the Urban Agglomeration in the

Yangtze River Delta region as an Example[J]. World Regional Studies, 2016, (6), 166.

[18] Wang, B., He, S., Min, Q., Cui, F., Wang, B., Liu, X., Bai, Y. Framework for evaluating the development suitability of tourism resources in agricultural heritage systems: A case study of Qingyuan County in Zhejiang Province[J]. Chinese Journal of Eco-Agriculture, 2020, 28(9), 1382-1396.

[19] Yang, S. H, Kong, X. T. Evaluation of rural tourism resources based on AHP-fuzzy mathematical comprehensive model[J]. Mathematical Problems in Engineering, 2022, (2022). DOI:10.1155/2022/7196163

[20] Su, J., Jun, H. Analysis on the tourism resource evaluation factors based on grey relational analysis-Taking Guizhou minority areas as an example[J]. Journal of Computational Methods in Sciences and Engineering 2019, 19(4): 1093-1099.

[21] Zhuo, F. C., Zhou, C. X., Lu, M. Y., Zhou, J., Tang, J. M. Evaluation of fruit traits of 27 dragon fruit germplasm resources based on correlation analysis and coefficient of variation method[J]. China Tropical Agriculture,2023,(02):21-27.

[22] Sun, Y., Liang, X. J., Xiao, C. L. Assessing the influence of land use on groundwater pollution based on coefficient of variation weight method: a case study of Shuangliao City[J]. Environmental science and pollution research, 2019, 26(34): 34964-34976.

[23] Shan, C., Dong, Z., Lu, D., Xu, C., Wang, H., Ling, Z., Liu, Q. Study on river health assessment based on a fuzzy matter-element extension model[J]. Ecological Indicators, 2021, 127, 107742.

Arabic Sentiment Analysis Using Optuna Hyperparameter Optimization and Metaheuristics Feature Selection to Improve Performance of LightGBM

Mostafa Medhat Nazier[1], Mamdouh M. Gomaa[2], Mohamed M. Abdallah[3], Awny Sayed[4]
Computer Science Department-Faculty of Science, Minia University, Minia 61519, Egypt[1, 2, 3]
Information Technology Department-Faculty of Computing and Information Technology,
King Abdulaziz University, Jeddah 21589, Saudi Arabia[4]

Abstract*—Sentiment Analysis (SA) effectively examines big data, such as customer reviews, market research, social media posts, online discussions, and customer feedback evaluation. Arabic Language is a complex and rich language. The main reason for the need to enhance Arabic resources is the existence of numerous dialects alongside the standard version (MSA). This study investigates the impact of stemming and lemmatization methods on Arabic sentiment analysis (ASA) using Machine Learning techniques, specifically the LightGBM classifier. It also employs metaheuristic feature selection algorithms like particle swarm optimization, dragonfly optimization, grey wolf optimization, harris hawks optimizer, and a genetic optimization algorithm to identify the most relevant features to improve LightGBM's model performance. It also employs the Optuna hyperparameter optimization framework to determine the optimal set of hyperparameter values to enhance LightGBM model performance. It also underscores the importance of preprocessing strategies in ASA and highlights the effectiveness of metaheuristic approaches and Optuna hyperparameter optimization in improving LightGBM model performance in ASA. It also applies different stemming and lemmatization methods, Metaheuristic Feature Selection algorithms, and the Optuna hyperparameter optimization on eleven datasets with different Arabic dialects. The findings indicate that metaheuristics feature selection with the LightGBM classifier, using suitable stemming and lemmatization or combining them, enhances LightGBM's accuracy by between 0 and 8%. Still, Optuna hyperparameter optimization with the LightGBM classifier, using suitable stemming and lemmatization or combining them, depending on data characteristics, improves LightGBM's accuracy by between 2 and 11%. It achieves superior results than metaheuristics feature selection in more than 90% of cases. This study is of significant importance in the field of ASA, providing valuable insights and directions for future research.

Keywords*—Arabic Sentiment Analysis (ASA); big data; Light Gradient Boosting Machine (LightGBM); Optuna hyperparameter optimization; metaheuristics feature selection; machine learning

I. INTRODUCTION

Sentiment analysis (SA), also known as opinion mining, is a technique within natural language processing (NLP) that involves several steps: data collection, preprocessing, feature extraction, and sentiment classification. It has recently become popular as an effective tool for examining big data, such as social media posts, customer reviews, market research, online discussions, and social media monitoring [1]. The popularity of SA has grown significantly among marketers and consumers. It enables them to gain insights into products and analyze market behavior. This method, further enhanced by machine learning (ML), data mining (DM), and deep learning (DL) algorithms, objectively assesses whether a text expresses positive or negative emotions or conveys sentiments about a specific issue, instilling confidence in its impartiality [2]. The internet offers valuable insights into Arabic Sentiment Analysis (ASA). However, analyzing Arabic content poses challenges due to the language's complexities, morphological features, inadequate resources, and the absence of suitable corpora [3]. Although plenty of resources exist to understand English social media content, Arabic resources still need improvement. This gap arises primarily from the variety of Arabic dialects in addition to Modern Standard Arabic (MSA). These dialects are linguistically exciting and widely used among Arabic speakers in everyday conversations and on social media. There is an urgent need to develop specialized AI models for language-specific applications in Arabic [4]. The main challenges of SA in Arabic dialects include morphological analysis [5], the scarcity of datasets [4], the complexity of dialects, and dialects scripted in Latin [5].

ASA presents significant challenges due to the linguistic complexities of the Arabic language, including its dialectical variations, morphological richness, and the lack of extensive labeled datasets. Despite the advancement of machine learning techniques, existing sentiment analysis models often struggle with the diversity of Arabic dialects and the need for effective text preprocessing methods. Additionally, optimizing machine learning models for these challenges requires not only efficient algorithms but also hyperparameter optimization and feature selection to enhance performance. Light Gradient Boosting Machine (LightGBM) has been shown to be a powerful model for classification tasks, but its performance in ASA, particularly with large and diverse Arabic datasets, has not been fully explored. This research aims to bridge this gap by enhancing LightGBM's performance using robust preprocessing techniques (ISRI stemmer and Qalsadi lemmatizer) and

advanced optimization methods (metaheuristic feature selection and Optuna hyperparameter tuning).

Feature Selection (FS), also called variable subset selection, is a crucial preprocessing step in ML. It helps reduce computational costs and improve classification accuracy [6] [7]. FS achieves this by discarding noisy, redundant, or irrelevant features, focusing instead on a smaller subset that is sufficient to describe the concept of interest. This process improves the predictor's performance, simplifies data processing, and reduces computational demands. Since the 1960s, FS research has emphasized developing efficient methods to handle high-dimensional datasets, which often include irrelevant or obsolete features [8]. Meta-heuristic techniques, particularly swarm-based optimization algorithms like Grey Wolf Optimization (GWO), Dragonfly Optimization (DFO), Particle Swarm Optimization (PSO), Harris Hawks Optimization (HHO), and Genetic Optimization (GO), have emerged as practical solutions for FS. These methods strike a balance between computational efficiency and solution quality, making them suitable for real-world applications where identifying the optimal feature subset is essential for accurate and cost-effective classification.

Hyperparameter optimization and tuning are critical steps in machine learning to enhance model performance by selecting the best combination of hyperparameters, which are parameters not learned from the data but set prior to training [9]. Effective tuning ensures improved model accuracy, stability, and generalization. Methods such as random search, grid search, and advanced algorithms like genetic algorithms, Bayesian optimization, and hyperband are widely used to efficiently explore the hyperparameter space. Optuna framework further streamline this process. Properly tuned hyperparameters can significantly impact both computational efficiency and predictive performance [10].

The motivation behind this study stems from the growing need for effective sentiment analysis tools for Arabic text, particularly as Arabic is a widely spoken language with various dialects. Traditional approaches to ASA often fail to account for the subtleties of the language, including regional variations and complex word forms. Leveraging LightGBM's strengths with tailored preprocessing techniques, such as the ISRI stemmer and Qalsadi lemmatizer, holds the potential to significantly improve the model's accuracy. Furthermore, applying metaheuristic feature selection algorithms and Optuna for hyperparameter optimization offers a promising way to enhance model performance by selecting the most informative features and fine-tuning the model's parameters. The use of large and diverse datasets is crucial to better understanding the impact of these techniques on ASA across different Arabic dialects. The combination of these methodologies could provide a substantial advancement in the field of ASA, making it more adaptable and accurate for real-world applications.

The paper's contributions include using LightGBM classification algorithm in ASA and improving its performance using ISRI stemmer and Qalsadi lemmatizer with metaheuristic feature selection algorithms and Optuna hyperparameter optimization. A key aspect of our research is using large Arabic datasets from different Arabic dialects in ASA. This diversity in the datasets could significantly enhance LightGBM's performance in ASA. The proposed approach involves using ISRI stemmer alone and Qalsadi lemmatizer alone for data preprocessing and combining them, followed by implementing metaheuristic feature selection algorithms and Optuna. We also compare Optuna hyperparameter optimization with metaheuristic feature selection algorithms to see the impact of improving LightGBM's performance in ASA. It also shows their effects on enhancing ASA.

The remainder of the paper is organized as follows: The "Related Work" in Section II examines previous research on ASA, LightGBM, Optuna hyperparameter, and metaheuristics feature selection. The "Proposed Methodology" in Section III provides detailed information about the proposed approach. In the "Results and Analysis" in Section IV, we present and compare the experimental findings with those of other methods. The "Discussion" in Section V interprets the results and contextualizes their significance, bridging the gap between the findings and the broader implications of the study. Finally, the "Conclusion" in Section VI summarizes the main points of the research.

II. Related Work

This section provides an overview of the existing research on ASA, focusing on key areas such as preprocessing techniques, ML and DL algorithms, FS, and hyperparameter optimization. Several studies have contributed to advancing ASA methodologies by exploring diverse approaches for improving sentiment classification accuracy. Research in ASA has addressed challenges related to the linguistic complexity of Arabic, including its morphological richness and dialectical diversity. In particular, studies have investigated various preprocessing techniques like word embedding, stemming, and lemmatization to enhance text representation. Additionally, the integration of advanced classification algorithms such as LightGBM, along with metaheuristic feature selection and hyperparameter optimization methods like Optuna, have emerged as crucial elements in improving the performance of ASA models. This review synthesizes the most prominent contributions in these areas, providing a comprehensive understanding of the current state-of-the-art in ASA and highlighting opportunities for further development. In study [3], this paper compared several ASA models and discussed the DL algorithms employed in ASA within the domain of e-marketing. The paper's contribution includes improving ASA using preprocessing techniques like word embedding. In a study referenced as study [11], the semantic orientation method was devised to determine the overall polarity of Arabic subjective texts. The technique involved using a specialized domain ontology and an established sentiment lexicon. This technique was evaluated using an Arabic dataset from the hospitality industry to construct the domain ontology. In study [12], this paper employs the Levenshtein distance algorithm for data preprocessing and implementing various classification models and introduces a novel method for conducting ASA using the mobile application comments dataset. This study thoroughly [13], reviews textual content analysis in the ASA domain, examining 133 ASA papers published from 2002 to 2020. It explores common themes, methodologies, technologies, and algorithms used in these studies. This paper's

key finding indicate various approaches, such as ML, lexicon-based, and hybrid methods, with algorithms like SVM, Naive Bayes (NB), and hybrid methods proving the most effective. The research presented in study [14] introduces an explainable sentiment classification framework tailored for Arabic. A noise layer is incorporated into various DL models, such as BiLSTM and CNN-BiLSTM, to mitigate overfitting. In study [15], the authors introduce LightGBM, a new GBDT algorithm featuring two innovative techniques: Gradient-based one-sided sampling (GOSS) and Exclusive Feature Bundling (EFB). These techniques are designed to handle large data instances and features, offering practical benefits concerning memory efficiency and processing speed. The study in [16] presents an innovative hybrid system for detecting fake news, which integrates a BERT-based model with LightGBM. The performance of this approach is assessed against four classification methods that utilize different word embedding techniques across three actual fake news datasets. In study [17], SA is used to determine sentiments in text. LightGBM is used for efficiency and scalability, but long and short-term memory is preferred to understand the deep context of the text. The LSTM model was trained using the Adam optimizer, and Text Blob was used to train the LightGBM. Short-term memory (92%) scores over a LightGBM (89%) accuracy. The F1 score for Long short-term memory (93%) and LightGBM (92%) is about comparable. In study [18], this paper focuses on calculating emotional scores for product features through comparative sentences and developing a clustering method to analyze the hierarchical relationships among brands. It utilizes an improved computing model that leverages a sentiment dictionary to generate weighted sentiment scores, enhancing the accuracy of an unsupervised algorithm. These scores are then organized into a design structure matrix, facilitating the clustering of brands with similar products. The research presented in study [19] developed four hybrid machine learning techniques for multi-class-based comparative SA using three datasets from diverse domains. The results revealed that the Multilayer Perceptron + Random Forest (MLP + RF) hybrid ML technique, employing a multilayer perceptron as the base estimator, achieved an F1-score of 93.0% and an average accuracy of 93.0%. The research presented in study [20] focused on the Optuna hyperparameter optimization framework in conjunction with the LightGBM algorithm. A 10-fold cross-validated model using Optuna and LightGBM was trained on the FHS dataset. The resulting model achieved an accuracy of 0.930, a sensitivity of 0.897, a specificity of 0.963, an F1 score of 0.929, a precision of 0.963, an area under the receiver operating characteristic curve (AUC-ROC) of 0.978, and a Matthews correlation coefficient (MCC) of 0.861. The study in [21] utilized an ensemble model that combined CNN and LSTM to predict the sentiment polarity of Arabic tweets using the ASTD dataset. The model achieved an F1-score of 64.46% and an accuracy of 65.05% on this dataset. In study [22], researchers explored various DL models, including LSTM and CNN, for ASA. They trained neural language models using two techniques based on word2vec: skip-gram and Continuous Bag of Words (CBOW). The experiments demonstrated that LSTM outperformed CNN in terms of performance. The research presented in study [23] developed a SA model by enhancing the ML approach (using complement Naive Bayes) with features derived from both an Arabic sentiment lexicon and the text itself. In study [24], an attempt was made to address ASA using a DL model. The LABR dataset in this study comprises book reviews.

In conclusion, the literature on ASA highlights the significant progress made in improving sentiment classification through advanced preprocessing techniques, effective feature selection, and optimized machine learning models. LightGBM has been demonstrated as a highly efficient algorithm for ASA, particularly when enhanced by hyperparameter optimization via Optuna and metaheuristic feature selection methods. Studies have shown that preprocessing methods such as word embedding, stemming, and lemmatization can effectively address the challenges posed by the Arabic language's unique structure. Hybrid models and deep learning approaches have demonstrated potential in improving the contextual understanding of sentiment in Arabic texts. However, challenges such as dialectal variation and the limited availability of large, labeled datasets persist. This indicates that future research should focus on addressing these issues while also exploring innovative ways to integrate different algorithms and preprocessing techniques. Ultimately, the continued development and optimization of ASA models will contribute to more accurate and efficient SA tools for Arabic-language applications.

III. Proposed Methodology

This section outlines the proposed methodology for conducting ASA using various preprocessing techniques, stemming, lemmatization, or a combination of them, tokenization, feature extraction, LightGBM as a classifier, metaheuristic feature selection methods, and Optuna hyperparameter optimization strategies to enhance the performance of the LightGBM classifier, as shown in Fig. 1.

A. Preprocessing Data

Preprocessing is a vital phase in ASA, as it prepares the input data for effective SA. This step significantly enhances the quality of SA [25]. In this research, a meticulous data cleansing process was conducted to prepare the datasets. This involved a series of steps, including the removal of non-Arabic words, numbers, symbols, Arabic and English stop words, duplicate characters, Arabic Tashkeel and Tanween, HTML tags, links, and the replacement of characters such as Hamza, Ha, and Ta Marbuta with their simplified equivalents. Tokenization includes breaking down text into smaller units, such as words or sub-words, with high precision. Stemming aims to reduce words to their base forms, known as stems or roots. At the same time, Lemmatization seeks to associate all word variations with their canonical form, called a lemma (the form found in dictionaries) [26]. Stemming and Lemmatization are essential for improving the consistency and accuracy of the SA process. This paper applies ISRI Light stemmer from NLTK to various datasets, successfully eliminating inflections and affixes to expose the base form. It uses a Qalsadi Arabic lemmatizer on data sets [27] [28], and its influence is compared with ISRI stemmers in improving LightGBM performance. It is combined with ISRI stemmer and compared with Qalsadi and ISRI alone. Table I shows comparison between the output of ISRI Stemmer, Qalsadi lemmatizer and ISRI-Qalsadi.

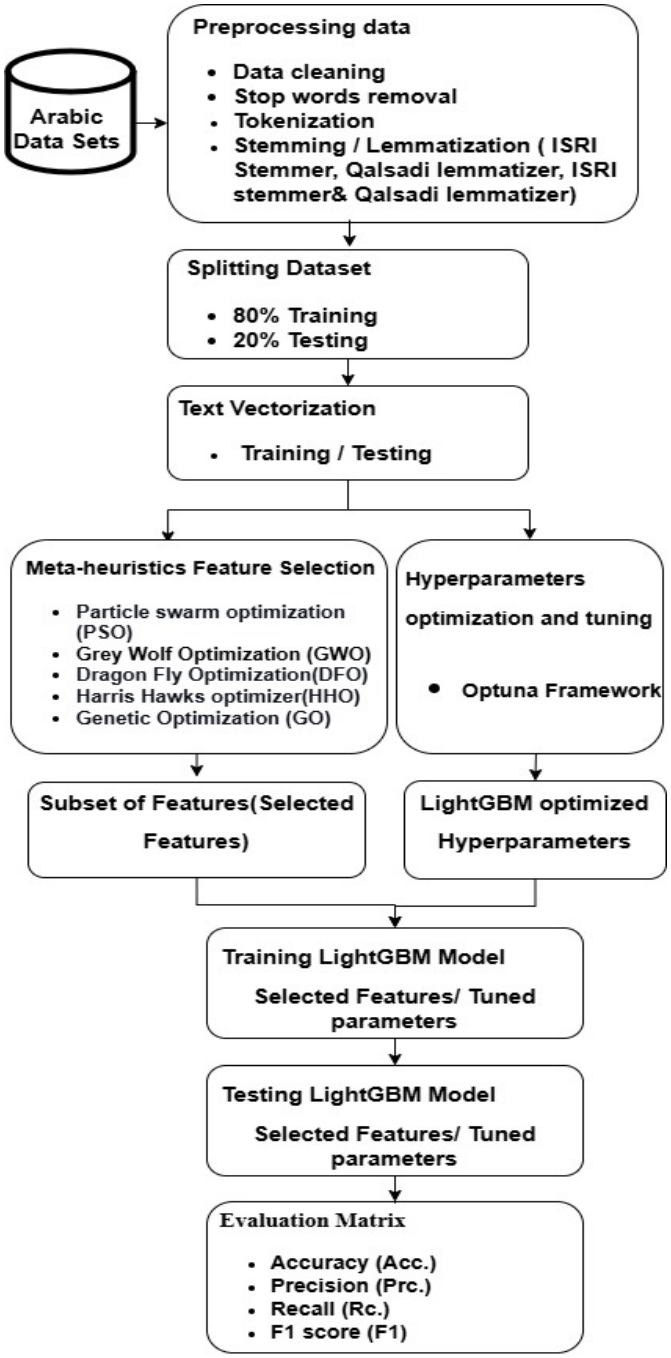


Fig. 1. Proposed methodology.

TABLE I. COMPARISON BETWEEN THE OUTPUT IF ISRI STEMMER, QALSADI LEMMATIZER AND ISRI-QALSADI

Word	ISRI Stemmer	Qalsadi lemmatizer	ISRI-Qalsadi
وبسواعدهما	وبسواعده	سواعد	سواعد
أوصيك	اوص	وصي	اوص
بالدقة	دقة	دقة	دقة
يعجبني	عجب	أعجب	عجب
النرجسيه	رجس	النرجسيه	رجس
الكتاب	كتب	كتاب	تب

B. Spliting Dataset

In this study, the dataset was divided into 80% training and 20% testing sets using the train_test_split method from Scikit-Learn. By using this method, it automatically shuffles the dataset. By shuffling the data, it will be distributed equally in the model, so it will be more accurate for predictions.

C. Text Vectorization

Text vectorization is the process of converting text into numerical representations, enabling their representation in a format suitable for ML models as a feature selection process [29]. In this research, the tokenizer class from Tensor Flow/Keras is used to build the vocabulary based on a list of input texts. It analyzes the texts, extracts unique words, assigns an integer index to each word in the vocabulary and convert it into sequences of indices, assign numerical indices to tokens, and pad sequences (usually zeros) to sequences shorter than the specified length to ensure that all sequences have the same length, which is necessary for processing by ML models [30].

Arabic texts = ["أنا أحب تعلم الآلة", "تحليل النصوص هو جزء من الذكاء الاصطناعي", "نحن نستخدم خوارزميات متعددة لتحليل البيانات"]

The result of text vectorization as below:

Word Index (Vocabulary): {'تحليل': 1, 'نحن': 2, 'نستخدم': 3, 'خوارزميات': 4, 'متعددة': 5, 'لتحليل': 6, 'البيانات': 7, 'أنا': 8, 'أحب': 9, 'تعلم': 10, 'الآلة': 11, 'النصوص': 12, 'هو': 13, 'جزء': 14, 'من': 15, 'الذكاء': 16, 'الاصطناعي': 17}.

Sequences: [[8, 9, 10, 11], [1, 12, 13, 14, 15, 16, 17], [2, 3, 4, 5, 6, 7]].

Padded Sequences: [[8 9 10 11 0 0] [1 12 13 14 15 16 17] [2 3 4 5 6 7 0]].

D. Arabic Text Classification using LightGBM Algorithm

LightGBM, developed by Microsoft, is an advanced gradient-boosting framework known for its faster training times and higher accuracy than other traditional gradient-boosting algorithms [29]. Its efficient memory usage allows it to handle large datasets with minimal resource requirements, resulting in improved performance and cost savings. LightGBM uses histogram-based learning and leaf-wise tree growth to enhance prediction accuracy [30]. It supports distributed GPU learning and parallel training on multi-core CPUs, making it suitable for big data applications. Additionally, its GOSS technique prioritizes critical data points during tree construction, reducing training time and memory usage. LightGBM is applicable for classification, regression, and ranking tasks [30].

The LightGBM algorithm can be represented mathematically as follows: Let X be the training dataset consisting of N examples and M features, and let Y represent the corresponding target values. $f(x_i)$ is defined as a function that maps the input features to the target values. The objective of LightGBM is to minimize the loss function $L(f)$, which measures the difference between the predicted values and the actual target values in relation to the function f as in Eq. (1).

$$L(f) = \sum[y_i - f(x_i)]^2 + \Omega(f) \quad (1)$$

An important regularization term, denoted as Ω(f), enhances the robustness of LightGBM by controlling the complexity of

the learned function and preventing overfitting. This term strikes a balance between effectively fitting the training data and generalizing to new data, ensuring a reliable and robust solution. LightGBM addresses this optimization problem by iteratively adding decision trees to the ensemble. At each iteration t, LightGBM constructs a decision tree h_t (x) that minimizes the loss function over a subset of the training examples St:

$$h_t(x) = argmin_h \sum[y_i - f_{t-1}(x_i) - h(x_i)]^2 + \Omega(h) \quad (2)$$

In LightGBM, the ensemble of decision trees from previous iterations, denoted as $f_{t-1}(x_i)$, is essential for the model's performance. Each new tree is trained to address the errors of the prior trees, enhancing predictions iteratively. This ongoing learning process ensures the model adapts and improves over time, reinforcing its reliability. LightGBM employs gradient boosting to optimize the loss function by sequentially adding decision trees. At each iteration (t), it calculates the negative gradient of the loss function based on the predictions from the existing ensemble, as expressed in Eq. (3).

$$g_i = -\partial L(f_{t-1}(x_i))/\partial f_{t-1}(x_i) \quad (3)$$

LightGBM utilizes the GOSS technique to enhance the training process by selecting a subset of examples. It prioritizes samples with large gradients to ensure their significance while under-sampling those with small gradients to lower computational costs and reduce the risk of overfitting. The algorithm employs a variant of the Gradient-based Decision Tree (GBDT), which constructs decision trees in a leaf-wise manner. In each split, it selects features that maximize loss reduction and prunes the tree based on a minimum gain threshold. This iterative method of adding trees continues until a stopping criterion is met, such as reaching a maximum number of trees or observing minimal improvement in validation error [16]. After training, LightGBM makes predictions by calculating the weighted average of the outputs from the individual trees as expressed in Eq. (4).

$$f(x) = \sum_{t=1}^{T} w_t h_t(x) \quad (4)$$

Where T is the number of trees in the ensemble, $\boldsymbol{w_t}$ is the weight of the t-th tree, and $\boldsymbol{h_t(x)}$ is the prediction of the t-th tree. The LightGBM determines their contribution to reducing the loss function as the weight.

E. Metaheuristics Feature Selection

FS is a vital step in ML that helps identify relevant variables related to target outcomes, improving model performance and control. Its key goals include enhancing generalization to reduce overfitting, eliminating redundant features for better inference, and enabling more efficient training with fewer features, shortening training times. Simpler models with fewer features are more easily interpreted [6] [7]. Metaheuristic algorithms, such as PSO, GWO, DFO, HHO, and GO, are practical tools for feature selection due to their reliability and efficiency [31]. However, they may not always guarantee global optimality. Among these, PSO is notable for its simplicity and efficiency in searching for optimal solutions without relying on gradients, making it a straightforward optimization tool with minimal hyperparameters. Inspired by natural behaviors like the collective movement of birds or fish, PSO effectively explores complex solution spaces to find optimal outcomes across various fields [31]. GWO is a recently developed evolutionary algorithm inspired by the social behavior of grey wolves, emphasizing the importance of pack dynamics in achieving reproductive success. In this model, a dominant male and female wolf hold higher ranks and guide the other pack members [32]. DFO is a new swarm intelligence algorithm inspired by the swarming behavior of dragonflies. It mimics five key principles: separation, cohesion, attraction, alignment, and distraction, which help dragonflies avoid collisions, maintain speed, connect with neighbors, seek food, and evade threats. DFO incorporates these behaviors into an optimization technique with two main phases: exploration and exploitation. These phases simulate the social interactions of dragonflies during navigation, food searching, and enemy avoidance in dynamic and static environments [33]. HHO enhances the effectiveness of wrapper-based FS techniques. As a fast and efficient swarm-based optimizer, HHO utilizes straightforward yet powerful exploratory and exploitative mechanisms, including Lévy flight and greedy selection. Additionally, it features a dynamic structure specifically designed for continuous problems. Its efficiency makes HHO a promising tool for a variety of optimization tasks, although it was originally developed for continuous search spaces [34]. GO is a highly effective computational method that is valuable in complex, poorly defined, or high-dimensional search spaces. Its primary goal in feature selection is to reduce the number of features by eliminating redundant and irrelevant ones while maintaining or improving classification accuracy. Various search algorithms have been employed for FS tasks [35].

F. Hyperparameter Optimization and Tuning

Hyperparameters play a crucial role in a model's functionality, performance, and structure, making their optimization essential for data scientists [9] [10]. The effectiveness of ML models, such as LightGBM, depends on selecting appropriate hyperparameter values, including learning rate, maximum depth, number of trees, and regularization parameters [36]. A systematic approach to hyperparameter tuning helps balance model complexity and generalization, improving accuracy and training speed. Optuna is noted as an advanced optimization framework that utilizes Bayesian techniques for more effective exploration of parameter spaces, allowing for fewer trials while managing experiments autonomously [10]. This capability enables Optuna to identify optimal hyperparameters that enhance model performance metrics like accuracy, precision, and recall, making hyperparameter optimization vital for maximizing a model's potential and achieving better results [36].

Algorithm 1: A simplified pseudocode for proposed methodology

```
# Step 1: Data Preprocessing
Input: Arabic Dataset
Output: Cleaned and tokenized dataset
BEGIN
  Perform Data Cleaning
    - Remove unnecessary data (e.g., duplicates, special characters)
  Remove Stop Words
  Perform Tokenization
  Apply Stemming/Lemmatization
    - Use ISRI Stemmer
    - Use Qalsadi Lemmatizer
```

```
    - Use ISRI Stemmer & Qalsadi Lemmatizer
END

# Step 2: Split Dataset
Input: Preprocessed dataset
Output: Training set (80%) and Testing set (20%)
BEGIN
  Split dataset into 80% Training and 20% Testing
END

# Step 3: Text Vectorization
Input: Training and Testing sets
Output: Vectorized text data
BEGIN
  Vectorize Text Data for Training and Testing
END

# Step 4: Feature Selection using Meta-heuristics
Input: Vectorized training data
Output: Subset of selected features
BEGIN
  Initialize meta-heuristic optimization algorithms:
    - Particle Swarm Optimization (PSO)
    - Grey Wolf Optimization (GWO)
    - Dragon Fly Optimization (DFO)
    - Harris Hawks Optimization (HHO)
    - Genetic Optimization (GO)
  Perform FS
    - Identify and retain the most relevant features
  Output Selected Features
END

# Step 5: Hyperparameter Optimization and Tuning
Input: Vectorized training data and feature subset
Output: Optimized hyperparameters
BEGIN
  Use Optuna Framework for Hyperparameter Tuning
  Optimize LightGBM's parameters
END

# Step 6: Train LightGBM's Model
Input: Selected features, optimized hyperparameters
Output: Trained LightGBM's model
BEGIN
  Train LightGBM's model using:
    - Selected Features
    - Optimized Hyperparameters
END

# Step 7: Test LightGBM's Model
Input: Testing data, trained model
Output: Evaluation metrics
BEGIN
  Test LightGBM's model
  Evaluate Performance using:
    - Accuracy (Acc.)
    - Precision (Prc.)
    - Recall (Rc.)
    - F1 Score (F1)
END
```

IV. Experiments and Results

This section presents and discusses datasets used in experiments, working environment and experiment setting and classification results and performance evaluation.

A. Datasets Description

The experiments and comparison results use eleven datasets in Table II from GitHub and Kaggle. "qrci " was downloaded from [37]. "Ar_reviews_100k" is a much larger dataset with 100,000 rows and 99999 tweets/reviews. It combines reviews from hotels, books, movies, products, and airlines and was downloaded from [38]. "ARABIC Dataset" was downloaded from [39]. It contains 58751 Arabic tweets. It has three classes (natural, negative, and positive). "ARABIC Dataset_2cat" is "ARABIC Dataset" after removing natural tweets. "mpqa-ar" is an Arabic opinion corpus containing articles from many news sources annotated for opinions downloaded from [37]. LABR is a large ASA dataset. It consists of over 63,000 book reviews [40]. After balancing it, it has 16448 reviews. It was downloaded as a balanced dataset from [37]. "Astd-artwitter" is a combined dataset between ASTD and Artwitter data sets downloaded from [37]. ASTD was downloaded from [37] with 1590 tweets. ASA_SS2030 is a dataset related to social events in the Arabic Saudi Dialect associated with Saudi Arabia's 2030 vision and downloaded from [41]. AJGT introduces an Arabic Jordanian General Tweets (AJGT) Corpus in MSA or Jordanian dialect [42]. "Company Reviews" were collected for SA to produce a score for companies [43]. It has 40K+ reviews in Arabic for SA. It has reviews for Ezz Steel, Talbat, Elsewedy, Hilton, Nestle, Raya, SWVL, Telecom Egypt, TMG, Venus, Domty, and Capiter companies. Table II shows Data Sets Information.

TABLE II. Data Sets Information

Date Set	No Rows	No. Tweets/Reviews	No Categories	No. Positives	No. Negatives	No. Neutral	No. Features
qrci	755	754	2	377	377	0	10
ar_reviews_100k	100000	99999	3	33333	33333	33333	41
ARABIC Dataset_2cat	56498	56497	2	29460	27037	0	10
ARABIC Dataset	58752	58751	3	29460	27037	2254	10
mpqa-ar	9997	9996	2	5399(subjectives)	0	4597	16
LABR-book-reviews	16449	16448	2	8224	8224	0	42
astd-artwitter	3543	3542	2	1771	1771	0	10
ASTD	1590	1591	2	777	812	0	14
ASA_SS2030	4253	4252	2	2436	1816	0	21
AJGT	1801	1800	2	900	900	0	8
Company Reviews	40046	40045	3	23921	14200	1925	8

As shown in Table II, the datasets vary widely in size, ranging from a few thousand to over 100,000 rows. The number of categories also varies, with most datasets having two categories but some having three. The distribution of positive, negative, and neutral examples is only sometimes balanced, especially in larger datasets. The number of features also varies across the datasets, with some having as few as eight features and others having as many as 42.

B. Working Environment and Experimental Setting

The experiments have been done in google Colab using several python libraries such as pandas, numpy, NLTK, Qalsadi, TensorFlow, Scikit-learn, pandas, LightGBM, PSO, GWO, DFO, HHO, GO, zoofs and Optuna framework. Colab is a hosted Jupyter Notebook service (SaaS Service) that offers free access to GPUs and TPUs among other computing resources. It does not require any setup. Colab works particularly effectively with ML, data science, and teaching.

TABLE III. HYPERPARAMETERS SPACE SEARCH CONFIGURATION FOR THE LIGHTGBM MODEL

Model	Hyperparameter settings
LightGBM	search_space ={ 'boosting_type': Categorical(['gbdt', 'dart', 'goss']),'max_depth': Integer(1, 750), 'num_leaves': Integer(2, 400),'learning_rate': Real(0.01, 1.0, 'log-uniform'), 'subsample': Real(0.1, 1.0, 'uniform'),'n_estimators': Integer(50, 1500), 'min_child_samples': Integer(1, 100),'colsample_bytree': Real(0.1, 1.0, 'uniform'), 'reg_alpha': Real(1e-9, 100, 'log-uniform'), 'max_bin': Integer(100, 700) 'reg_lambda': Real(1e-9, 100, 'log-uniform'), , 'max_delta_step': Real(0, 10, 'uniform') }

Table III presents the hyperparameters search space configuration for the LightGBM model, covering a wide range of values to optimize performance. The search space includes categorical choices for boosting_type (GBDT, DART, GOSS), and numerical ranges for key parameters such as num_leaves (2 to 400), max_depth (1 to 750), and learning_rate (0.01 to 1.0) with a logarithmic uniform distribution to explore both small and large values effectively. Other parameters include n_estimators, min_child_samples, subsample, and colsample_bytree, which control model complexity and generalization. Additionally, regularization parameters such as reg_alpha and reg_lambda are optimized within a logarithmic scale to prevent overfitting. The inclusion of max_bin and max_delta_step further refines the model's handling of data granularity and convergence stability. This comprehensive hyperparameter tuning strategy aims to enhance LightGBM's adaptability and accuracy across diverse Arabic sentiment analysis datasets.

C. Classification Results and Performance Evaluation

This section presents and discusses the experiments of the LightGBM classification model, metaheuristic FS algorithms, and Optuna hyperparameter optimization. The experiments in this study are divided into three main dimensions: studying the effect of ISRI stemming and Qalsadi lemmatization methods on LightGBM' s classification, both individually and in combination with the classification efficiency on different datasets; studying the effects of metaheuristic FS algorithms; and studying the impact of Optuna hyperparameter optimization in the classification task.

Experiment 1: In the first experiment, the ISRI stemming and Qalsadi lemmatization methods and their combination with LightGBM are applied to eleven datasets as shown in table IV.

Table V outlines the hyperparameter settings used for running various metaheuristic algorithms to optimize feature selection in the sentiment analysis task. Each algorithm is configured with a common objective function, log loss, to minimize classification error, and a consistent number of iterations (20) and population size (20) to ensure fair comparisons. Specific parameter settings are applied to individual algorithms to enhance their optimization efficiency. For instance, PSO includes acceleration constants and an inertia weight for balancing exploration and exploitation. GO incorporates selective pressure, elitism, and mutation rate to guide the search process. Meanwhile, GWO, DFO, and HHO follow standard configurations focused on convergence towards optimal feature subsets. These settings ensure a robust evaluation of different optimization techniques in improving the model's performance.

Experiment 2: The second experiment is conducted to study the effects of different metaheuristic feature selection algorithms as shown in Tables VI, VII, VIII, XI, IX, X, XI.

TABLE IV. COMPARISON BETWEEN ISRI-LIGHTGBM, QALSADI-LIGHTGBM, AND ISRI-QALSADI-LIGHTGBM

Datasets	ISRI-LightGBM				Qalsadi-LightGBM				ISRI-Qalsadi-LightGBM			
	Acc.	Prc	Rc	F1	Acc.	Prc	Rc.	F1	Acc.	Prc	Rc	F1
qrci	55.6	56	56	56	56.3	56	56	56	**59**	**59**	**59**	**59**
ar_reviews_100k	62.1	63	62	62	**66.6**	**67**	**67**	**67**	57.3	58	57	57
ARABIC Dataset_2cat	**70.2**	**70**	**70**	**70**	69	69	69	69	69.6	70	70	70
ARABIC Dataset	**70**	**71**	**70**	**70**	69.8	71	70	70	69	70	69	69
mpqa-ar	60.9	61	61	61	**61.2**	**61**	**61**	**61**	59.8	59	60	59
LABR-book-reviews	64.6	65	65	64	65.3	65	65	65	**65.7**	**66**	**66**	**65**
astd-artwitter	67.3	68	67	67	**67.4**	**68**	**67**	**67**	63.1	63	63	63
ASTD	**58.5**	**59**	**58**	**58**	58	58	58	58	57.6	58	58	57
ASA_SS2030	71.1	71	71	71	71.9	72	72	72	**74.4**	**74**	**74**	**74**
AJGT	**73.9**	**74**	**74**	**74**	67.8	68	68	68	63.6	64	64	64
Company Reviews	**76.7**	72	**77**	**74**	75.2	**75**	75	73	76.6	73	77	74

TABLE V. HYPERPARAMETERS SETTINGS FOR RUNNING METAHEURISTIC ALGORITHMS

Algorithm	Hyper parameters
PSO	objective_function= log_loss, population_size=20, n_iteration=20, minimize=True, constant accelerator 1=2, constant accelerator 2=2,weight=0.9
GWO	objective_function= log_loss, population_size=20, n_iteration=20, minimize=True
DFO	objective_function= log_loss, population_size=20, n_iteration=20,method='linear', minimize=True
HHO	objective_function= log_loss, population_size=20,n_iteration=20, minimize=True
GO	objective_function= log_loss, population_size=20, n_iteration=20,selective_pressure=2 ,elitism=2,mutation_rate=0.05,minimize=True

TABLE VI. FEATURE SELECTION (FS) OF EACH ALGORITHM IN EACH DATASET

Data Set	Stemming / Lemmatization	Light GBM No .F.	PSO-Light GBM FS.	GWO –Light GBM FS.	DFO-Light GBM FS.	HHO-Light GBM FS.	GO –Light GBM FS.
qrci	ISRI	10	1	6	1	7	1
	Qalsadi	10	1	7	3	6	2
	ISRI- Qalsadi	10	3	8	5	3	5
ar_reviews_100k	ISRI	41	30	41	32	36	39
	Qalsadi	41	33	41	32	39	33
	ISRI- Qalsadi	41	30	38	31	37	31
ARABIC Dataset_2cat	ISRI	10	10	9	10	10	10
	Qalsadi	10	8	10	9	8	8
	ISRI+ Qalsadi	10	9	8	9	9	8
ARABIC Dataset	ISRI	10	10	10	9	10	8
	Qalsadi	10	10	9	10	9	8
	ISRI- Qalsadi	10	8	8	8	7	9
mpqa-ar	ISRI	16	8	12	12	9	8
	Qalsadi	16	9	15	11	11	12
	ISRI- Qalsadi	16	10	16	9	12	9
LABR-book-reviews	ISRI	42	17	34	20	26	22
	Qalsadi	42	19	39	22	31	16
	ISRI- Qalsadi	42	14	41	21	30	23
astd-artwitter	ISRI	10	7	9	10	10	9
	Qalsadi	10	5	9	5	7	5
	ISRI- Qalsadi	10	3	10	3	7	6
ASTD	ISRI	14	6	10	4	8	6
	Qalsadi	14	6	6	6	6	6
	ISRI- Qalsadi	14	2	11	5	8	6
ASA_SS2030	ISRI	21	14	18	10	13	11
	Qalsadi	21	14	31	15	16	27
	ISRI-Qalsadi	21	13	15	13	12	12
AJGT	ISRI	8	3	7	3	4	6
	Qalsadi	8	3	6	4	5	3
	ISRI- Qalsadi	8	1	5	1	4	3
Company Reviews	ISRI	8	7	8	7	7	5
	Qalsadi	8	6	7	6	6	7
	ISRI- Qalsadi	8	5	7	5	5	5

TABLE VII. COMPARISON BETWEEN ISRI-PSO - LIGHTGBM, QALSADI-PSO - LIGHTGBM, AND ISRI-QALSADI-PSO - LIGHTGBM

Datasets	ISRI-PSO –LightGBM				Qalsadi-PSO- LightGBM				ISRI-Qalsadi-PSO - LightGBM			
	Acc.	Prc	Rc	F1	Acc.	Prc	Rc	F1	Acc.	Prc	Rc	F1
qrci	58.3	59	58	57	60.9	62	61	59	**64.9**	**65**	**65**	**65**
ar_reviews_100k	61.4	62	61	61	**65.6**	**66**	**66**	**66**	57.3	58	57	57
ARABIC Dataset_2cat	**70.2**	**70**	**70**	**70**	69	69	69	69	69.6	70	70	70
ARABIC Dataset	**70**	**71**	**70**	**70**	69.8	71	70	70	69.5	70	70	69
mpqa-ar	61.4	61	61	61	**61.7**	**61**	**62**	**61**	**61.7**	**61**	**62**	**61**
LABR-book-reviews	65.8	66	66	65	65.3	65	65	65	**67.7**	**68**	**68**	**67**
astd-artwitter	67.6	68	68	68	**69.4**	**70**	**69**	**69**	66.9	67	67	67
ASTD	62	62	62	62	**63.8**	**64**	**64**	**64**	61.6	62	62	62
ASA_SS2030	75	75	75	75	**75.7**	**75**	**77**	**75**	**75.7**	**75**	**76**	**76**
AJGT	**76.1**	**76**	**76**	**76**	71.1	72	71	71	65.8	66	66	66
Company Reviews	**76.7**	**76**	**77**	**74**	75.4	74	75	73	76.6	74	77	74

TABLE VIII. COMPARISON BETWEEN ISRI- GWO - LIGHTGBM, QALSADI- GWO- LIGHTGBM, AND ISRI-QALSADI- GWO - LIGHTGBM

Datasets	ISRI-GWO - LightGBM				Qalsadi-GWO - LightGBM				ISRI-Qalsadi-GWO - LightGBM			
	Acc.	Prc	Rc	F1	Acc.	Prc	Rc	F1	Acc.	Prc	Rc	F1
qrci	60.9	61	61	61	57	57	57	57	**62.9**	**63**	**63**	**63**
ar_reviews_100k	62.1	63	62	62	**66.6**	**67**	**67**	**67**	57.1	58	57	57
ARABIC Dataset_2cat	**69.8**	**70**	**70**	**70**	69	69	69	69	68.5	69	68	68
ARABIC Dataset	**70**	**71**	**70**	**70**	69.2	70	69	69	69.6	70	70	69
mpqa-ar	60.5	60	60	60	**60.9**	**60**	**61**	**60**	59.8	59	60	59
LABR-book-reviews	64.7	65	65	64	64.3	64	64	64	**65.8**	**66**	**66**	**66**
astd-artwitter	**67.1**	**68**	**67**	**67**	64.7	65	65	65	63.1	63	63	63
ASTD	60.7	61	61	61	**63.8**	**64**	**64**	**64**	55.7	56	56	56
ASA_SS2030	73.9	74	74	74	72.4	72	72	72	**75.1**	**75**	**75**	**75**
AJGT	**75.3**	**75**	**75**	**75**	70.6	71	71	71	64.7	65	65	65
Company Reviews	**76.7**	72	**77**	74	75.4	73	75	73	76.4	**75**	77	**75**

TABLE IX. COMPARISON BETWEEN ISRI- DFO - LIGHTGBM, QALSADI- DFO - LIGHTGBM, AND ISRI-QALSADI- DFO - LIGHTGBM

Datasets	ISRI- DFO - LightGBM				Qalsadi- DFO - LightGBM				ISRI-Qalsadi- DFO - LightGBM			
	Acc.	Prc	Rc	F1	Acc.	Prc	Rc	F1	Acc.	Prc	Rc	F1
qrci	58.3	59	58	57	**63.6**	**64**	**64**	**64**	61.6	62	62	62
ar_reviews_100k	61.2	62	61	61	**64.8**	**65**	**65**	**65**	57.3	58	57	57
ARABIC Dataset_2cat	**70.2**	**70**	**70**	**70**	69.3	69	69	69	69.6	70	70	70
ARABIC Dataset	**70.4**	**71**	**70**	**70**	69.8	71	70	70	69.5	70	70	69
mpqa-ar	61.9	62	62	62	**62.2**	**62**	**62**	**62**	60.8	60	61	61
LABR-book-reviews	64.8	65	65	65	64.6	65	65	64	**67.5**	**68**	**67**	**67**
astd-artwitter	**67.3**	**68**	**67**	**67**	69.4	70	69	69	66.9	67	67	67
ASTD	**66.7**	**67**	**67**	**67**	63.8	64	64	64	64	64	64	64
ASA_SS2030	**75.1**	**75**	**75**	**75**	73.7	74	74	74	74	74	74	74
AJGT	**76.1**	**76**	**76**	**76**	70.8	71	71	71	65.8	66	66	66
Company Reviews	**76.7**	**76**	**77**	**74**	75.4	74	75	73	**76.7**	**74**	**77**	**75**

TABLE X. COMPARISON BETWEEN ISRI- HHO - LIGHTGBM, QALSADI- HHO - LIGHTGBM, AND ISRI-QALSADI- HHO - LIGHTGBM

Datasets	ISRI- HHO - LightGBM				Qalsadi- HHO - LightGBM				ISRI-Qalsadi- HHO - LightGBM			
	Acc.	Prc	Rc	F1	Acc.	Prc	Rc	F1	Acc.	Prc	Rc	F1
qrci	57	57	57	57	60.3	60	60	60	**62.9**	**63**	**63**	**63**
ar_reviews_100k	62	62	62	62	**66.4**	**67**	**66**	**66**	57.5	58	57	57
ARABIC Dataset_2cat	**70.2**	**70**	**70**	**70**	69	69	69	69	69.6	70	70	70
ARABIC Dataset	**70**	**71**	**70**	**70**	68.8	70	69	69	68.6	69	69	68
mpqa-ar	**60.4**	**60**	**60**	**60**	58.4	58	58	58	59.5	59	59	59
LABR-book-reviews	63	63	63	63	64.2	64	64	64	**65.9**	**66**	**66**	**66**
astd-artwitter	**67.3**	**68**	**67**	**67**	66.3	67	66	66	62.9	63	63	63
ASTD	63.8	64	64	64	63.8	64	64	64	54.1	54	54	54
ASA_SS2030	71.8	72	72	72	74.4	74	74	74	**74.9**	**75**	**75**	**75**
AJGT	58.6	59	59	59	**66.9**	**67**	**67**	**67**	65.3	65	65	65
Company Reviews	**76.7**	**76**	**77**	**74**	71.8	71	72	69	76.6	74	77	74

TABLE XI. COMPARISON BETWEEN ISRI- GO - LIGHTGBM, QALSADI- GO - LIGHTGBM, AND ISRI-QALSADI- GO - LIGHTGBM

Datasets	ISRI- GO - LightGBM				Qalsadi- GO - LightGBM				ISRI-Qalsadi- GO - LightGBM			
	Acc.	Prc	Rc	F1	Acc.	Prc	Rc	F1	Acc.	Prc	Rc	F1
qrci	58.3	59	58	57	60.3	60	60	60	**64.2**	**66**	**64**	**63**
ar_reviews_100k	60.6	61	61	60	**64.9**	**65**	**65**	**65**	57.4	58	57	57
ARABIC Dataset_2cat	**70.2**	**70**	**70**	**70**	69	69	69	69	69.5	70	70	70
ARABIC Dataset	**70**	**71**	**70**	**70**	69.5	70	69	69	69	70	69	69
mpqa-ar	61.7	61	62	62	**62.2**	**62**	**62**	**62**	60.8	60	61	60
LABR-book-reviews	66	66	66	66	65.7	66	66	65	**66.7**	**67**	**67**	**66**
astd-artwitter	66.4	67	66	66	**65.6**	**66**	**66**	**66**	65.4	66	65	65
ASTD	62.6	63	63	63	**63.8**	**64**	**64**	**64**	59.4	60	59	59
ASA_SS2030	73	74	74	74	**75.7**	**75**	**76**	**76**	74.9	75	75	75
AJGT	**74.7**	**75**	**75**	**75**	71.1	72	71	71	66.1	66	66	66
Company Reviews	**76.7**	**73**	**77**	**74**	75.7	74	76	74	76.1	73	76	74

TABLE XII. HYPERPARAMETER SETTINGS FOR RUNNING OPTUNA FRAMEWORK

Method	Hyperparameters
Optuna study	'objective': 'binary', 'metric': 'binary_logloss', 'num_leaves': trial.suggest_int('num_leaves', 2, 256) 'lambda_l1': trial.suggest_loguniform('lambda_l1', 1e-8, 10.0), 'lambda_l2': trial.suggest_loguniform('lambda_l2', 1e-8, 10.0), 'bagging_fraction': trial.suggest_uniform('bagging_fraction', 0.4, 1.0), ,'feature_fraction': trial.suggest_uniform('feature_fraction', 0.4, 1.0), 'bagging_freq': trial.suggest_int('bagging_freq', 1, 7),'min_child_samples': trial.suggest_int('min_child_samples', 5, 100)

Table XII presents the hyperparameter settings used for running the Optuna framework, which is employed to optimize the LightGBM model for Arabic sentiment analysis. The optimization process is guided by the binary classification objective with the evaluation metric set to binary_logloss, ensuring a focus on minimizing classification errors. The search space for key hyperparameters includes lambda_l1 and lambda_l2 for regularization, both explored within a logarithmic range to prevent overfitting. Structural parameters such as num_leaves (ranging from 2 to 256) and min_child_samples (ranging from 5 to 100) are fine-tuned to balance model complexity and generalization. Additionally, feature and data sampling parameters, including feature_fraction and bagging_fraction, are optimized within uniform distributions to enhance model robustness. The bagging_freq parameter, which controls the frequency of bagging operations, is also explored to improve model stability. These hyperparameter settings enable efficient and automated tuning to achieve optimal model performance.

D. Evaluation Matrix

Evaluation metrics are used to assess the effectiveness and performance of statistical or ML models. They help illustrate how well the model's predictions align with the true patterns in the dataset. The key metrics for evaluating ML models include Accuracy (Acc.), Precision (Prc.), Recall (Rc.), and the F1 score (F1). These metrics, calculated using specific equations, are critical in the context of Deep ASA evaluation metrics [16]:

$$\text{Accuracy (Acc.)} = (TP+ TN) / (TP + TN + FP + FN) \quad (1)$$

$$\text{Precision (Prc.)} = TP\ (TP + FP) \quad (2)$$

$$\text{Recall (Rc.)} = TP\ (TP + FN) \quad (3)$$

$$\text{F1 Score(F1.)} = 2 * \text{Prc.} * \text{Rc.}\ (\text{Prc.} + \text{Rc.}) \quad (4)$$

Where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively.

The primary goal of Experiment 2 is to compare the effectiveness of different metaheuristics FS algorithms. Analyzing values allows us to identify which algorithms and feature selection techniques selected the most relevant features for each dataset that increase ASA accuracy, precision, recall, and F1 score. Comparing the results with stemming, lemmatization or a combination of them can help assess the effect of text preprocessing on feature selection. Comparing the results for all datasets can provide insights into how the algorithms s perform on different data characteristics.

Experiment 3: The second experiment is conducted to study studying the impact of Optuna hyperparameter optimization in ASA as shown in Table XIII.

TABLE XIII. COMPARISON BETWEEN ISRI-OPTUNA - LIGHTGBM, QALSADI-OPTUNA - LIGHTGBM, AND ISRI-QALSADI-OPTUNA - LIGHTGBM

Datasets	ISRI-Optuna - LightGBM				Qalsadi-Optuna - LightGBM				ISRI-Qalsadi-Optuna - LightGBM			
	Acc.	**Prc**	**Rc**	**F1**	**Acc.**	**Prc**	**Rc**	**F1**	**Acc.**	**Prc**	**Rc**	**F1**
qrci	64.2	64	64	64	66	65	65	65	**66.9**	**67**	**67**	**67**
ar_reviews_100k	**78**	**78**	**78**	**78**	68.2	68	68	68	59.9	60	60	60
ARABIC Dataset_2cat	**77.2**	**77**	**77**	**77**	76	76	76	76	76.5	77	77	77
ARABIC Dataset	**76.7**	**77**	**76**	**76**	75.2	76	76	76	76.2	77	76	76
mpqa-ar	64	62	62	62	**64.2**	**63**	**63**	**63**	62.5	62	62	62
LABR-book-reviews	67.6	66	66	66	68.1	68	68	68	**70.9**	**71**	**71**	**71**
astd-artwitter	70	68	68	68	**71.1**	**72**	**71**	**71**	70	70	70	70
ASTD	**64.2**	**64**	**64**	**64**	63.5	64	64	63	63	64	63	62
ASA_SS2030	75	75	75	75	76.1	76	76	76	**78.5**	**78**	**78**	**78**
AJGT	**77.8**	**78**	**78**	**78**	74.2	75	74	74	70.5	71	71	71
Company Reviews	78.2	76	78	76	77.3	75	77	75	**78.4**	**76**	**78**	**76**

Table XIV shows the optimal hyperparameter values found by Optuna for the LightGBM algorithm using stemming, lemmatization, or both methods for each dataset. The optimal hyperparameter values and accuracy scores vary between datasets, highlighting the importance of dataset-specific tuning. The choice of stemming or lemmatization can influence the optimal hyperparameters and accuracy. The relative importance of different hyperparameters can vary depending on the dataset and algorithm.

TABLE XIV. OPTUNA-LIGHTGBM HYPERPARAMETERS WITH THE BEST ACCURACY USING STEMMING / LEMMATIZATION OR BOTH METHODS FOR EACH DATASET

Data set	Algorithms	Optuna-LightGBM Hyperparameters		Trial	Acc.
qrci	ISRI-Qalsadi-Optuna - LightGBM	learning_rate	0.07899	118	66.9
		num_leaves	119		
		max_depth	17		
		min_child_samples	76		
		subsample	0.696647		
		colsample_by tree	0.865725		
		n_estimators	693		
ar_reviews_100k	ISRI-Optuna - LightGBM	learning_rate	0.035947	165	77.9
		num_leaves	119		
		max_depth	43		
		min_child_samples	53		
		subsample	0.674727		
		colsample_by tree	0.524262		
		n_estimators	793		
ARABIC Dataset_2cat	ISRI-Optuna - LightGBM	learning_rate	0.072340	501	77.3
		num_leaves	227		
		max_depth	15		
		min_child_samples	26		
		subsample	0.512467		
		colsample_by tree	0.553126		
		n_estimators	860		
ARABIC Dataset	ISRI-Optuna - LightGBM	learning_rate	0.027153	424	76.7
		num_leaves	214		
		max_depth	22		
		min_child_samples	9		
		subsample	0.684298		
		colsample_by tree	0.723168		
		n_estimators	872		
mpqa-ar	Qalsadi-Optuna - LightGBM	learning_rate	0.015316	281	64.2
		num_leaves	24		

		max_depth	33		
		min_child_samples	35		
		subsample	0.766491		
		colsample_by tree	0.760554		
		n_estimators	955		
LABR-book-reviews	ISRI-Qalsadi-Optuna - LightGBM	learning_rate	0.083975	112	70.3
		num_leaves	237		
		max_depth	11		
		min_child_samples	6		
		subsample	0.620707		
		colsample_by tree	0.677474		
		n_estimators	843		
astd-artwitter	Qalsadi-Optuna - LightGBM	learning_rate	0.070423	257	71.1
		num_leaves	94		
		max_depth	6		
		min_child_samples	9		
		subsample	0.743444		
		colsample_by tree	0.774086		
		n_estimators	603		
ASTD	ISRI-Optuna - LightGBM	learning_rate	0.034131	919	64.5
		num_leaves	7		
		max_depth	48		
		min_child_samples	5		
		subsample	0.82711		
		colsample_by tree	0.581206		
		n_estimators	920		
ASA_SS2030	ISRI-Qalsadi-Optuna - LightGBM	learning_rate	0.095296	676	78.5
		num_leaves	29		
		max_depth	39		
		min_child_samples	26		
		subsample	0.946048		
		colsample_by tree	0.804287		
		n_estimators	917		
AJGT	ISRI-Optuna - LightGBM	learning_rate	0.040334	593	77.8
		num_leaves	84		
		max_depth	17		
		min_child_samples	7		
		subsample	0.736873		
		colsample_by tree	0.587376		
		n_estimators	263		
		learning_rate	0.0256	212	78.4
Company Reviews	ISRI-Qalsadi-Optuna - LightGBM	num_leaves	179		
		max_depth	26		
		min_child_samples	72		
		subsample	0.83006		
		colsample_by tree	0.71024		
		n_estimators	673		

The figures below depict bar charts summarizing the classification results and performance evaluation for the best models in the three experiments mentioned above on eleven data sets. It summarizes all classification results and performance evaluation tables. For the "qrci" dataset, the model ISRI-Qalsadi-Optuna– LightGBM achieved an accuracy score of approximately 67%. Optuna with ISRI-Qalsadi increase LightGBM's overall accuracy by 8%, but PSO metaheuristics feature selection, with ISRI stemming, increases LightGBM's overall accuracy by 6% as shown in Fig. 2.

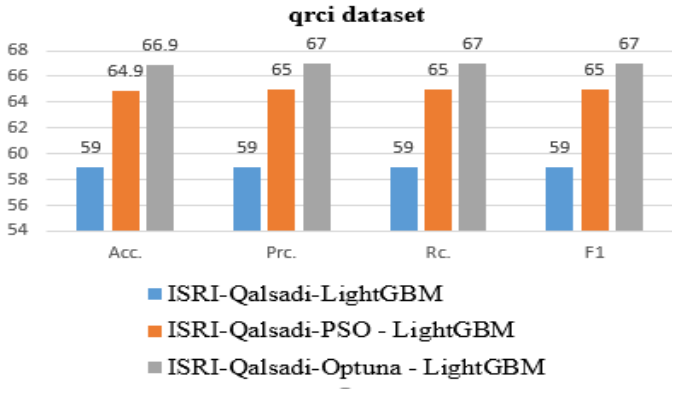


Fig. 2. Comparison between the best in the three experiments for "qrci" dataset.

For the "ar_reviews_100k" dataset, the model ISRI-Optuna - LightGBM achieved an accuracy score of approximately 78%. Optuna with ISRI stemming increase LightGBM's overall accuracy by 11% despite GWO metaheuristics feature selection, with Qalsadi lemmatization having the same value as Qalsadi-LightGBM with an accuracy score of approximately 67% as shown in Fig. 3.

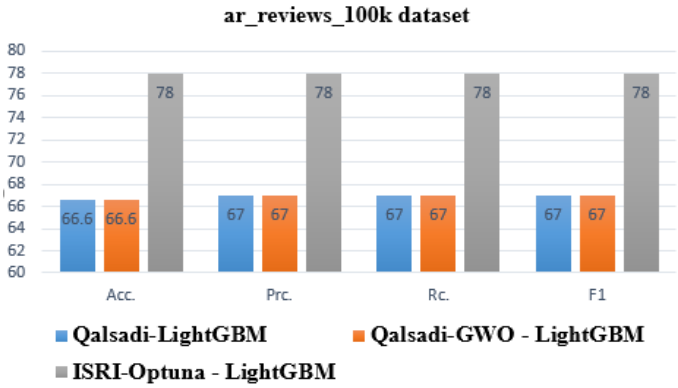


Fig. 3. Comparison between the best in the three experiments for "ar_reviews_100k" dataset.

For the "ARABIC Dataset_2cat" dataset, the model ISRI-Optuna - LightGBM achieved an accuracy score of roughly 77%. Optuna with ISRI stemming increase LightGBM's overall accuracy by 7%. Despite PSO metaheuristics feature selection with ISRI stemming, it has the same value as ISRI-LightGBM with an accuracy score of approximately 70% as shown in Fig. 4.

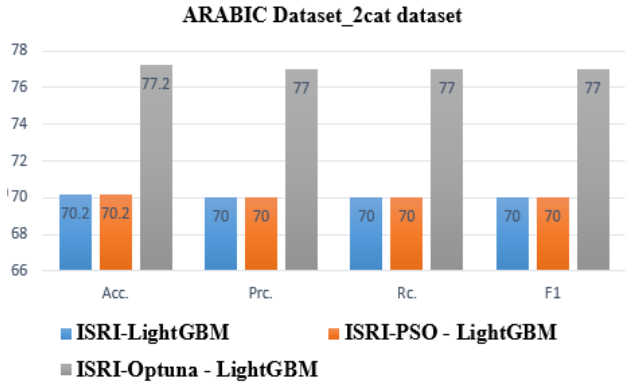


Fig. 4. Comparison between the best in the three experiments for "ARABIC Dataset_2cat" dataset.

For the "ARABIC Dataset" dataset, the model ISRI-Optuna - LightGBM achieved an accuracy score of roughly 77%. Optuna with ISRI stemming increase LightGBM's overall accuracy by 7% despite DFO metaheuristics feature selection, with ISRI stemming increasing LightGBM's overall accuracy by 0.4% as shown in Fig. 5.

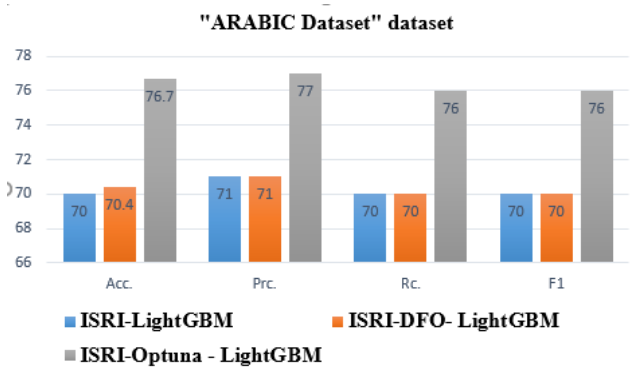


Fig. 5. Comparison between the best in the three experiments for "ARABIC Dataset" dataset.

For the "mpqa-ar" dataset, the model Qalsadi-Optuna - LightGBM achieved an accuracy score of approximately 64%. Optuna with Qalsadi lemmatization increase LightGBM's overall accuracy by 3% despite DFO metaheuristics feature selection, with Qalsadi lemmatization increasing LightGBM's overall accuracy by 1% as shown in Fig. 6.

For the "LABR-book-reviews" dataset, the model ISRI-Qalsadi-Optuna – LightGBM achieved an accuracy score of approximately 71%. Optuna with ISRI-Qalsadi increase LightGBM's overall accuracy by 5% despite PSO metaheuristics feature selection, with ISRI-Qalsadi increasing LightGBM's overall accuracy by 2% as shown in Fig. 7.

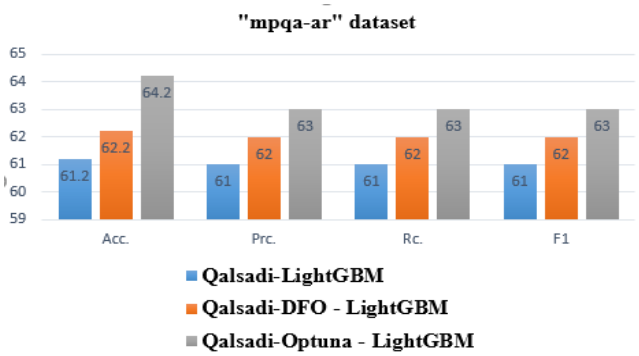


Fig. 6. Comparison between the best in the three experiments for "mpqa-ar" dataset.

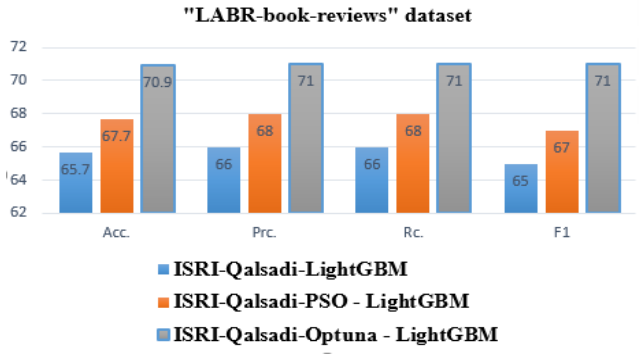


Fig. 7. Comparison between the best in the three experiments for "LABR-book-reviews" dataset.

For the "astd-artwitter" dataset, the model Qalsadi-Optuna - LightGBM achieved an accuracy score of approximately 71%. Optuna with ISRI-Qalsadi increase LightGBM's overall accuracy by 4% despite PSO and DFO metaheuristics feature selection, with Qalsadi increasing LightGBM's overall accuracy by 2% as shown in Fig. 8.

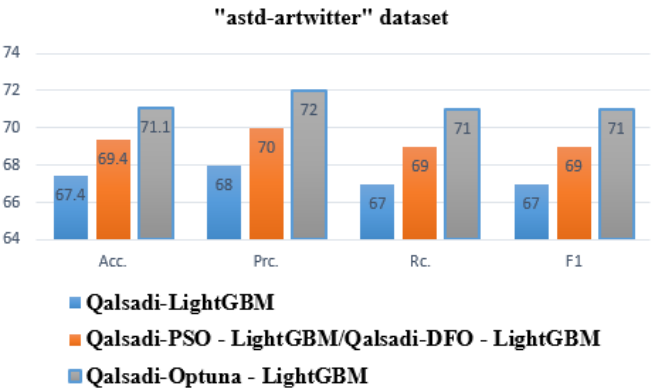


Fig. 8. Comparison between the best in the three experiments for "astd-artwitter" dataset.

For the "ASTD" dataset, the model ISRI-DFO- LightGBM achieved an accuracy score of approximately 64%. DFO metaheuristics feature selection with ISRI increases LightGBM's overall accuracy by 8% despite Optuna with ISRI increasing LightGBM's by 6% as shown in Fig. 9.

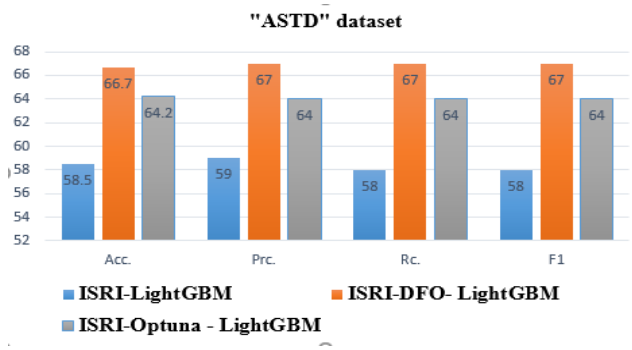


Fig. 9. Comparison between the best in the three experiments for "ASTD" dataset.

For the "ASA_SS2030" dataset, the model ISRI-Qalsadi-Optuna - LightGBM achieved an accuracy score of approximately 78.5%. Optuna with ISRI-Qalsadi increase LightGBM's overall accuracy by 4% despite DFO metaheuristics feature selection with ISRI-Qalsadi and GO metaheuristics feature selection, with Qalsadi increasing LightGBM's overall accuracy by 1.5% as shown in Fig. 10.

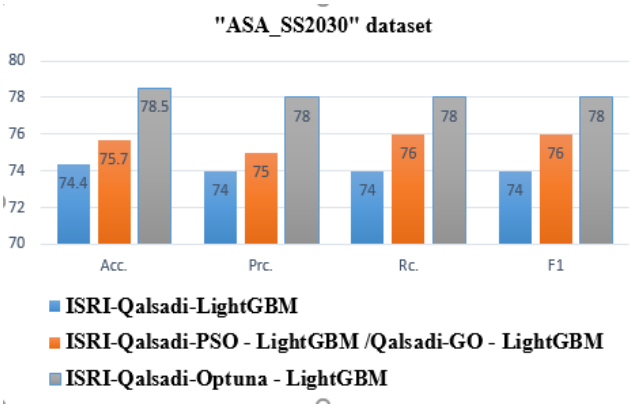


Fig. 10. Comparison between the best in the three experiments for "ASA_SS2030" dataset.

For the "AJGT" dataset, the model ISRI-Optuna - LightGBM achieved an accuracy score of approximately 78%. Optuna with ISRI increase LightGBM's overall accuracy by 4% despite DFO metaheuristics feature selection, with ISRI and PSO metaheuristics feature selection, with ISRI increasing LightGBM's overall accuracy by 2% as shown in Fig. 11.

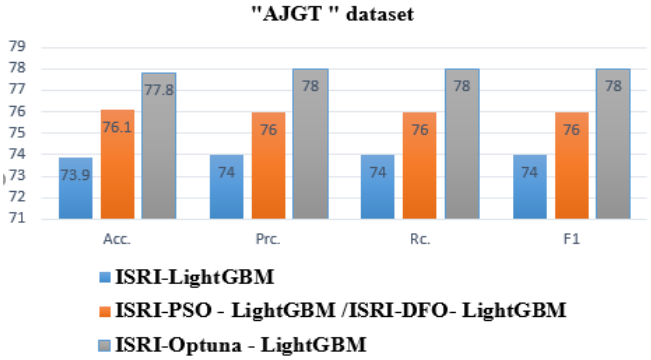


Fig. 11. Comparison between the best in the three experiments for "AJGT" dataset.

For the "Company Reviews" dataset, the model ISRI-Qalsadi-Optuna - LightGBM achieved an accuracy score of approximately 78.4%. Optuna with ISRI increase LightGBM's overall accuracy by 1.7% as shown in Fig. 12. Still, DFO metaheuristics feature selection with ISRI, PSO metaheuristics feature selection with ISRI, and HHO metaheuristics feature selection with ISRI have the same value as ISRI-LightGBM. The results demonstrate the effectiveness of hyperparameter optimization using Optuna-LightGBM in improving LightGBM's performance on ASA. By carefully tuning the hyperparameters, you can significantly improve accuracy and generalization.

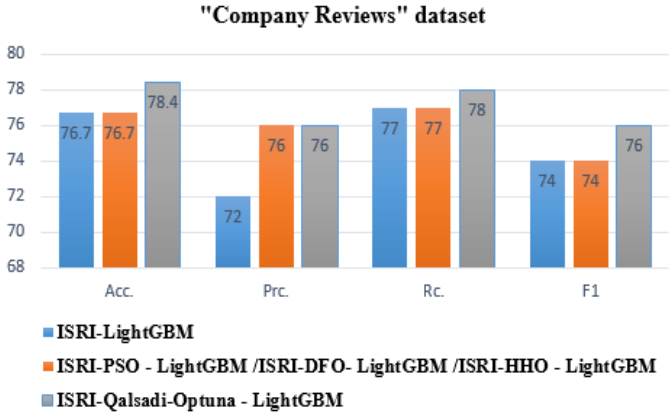


Fig. 12. Comparison between the best in the three experiments for "Company Reviews" dataset.

V. Discussion

This study contributes to the field of ASA by presenting an integrated approach to enhance the accuracy of the LightGBM model through advanced text preprocessing, feature selection, and hyperparameters optimization techniques. The complexity of the Arabic language, with its numerous dialects and MSA, poses a significant challenge for sentiment analysis applications. To address this, the study employs the ISRI stemmer and Qalsadi lemmatizer for effective text preprocessing, alongside metaheuristic FS algorithms such as PSO, GWO, and others to identify the most informative features and reduce noise in the data. Additionally, the study leverages the Optuna framework for hyperparameters tuning, aiming to achieve an optimal balance between computational efficiency and model performance. The findings demonstrate that combining these methodologies can enhance the classification accuracy of LightGBM by up to 11%, highlighting the effectiveness of these strategies in improving ASA.

However, ASA faces numerous challenges that contribute to low accuracy compared to other languages, such as English. These challenges include the morphological complexity of the language, characterized by extensive inflection, derivation, and multiple word forms that increase the difficulty of automated text analysis. Additionally, the scarcity of high-quality labeled datasets covering diverse Arabic dialects limits the model's ability to generalize effectively across various user demographics. The coexistence of MSA and dialectal Arabic, informal writing styles on social media, and the lack of standardized linguistic resources further complicate the

analysis. This study provides a valuable contribution by exploring potential solutions to these issues, such as using different stemming and lemmatization methods, optimizing models through feature selection, and fine-tuning hyperparameters to achieve higher accuracy in real-world applications.

VI. Conclusion

Analyzing Arabic content poses challenges due to the language's complexities, morphological features, inadequate resources, and the absence of suitable corpora. This study delves into the effectiveness of various preprocessing techniques—ISRI stemming, Qalsadi lemmatization, and their combination—on ASA. The study's primary focus is on the LightGBM classifier, and It systematically compared these methods to find that each preprocessing approach contributes positively to sentiment classification accuracy. Depending on the dataset, using metaheuristic feature selection algorithms significantly enhanced the performance of the LightGBM model by identifying the most relevant features, thus reducing noise and improving LightGBM's classification efficiency between 0 and 8%. PSO metaheuristic feature selection algorithm with suitable stemming between ISRI and Qalsadi or a combination for LightGBM achieves superior results than GWO, DFO, HHO, and GO metaheuristic feature selection in more than 60% of used datasets. DFO metaheuristic feature selection algorithm with suitable stemming between ISRI and Qalsadi or a combination for LightGBM achieves superior results than other metaheuristic feature selection in more than 35% of used datasets. Applying the Optuna hyperparameter optimization framework further demonstrated the potential to refine LightGBM model parameters, effectively resulting in substantial performance gains. Depending on the dataset, Optuna using suitable stemming between ISRI and Qalsadi or a combination improves LightGBM's accuracy by between 2 and 11% and achieves superior results than PSO, GWO, DFO, HHO, and GO metaheuristic feature selection in more than 90% of used datasets. Our findings highlight the critical role that preprocessing and optimization strategies play in ASA. These methodologies improve classification accuracy and highlight the LightGBM model's robustness in this domain. ASA faces challenges such as the morphological complexity of the language, the scarcity of high-quality labeled datasets, and the coexistence of MSA and dialectal Arabic, which hinder its classification accuracy compared to languages like English, but this research explores solutions like advanced stemming, metaheuristics feature selection, and Optuna hyperparameter fine-tuning to improve performance. This research underscores the necessity for continued exploration of advanced techniques in ASA. The potential for future research to explore additional ML, DL models, transformers, and large language models to enhance ASA applications across diverse contexts and rebalance unbalanced used datasets to have higher accuracy is vast and inspiring.

Acknowledgment

This Project was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, under grand no. (GPIP: 92-611-2024). The authors therefore, acknowledge with thanks, DSR for technical and financial support.

References

[1] A. Harjai, A. Charan, and S. Singhal, “Sentiment Analysis of Medications Review Using Deep Learning Algorithms,” in Proceedings of the 5th International Conference on Information Management & Machine Intelligence, November 2023, pp. 1–5.

[2] A. Sayed, M. M. Gomaa, and M. M. Nazier, “Sentiment Analysis on Twitter's Big Data Against the Covid-19 Pandemic Using Machine Learning Algorithms,” Inf. Sci. Lett., vol. 12, no. 8, pp. 2747–2756, 2023.

[3] S. M. Almutairi and F. M. Alotaibi, “A Comparative Analysis of Arabic Sentiment Analysis Models in E-Marketing Using Deep Learning Techniques,” Journal of Engineering and Applied Sciences, vol. 10, no. 1, 2023.

[4] F. Alzamzami and A. E. Saddik, “OSN-MDAD: Machine Translation Dataset for Arabic Multi-Dialectal Conversations on Online Social Media,” arXiv preprint arXiv:2309.12137, 2023.

[5] M. Jbel, I. Hafidi, and A. Metrane, “Sentiment Analysis Dataset in Moroccan Dialect: Bridging the Gap Between Arabic and Latin Scripted Dialect,” arXiv preprint arXiv:2303.15987, 2023.

[6] B. K. Lavine, “Feature Selection: Introduction,” Elsevier EBooks, 2009. [Online]. Available: https://doi.org/10.1016/b978-044452701-1.00028-4.

[7] B. Tran, B. Xue, and M. Zhang, “Overview of Particle Swarm Optimisation for Feature Selection in Classification,” in Simulated Evolution and Learning: 10th International Conference, SEAL 2014, Dunedin, New Zealand, Dec. 15-18, 2014, Proceedings, pp. 605-617. Springer International Publishing.

[8] A.G. Gad, K.M. Sallam, R.K. Chakrabortty, M.J. Ryan, and A.A. Abohany, “An improved binary sparrow search algorithm for feature selection in data classification,” Neural Computing and Applications, vol. 34, no. 18, pp. 15705–15752, 2022.

[9] B. Bischl et al., “Hyperparameter Optimization: Foundations, Algorithms, Best Practices, and Open Challenges,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 13, no. 2, p. e1484, 2023.

[10] T. Akiba et al., “Optuna: A Next-Generation Hyperparameter Optimization Framework,” in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2623-2631, Jul. 2019.

[11] S. M. Khabour, Q. A. Al-Radaideh, and D. Mustafa, “A New Ontology-Based Method for Arabic Sentiment Analysis,” Big Data and Cognitive Computing, vol. 6, no. 2, p. 48, 2022.

[12] S. Al-Hagree and G. Al-Gaphari, “Arabic Sentiment Analysis on Mobile Applications Using Levenshtein Distance Algorithm and Naive Bayes,” in 2022 2nd International Conference on Emerging Smart Technologies and Applications (eSmarTA), October 2022, pp. 1–6. IEEE.

[13] L. Almurqren, R. Hodgson, and A. Cristea, “Arabic Text Sentiment Analysis: Reinforcing Human-Performed Surveys with Wider Topic Analysis,” arXiv preprint arXiv:2403.01921, 2024.

[14] M. Atabuzzaman, M. Shajalal, M. B. Baby, and A. Boden, “Arabic Sentiment Analysis with Noisy Deep Explainable Model,” in Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval, December 2023, pp. 185–189.

[15] G. Ke et al., “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” Advances in Neural Information Processing Systems, vol. 30, 2017.

[16] E. Essa, K. Omar, and A. Alqahtani, “Fake News Detection Based on a Hybrid BERT and LightGBM Models,” Complex & Intelligent Systems, vol. 9, no. 6, pp. 6581-6592, 2023.

[17] A. Harjai, A. Charan, and S. Singhal, “Sentiment Analysis of Medications Review Using Deep Learning Algorithms,” in Proceedings of the 5th International Conference on Information Management & Machine Intelligence, pp. 1-5, Nov. 2023.

[18] F. Yang, L. Hu, and Y. Chen, “Product Competitiveness and Sentiment Analysis Based on DSM Clustering of Chinese Online Comparative Comments,” U.P.B. Sci. Bull., Series C, vol. 83, no. 4, 2021.

[19] B. O. Ondara et al., "Hybrid Machine Learning Techniques for Comparative Opinion Mining," Indonesian Journal of Artificial Intelligence and Data Mining, vol. 6, no. 2, pp. 131-143.

[20] T. O. Omotehinwa, D. O. Oyewola, and E. G. Moung, "Optimizing the Light Gradient-Boosting Machine Algorithm for an Efficient Early Detection of Coronary Heart Disease," Informatics and Health, vol. 1, no. 2, pp. 70-81, 2024.

[21] M. Heikal, M. Torki, and N. El-Makky, "Sentiment Analysis of Arabic Tweets Using Deep Learning," Procedia Computer Science, vol. 142, pp. 114-122, 2018.

[22] S. Al-Azani and E. S. M. El-Alfy, "Hybrid Deep Learning for Sentiment Polarity Determination of Arabic Microblogs," in Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, Nov. 14-18, 2017, Proceedings, Part II, pp. 491-500. Springer International Publishing.

[23] S. R. El-Beltagy et al., "Combining Lexical Features and a Supervised Learning Approach for Arabic Sentiment Analysis," in CICLing 2016, Konya, Turkey, 2016.

[24] M. Aly and A. Atiya, "Labr: A Large Scale Arabic Book Reviews Dataset," in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 494-498, Aug. 2013.

[25] M. Abdou, S. AbdelGaber, M. Farhan, and O. Othman, "Predicting Human Behavior Using Arabic Sentiment Analysis on Social Media: Approaches and Challenges," Informatics Bulletin, Faculty of Computers and Artificial Intelligence, Helwan University, Published Online, 2023.

[26] I. Zeroual and A. Lakhouaja, "Arabic Information Retrieval: Stemming or Lemmatization?" in 2017 Intelligent Systems and Computer Vision (ISCV), pp. 1-6, Apr. 2017.

[27] T. Zerrouki, "Towards an Open Platform for Arabic Language Processing," 2020.

[28] T. Zerrouki, "Qalsadi: Arabic Morphological Analyzer Library for Python," 2012. [Online]. Available: https://pypi.python.org/pypi/qalsadi.

[29] "LightGBM: A Comprehensive Guide," [Online]. Available: https://medium.com/@pelinokutan/lightgbm-a-comprehensive-guide-cb773cfc23b3.

[30] "LightGBM: Light Gradient Boosting Machine," [Online]. Available: https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/.

[31] N. Van Thieu and S. Mirjalili, "MEALPY: An Open-Source Library for Latest Meta-Heuristic Algorithms in Python," Journal of Systems Architecture, vol. 139, p. 102871, 2023.

[32] E. Emary, H. M. Zawbaa, and A. E. Hassanien, "Binary Grey Wolf Optimization Approaches for Feature Selection," Neurocomputing, vol. 172, pp. 371-381, 2016.

[33] A. I. Hammouri et al., "An Improved Dragonfly Algorithm for Feature Selection," Knowledge-Based Systems, vol. 203, p. 106131, 2020.

[34] T. Thaher et al., "Binary Harris Hawks Optimizer for High-Dimensional, Low Sample Size Feature Selection," in Evolutionary Machine Learning Techniques: Algorithms and Applications, pp. 251-272, 2020.

[35] F. Hussein, N. Kharma, and R. Ward, "Genetic Algorithms for Feature Selection and Weighting: A Review and Study," in Proceedings of the Sixth International Conference on Document Analysis and Recognition, pp. 1240-1244, Sep. 2001.

[36] B. Xiang et al., "Based on the WSP-Optuna-LightGBM Model for Wind Power Prediction," in Journal of Physics: Conference Series, vol. 2835, no. 1, p. 012011, Aug. 2024.

[37] "Arabic Embeddings Datasets," [Online]. Available: https://github.com/iamaziz/ar-embeddings/tree/master/datasets.

[38] "Arabic 100K Reviews Dataset," [Online]. Available: https://www.kaggle.com/datasets/abedkhooli/arabic-100k-reviews.

[39] M. Saad, "Arabic Sentiment Analysis Dataset," GitHub repository, 2019. [Online]. Available: https://github.com/motazsaad/arabic-sentiment-analysis/tree/master/arabic_tweets_tsv/20190413/3labels. [Accessed: Oct. 15, 2023]

[40] "LABR Dataset," [Online]. Available: https://paperswithcode.com/dataset/labr.

[41] "Arabic Sentiment Analysis Dataset SS2030," [Online]. Available: https://www.kaggle.com/datasets/snalyami3/arabic-sentiment-analysis-dataset-ss2030-dataset.

[42] "Arabic Twitter Corpus AJGT," [Online]. Available: https://github.com/komari6/Arabic-twitter-corpus-AJGT.

[43] "Arabic Company Reviews Dataset," [Online]. Available: https://www.kaggle.com/datasets/fahdseddik/arabic-company-reviews/code.

Flexible Framework for Lung and Colon Cancer Automated Analysis Across Multiple Diagnosis Scenarios

Marwen SAKLI[1], Chaker ESSID[2], Bassem BEN SALAH[3], Hedi SAKLI[4]
SERCOM Laboratory-Tunisia Polytechnic School, University of Carthage, La Marsa 2078, Tunisia[1, 2, 3]
FST, Campus Universitaire El-Manar, 2092 El Manar Tunis, Tunisia[2]
National Engineering School of Tunis-ENIT Communication Systems Research Laboratory SYSCOM-LR-99-ES21, University of Tunis El-Manar, Tunisia[4]
EITA Consulting, 7 Rue Du Chant Des Oiseaux, 78360 Montesson, France[4]

***Abstract*—Among humans, lung and colon cancers are regarded as primary contributors to mortality and morbidity. They may grow simultaneously in organs, having a harmful influence on the lives of people. If tumor is not diagnosed early, it is likely to spread to both of those organs. This research presents a flexible framework that employs lightweight Convolutional Neural Networks architecture for automating lung and colon cancer diagnosis in histological images across multiple diagnosis scenarios. The LC25000 dataset is commonly used for this task. It includes 25000 histopathological images belonging to 5 distinct classes, which are lung adenocarcinoma, lung squamous cell carcinoma, benign lung tissue, colon adenocarcinoma, and benign colonic tissue. This work includes three diagnosis scenarios: (S1) evaluates lung or colon samples, (S2) distinguishes benign from malignant images, and (S3) classifies images into five categories from the LC25000 dataset. Across all the scenarios, the scored accuracy, recall, precision, F1-score, and AUC exceeded 0.9947, 0.9947, and 0.9995, respectively. This investigation with a lightweight Convolutional Neural Network containing only 1.612 million parameters is extremely efficient for automated lung and colon cancer diagnosis, outperforming several current methods. This method might help doctors provide more accurate diagnoses and improve patient outcomes.**

Keywords*—Lung and colon cancers; histopathological images; LC25000 dataset; lightweight convolutional neural networks; multiple diagnosis scenarios

I. INTRODUCTION

Statistical analysis undertaken in the United States showed that lung and colon (LC) cancers are expected to be among the three most prevalent cancer types in 2020. Moreover, these malignancies were expected to have the most fatality rates of any cancer diagnosis. The GLOBOCAN 2020 data indicated LC cancer incidence rates of 11.4% and 18.0%, respectively [1]. The World Health Organization (WHO) anticipated that roughly 4 Million persons on a global scale will acquire lung or colon cancer in 2020, resulting in approximately 2.7 Million deaths. The presented information highlights the substantial worldwide health effects of lung and colon cancers. It is worth noting that LC cancers can coexist, with roughly 17% of cases containing both tumors concurrently [2].

Lung cancer, a malignant disease, arises from the excessive and unregulated multiplication of atypical cells within the lung [3]. This can result in tumor formation, which may spread to other parts of the body. Various factors play a role in the increasing incidence of lung cancer, including exposure to harmful substances, such as tobacco smoke, and aging. Early-stage lung cancer often presents with subtle or no symptoms, making early detection challenging [4]. Consequently, diagnosis frequently occurs at a late stage, when therapeutic options are curtailed. Adenocarcinoma and squamous cell carcinoma constitute the majority of lung cancer cases [5]. Adenocarcinoma, which can affect both smokers and non-smokers, is more prevalent in women and younger individuals It often originates in the outermost regions of the pulmonary tissue and can spread rapidly. Squamous cell carcinoma, primarily associated with smoking, can develop anywhere in the lungs and tends to grow and spread aggressively [6,7].

The causes of LC cancer are multiple and complex. Smoking is the main risk factor for lung cancer, while for colon cancer, a low-fiber diet, prolonged sedentary lifestyle obesity, and certain genetic factors can increase the risk of other environmental factors such as exposure to certain chemical substances or air pollution contribute in the multiplication of these cancers. It is crucial to note that these risk factors are not synonymous with inevitable cancer development, but adopting them can considerably increase the chances of developing these diseases.

Generally, to detect and diagnose cancer, a variety of diagnostic tests are employed, including imaging modalities such as Magnetic Resonance Imaging (MRI) [8-11], X-rays [12], CT scans, and dermoscopy [13-16], as well as tissue sampling procedures such as biopsies. Histological images offer considerable advantages over other types of medical imaging in the analysis and characterization of LC cancers. Histology enables microscopic analysis of tissues removed during biopsy or surgery. This enables us to observe cancer cells directly, and determine their type, stage, and aggressiveness. This information is crucial for making an accurate diagnosis, choosing the most appropriate treatment, and assessing prognosis. Additional imaging techniques, such as radiography, computed tomography, or MRIs, provide information on the anatomy and size of tumors but do not allow detailed analysis of

cellular characteristics. The microscopic examination of tissue parts by experienced pathologists is crucial for determining the presence of cancer cells and classifying their type and subtype [17,18]. Although, manual analysis of histopathological images consumes time. Also, it is labor-intensive and subjective process, prone to inter-observer variability. Pathologists may have differing interpretations of the same image, leading to potential diagnostic errors, especially when dealing with subtle morphological features. Additionally, the growing number of medical images and the complexity of certain cases further exacerbate the challenge.

To overcome these constraints [19], this work investigates the application of Deep Learning (DL) and machine learning (ML) approaches to automate cancer analysis in medical images [12,20]. DL techniques provide potential solutions to mitigate these challenges. Through the utilization of deep neural networks, DL models can process high quantities of histopathological images, learning to recognize complex characteristics and features associated with LC cancers. This can lead to improved diagnostic and efficiency and precision in comparison with traditional approaches.

This study tries to balance dependability as well as precision in LC cancer classification. The main contributions of this study include:

- Proving that DL approaches can effectively diagnose and analyze LC cancers.
- Employing a huge dataset of 25000 histopathological images to classify LC cancers.
- Illustrating three diagnosis scenarios to ensure the flexibility of the presented framework.
- Designing a lightweight CNN model with only 1.6 million parameters, assessing its performance, while comparing it to current approaches.
- Achieving accuracy, F1-score, and AUC over 99.47%, 99.47%, and 99.95%, respectively, throughout all analytic phases and diagnosis scenarios.
- Scoring F1-score and accuracy of 99.17% and 99.47%. Also, a sensitivity and specificity of 99.07%, and 99.65% across all classes of overall diagnosis scenarios, respectively.

This investigation employs a classification approach to analyze a dataset containing lung and colon cancer images. The subsequent sections of this investigation are organized as follows: the second section presents a review of pertinent literature. The third section delineates the proposed methodology. The fourth section provides a detailed performance evaluation of the presented lightweight CNN. The fifth section offers a thorough discussion of the findings. This paper will be ended by the last section, Section VI.

II. Related Work

This study contributes to the research effort aimed at improving diagnostic support for LC cancer using artificial intelligence techniques. A new method for diagnosing histopathological images on the LC25000 dataset [21], a reference in the field. Several academics have recently used this dataset to develop AI-based applications.

Sakr et al. [22] introduce a lightweight DL approach using a CNN for powerful categorization of cancer of the colon. Histopathological images were normalized before being processed by the CNN model. Across two classes, their proposed system attained a high accuracy of 99.50%. Using the same data, Hasan et al. [23] presented an innovative DL strategy for the automated identification of colon adenocarcinomas. Their approach involved a DCNN model, coupled with several image preprocessing techniques, to obtain meaningful features from digital histological images. The proposed system demonstrated impressive performance, achieving a maximum accuracy of 99.80% in differentiating between non-cancerous and cancerous tissues. For the same target, Gabralla et al [24] introduced a novel stacking-based deep learning framework. Their approach involved integrating multiple pre-trained CNN models (ResNet50 [25], DenseNet121 [26], InceptionV3, and VGG16 [27]) with a meta-learner. The meta-learner was trained to effectively combine the predictions of the individual models, resulting in a significant improvement in overall performance. The proposed method achieved an ideal score of 100% in terms of F1-score and accuracy, using the LC25000 dataset [21]. In addition, they attained F1-score and accuracy of 98% when employing the WCE dataset [28,29], surpassing the performance of the individual base models. To enhance the accuracy of colon cancer prediction, Di Giammarco et al. [30] employed different pre-trained models. The metrics of the proposed method was assessed on a sub-dataset of LC25000 comprising 10,000 colon images. The experimental results demonstrated that MobileNet [31] had the highest, f1-score, accuracy, recall, and precision of 99.9%, indicating the model's capability of the model to efficiently categorize colon issues.

Concerning lung cancer, Hatuwal and Thapa [32] aimed to classify three classes of tissue: benign, squamous cell carcinoma, and adenocarcinoma. A CNN approach was trained and validated on a sub-dataset of histological data, LC25000 dataset [21]. The model demonstrated strong performance, achieving an accuracy of 96.11% and 97.2%, during the training and validation phases. Nishio et al. [33] established a CAD system to automate the analysis of lung tissue in histological images. The system employed a multi-stage approach involving image feature extraction and ML classification. Two feature extraction techniques were investigated: conventional texture analysis (TA) and homology-based image processing (HI). Eight ML algorithms were trained and evaluated using the extracted features. The results found in the experiments demonstrated the higher efficiency of the HI-based approach over the TA-based system, achieving an accuracy of 99.33%. The study of Hamed et al. [34] is about a novel system for the rapid and precise classification of lung tissue histological data. The treated tissue types were only benign and squamous cell carcinoma. The proposed approach involves a two-stage process: feature extraction using a lightweight CNN model and classification using a LightGBM classifier. The CNN technique, designed with a minimal number of parameters, efficiently extracts discriminative features from the preprocessed images. Subsequently, the LightGBM classifier, leveraging multiple threads, effectively classifies the input data into various tissue

types. When evaluated on the LC25000 dataset, the approach achieved a remarkable accuracy of 99.6% and a sensitivity of 99.6%. To increase lung cancer classification accuracy, Noaman et al [35] propose a novel hybrid feature extraction technique where the powerful capabilities of charateristic extraction of DenseNet201 was combined with the complementary information provided by color histograms. A comprehensive evaluation of eight machine learning algorithms, including, SVM, MultinomialNB, LGBM, CatBoost, XGBoost, KNN, and RF, was conducted on the LC25000 dataset [21]. The outcomes demonstrate that the established hybrid feature set, when coupled with an appropriate ML algorithm, achieves a remarkable accuracy of 99.683%. To further validate the generalizability of our approach, we applied it to the task of the analysis of the breast cancer utilizing the images of the BreakHis dataset [36]. The model achieved a high accuracy of 94.808%, highlighting the advantage of their hybrid feature extraction technique for various medical image analysis tasks.

Several researchers tried to classify the whole of the five types of tissues figured in the LC25000 dataset [21]. In fact, Ali et al. [37] reached 99.04% and 99.58% as F1-score and accuracy LC cancer classification. They employed a multi-input dual-stream Capsule Network (CapsNet) [38]. It consists of two major blocks: Convolutional Layers Block (CLB) and Separable Convolutional Layers Block (SCLB). CLB and SCLB uses traditional and separable convolutional layers. The SCLB block takes uniquely preprocessed images using gamma correction and color balancing. Also, it takes multi-scale fusion and image sharpening. This dual-input approach enhances feature learning. Besides CapsNet, numerous works applied Efficient Networks (EfficientNets) [39] for LC cancer classification. Masud et al [40] present a novel DL-based framework for the diagnostic of five distinct classes of LC tissues, containing both benign and malignant conditions. By leveraging advanced digital image processing techniques and DL models, the proposed framework effectively extracts relevant features from histopathological images and accurately classifies them. Experimental results demonstrate that the developed tool can detect cancer tissues with a high accuracy and F1-score with values of 96.33% and 96.38%, respectively. In Mehmood et al.'s study [41], a pre-trained AlexNet was adapted for the task of histological image classification. The initial model, trained using a generic dataset, achieved promising results for most image classes, except for one class where the accuracy was relatively low. To address this issue, the simple and effective technique of contrast enhancement was applied to enhance the quality of images from the underperforming class. This targeted approach significantly boosted the overall accuracy of the model to 98.4% while maintaining computational efficiency. Attallah et al. [42] developed a novel framework that integrated DL and feature reduction techniques to ameliorate the accuracy of the classification of histopathology images. To extract relevant features, histopathology scans were processed using three models: ShuffleNet [43], MobileNet [31], and SqueezeNet [44]. The high-dimensional feature vectors obtained from these models are then subjected to Fast Walsh-Hadamard Transform (FWHT) and dimensionality reduction by the application of Principal Component Analysis (PCA). To further enhance feature representation, Discrete Wavelet Transform (DWT) was employed to combine the reduced characteristics from the three DL models. The resulting reduced and fused feature sets are subsequently passed into four different ML algorithms for classification. The established framework achieves F1-score and accuracy of 99.6% on the given dataset. The study of Al-Jabbar et al. [45] introduces three novel strategies for the early diagnostic of the lung cancer based on the LC25000 dataset. To enhance image quality and improve diagnostic accuracy, preprocessing techniques were applied to enhance the contrast of affected areas. Subsequently, high-dimensional patterns were determined using the VGG-19 [27] and GoogLeNet [46] models. To reduce dimensionality and retain crucial information, Principal Component Analysis (PCA) was employed. The first strategy involved training separate Artificial Neural Networks (ANN) models using the features extracted from VGG-19 and GoogLeNet. The other approach combined the patterns from both models before applying dimensionality reduction and ANN classification. The third strategy, which yielded the best performance, involved fusing the features extracted from VGG-19, GoogLeNet, and handcrafted characteristics before training the ANN model. This approach achieved a sensitivity, specificity, precision, accuracy, and AUC surpassing 99.64%. Kumar et al. [47] did a comparative analysis to measure the effectiveness of handcrafted and DL-based feature extraction techniques for LC cancer classification. In this research, six handcrafted pattern extraction methods were employed to capture color, texture, shape, and structural information from histopathological images. These handcrafted features were then used to train and evaluate 4 ML classifiers: Gradient Boosting, MLP, Random Forest, and SVM-RBF. In another approach, seven pre-trained DL models are utilized to determine high-level patterns from data. These deep features were subsequently fed into the same four ML classifiers. The findings demonstrated that the classification performance was significantly enhanced when using DL-based features compared to handcrafted features. Notably, the Random Forest categorizer combined with DenseNet-121 achieved the highest ROC-AUC, accuracy, and F1-score with values exceeding 91%. In Anjum et al's work [48] EfficientNet models (B0 to B7) [39] were applied for the diagnostic of LC cancer in histopathological data. To improve model performance and mitigate overfitting, transfer learning, and parameter tuning techniques were employed. After preprocessing the LC25000 dataset [21] to remove noise and standardize image formats, experiments were conducted using different image resolutions, ranging from 224x224 pixels to 600x600 pixels. The models were evaluated based on classification accuracy and loss. While all EfficientNet [39] variants achieved promising results, EfficientNetB2 demonstrated the highest performance, attaining an accuracy of 97.24% when trained on 260x260 pixel images.

Some studies did not concentrate on creating automated diagnostic approaches specifically for colon or lung, or lung and colon cancer. Rather, they developed methods that can comprehensively address the diagnosis of colon cancer, lung cancer, or both. The work of Talkuder et al. [49] identified efficiently LC cancers by the employment of a hybrid ensemble method. The proposed model integrates powerful feature extraction approaches with ensemble learning and high-performance filtering to effectively analyze histopathological images from the LC25000 dataset [21]. The results demonstrate the superior performance of their hybrid model, reaching

accuracies of 100%, 99.05%, and 99.30% for colon, lung, and combined LC cancer classification, respectively. In the research of Hage Chehade et al. [50], they aimed to develop a computerized diagnostic system capable of efficiently categorizing the five distinct classes of LC tissues, including two types of colon cancer and three categories of lung cancer. When leveraging ML techniques, feature engineering, and image processing methods, meaningful information was extracted from histopathological data. The LC25000 dataset [21] was utilized to train and evaluate five ML models: Random Forest, XGBoost, SVM, Multilayer Perceptron, and Linear Discriminant Analysis. The best results were obtained using XGBoost. For colon cancer, they reached accuracy and F1-score of 99.3% and 99.5%, respectively. Concerning lung cancer, the accuracy, precision, recall, and F1-score reached were 99.53%, 99.33%, 99.33%, and 99.33%, respectively. Also, for LC cancer, they got an accuracy of 99%, precision of 98.6%, recall of 99%, and F1-score of 98.8%.

This work aims to develop a flexible based on multi-scenario diagnosis for LC cancer automated analysis. This framework is based on a lightweight CNN and uses the LC25000 dataset to evaluate its performance in this task. For the first scenario, a method was developed to detect if the input image corresponds to a lung or a colon. The results were perfect since all the performance metrics had attained 100%. Based on this result, two methods were established. The first approach is dedicated to determining the nature of the colon tissue type. Its AUC, F1-score, and accuracy reached 100%. The second method aims to classify lung cancer. It got AUC, F1-score, and accuracy with values of 99. 95%, 99.47%, and 99. 47%, respectively. The target of the second scenario is LC malignancy detection in histological images. This approach scored perfect AUC, F1-score, and accuracy with values of 100%, respectively. The purpose of the third scenario is to classify the totality of LC tissue types. AUC, F1-score, and accuracy achieved values of 99.96%, 99.76%, and 99.76%, respectively. Table I presents the description of the cited methods from the literature and the proposed approaches.

III. Proposed Method

In this study, three diagnosis scenarios are presented. The first scenario (S1) is composed of two distinct stages: the initial stage (S1-1) aims to evaluate whether the input image corresponds to either a lung or colon sample. Based on the outcome of this stage, the second stage (S1-2) is designed to determine the specific class associated with the identified organ. In the second scenario (S2), the target is to identify whether the input data belongs to either the benign or malignant class, irrespective of its origin. The third scenario (S3) involves categorizing the input data among all the type of the tissues (5 classes) of LC25000 dataset. Fig. 1 illustrates an overview of the three scenarios presented. Fig. 2 presents the workflow of the presented strategies of analysis.

A. Dataset

This investigation uses the LC25000 dataset [21], which includes 25000 histopathological images whose size is 768x768 pixels in JPEG format. This dataset is an invaluable resource for training and assessing ML models in cancer diagnosis, particularly for LC cancer. The dataset is distributed across five classes: lung squamous cell carcinoma (lung_scc), lung adenocarcinoma (lung_aca), benign lung tissue (lung_n), benign colonic tissue(colon_n), and colon adenocarcinoma (colon_aca). It was methodically obtained from an initial pool of 750 HIPAA-compliant and verified data, which included 250 cases of each class. Then, it was subsequently augmented using the Augmentor software where the used techniques are random horizontal/vertical flips and left/right rotations (higher than 25 degrees). Consequently, the number of images was enlarged to 25000 and 5000 histopathological images per class.

B. Data Preparation

This section is reserved to detail the data preparation steps, including data pre-processing and data splitting.

1) Data Pre-processing: This step is about resizing images from the original shape to (64, 64, 3). Then, for each scenario, the images belonging to the appropriate classes were selected.

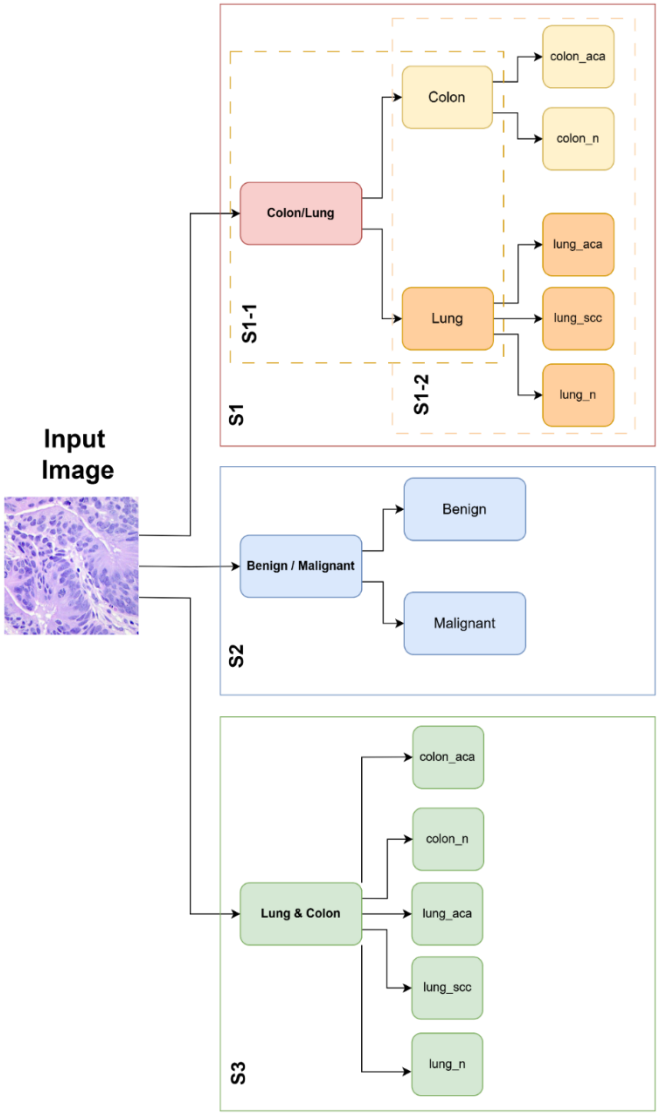


Fig. 1. Synoptic of the three scenarios presented.

TABLE I. STATE-OF-THE-ART AND PROPOSED METHODS DESCRIPTION

Method	Year	Organ	classes	Params(M)	Description	Accuracy
[33]	2021	Lung	3	-	- Image Feature Extraction: HI, TA, Multiscale Analysis - ML Models: KNN, SVM, DT, RF…	0.9933
[40]	2021	LC	5	-	- Digital Image Processing (DIP) - CNN	0.9633
[37]	2021	LC	5	-	- Image Preprocessing: Color Balancing, Image Sharpening, Gamma Correction, and Multi-Scale Fusion. - Multi-Input Capsule Network. - Dual-Input Learning.	0.9958
[50]	2022	Lung Colon LC	3 2 5	-	- Feature Engineering: texture, shape, color histograms… - Image Processing: image normalization, noise removal, and contrast enhancement. -ML Models: XGBoost, SVM, RF, LDA…	0.9953 0.993 0.99
[42]	2022	LC	5	-	- DL models: ShuffleNet, MobileNet, SqueezeNet. - Feature Reduction: PCA, FHWT - Feature Fusion: Discrete Wavelet Transform (DWT) - ML Algorithms: SVM, RF, KNN, LR	0.9960
[23]	2022	Colon	2	-	- Digital Image Processing: Noise reduction, data normalization - Deep CNN	0.9980
[47]	2022	LC	5	-	- Handcrafted Feature Extraction: color, texture, shape, and structure - Deep Feature Extraction: Transfer Learning models - Classifiers: GB, SVM-RBF, MLP, RF	0.9860
[49]	2022	Lung Colon LC	3 2 5	-	-Hybrid Ensemble Feature Extraction strategy: Deep Feature Extraction using DL models and integration of multiple classifiers. - High-Performance Filtering for enhancing image features	0.9905 1.0000 0.9930
[22]	2022	Colon	2	4.6	- Image Preprocessing: Image normalization - DL Model: Convolutional Neural Networks (CNNs)	0.9950
[41]	2022	LC	5	-	- Image Preprocessing: Contrast enhancement. - DL Model: Fine-tuned AlexNet (a pretrained CNN model).	0.984
[34]	2023	Lung	2	-	- Image Preprocessing: Data normalization, resizing, and potentially other enhancements. - Deep Feature Extraction: Convolutional Neural Networks (CNNs) for feature extraction. - ML Models: LightGBM for classification of extracted features.	0.996
[48]	2023	LC	5	9.2	- Image Preprocessing: Image cleaning, resizing, and normalization. - Transfer Learning: EfficientNet and its variants (B0 to B7) for image classification. - Parameter Tuning	0.9724
[45]	2023	LC	5	-	- Deep Features Extraction: GoogLeNet and VGG-19 for highlighting characteristics. - Dimensionality Reduction: PCA - Feature Fusion: Combining features from different models VGG-19 and GoogLeNet, and handcrafted patterns. - Classifier: Artificial Neural Network (ANN)	0.9964
[24]	2023	Colon	2	-	- Ensemble Learning: Stacking DL models to combine predictions from multiple base models. - Pretrained CNNs: InceptionV3, ResNet50, VGG16, DenseNet121 for feature extraction and initial predictions. - Meta-learner: Combined prediction based on SVM and base models. - Explainable AI (XAI)	1.0000
[30]	2024	Lung	3	-	- Deep Learning: CNN - Explainable AI (XAI)	0.999
[35]	2024	Lung	3		- Deep Features extraction: DenseNet201 - Image Preprocessing: Color Histogram Technique. - ML Algorithms: SVM, MultinomialNB, LGBM, CatBoost, XGBoost, KNN, RF	**0.9968**
Proposed Methods	2024	Lung Colon LC LC	3 2 2 5	1.6	- Image Preprocessing: Resizing - DL Models: Lightweight CNN	0.9947 **1.0000** **1.0000** **0.9976**

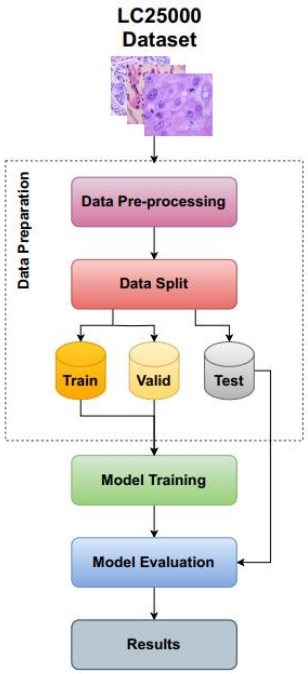


Fig. 2. Workflow of the proposed methods.

2) Data split: DL studies often divide the totality of the data into three sets which are training set, validation set, and testing set. The training set determines model parameters, whereas the validation set adjusts hyperparameters and measures overall performance. The testing dataset measures the model's effectiveness using previously unknown data. The LC25000 dataset was partitioned into training and validation subsets, and also a test set, with an 80:20 split.

The training and validation sets were partitioned using a 90:10 ratio.

C. Model

The proposed model's design belongs to CNN architecture. It consists of two main components: the Features Extractor (FE) and the Classifier. The input data (histological images) will be transformed by FE to advanced characteristics that identify shapes and correlations. This is achieved by the use of three repetitious blocks, including Convolution (Conv2D) and Max Pooling (MaxPool2D) bidimensional layers. The repeated blocks consist of four layers: 2 Conv2D, MaxPool2D, and Batch Normalization (BN). The extracted characteristics will be turned by the classifier to predicted labels of classes. It comprises Dense, Dropout, and Flatten layers, and 4x repeated blocks. The repeated blocks consist of BN and Dense layers. The designed lightweight CNN is depicted in Fig. 3.

D. Hyperparameters

For the training process, this research uses the trial-and-error process to determine the most optimal method. The model's input parameters were set at (64, 64, 3), matching the dimensions of the given image. 32 was the set batch size. The optimal number of epochs (ep) was 50. Adam which has a learning rate of 1.10-3 was the employed optimizer. In addition, the categorical cross-entropy loss was utilized as a loss function. During training, a learning rate schedule was used to manage this phase. It was set as follows:

$$lr = \begin{cases} lr, & ep \leq 10 \\ lr.\exp(-10^{-1}), & ep > 10 \end{cases} \quad (1)$$

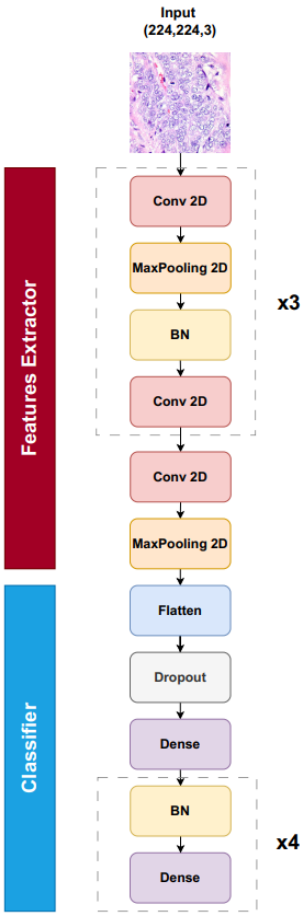


Fig. 3. Architecture of the proposed lightweight CNN.

E. Metrics

The proposed methods were evaluated using different metrics such as AUC, accuracy, F1 score, and. AUC calculates the ability of the system to differentiate between positive and negative data. F1-score incorporates precision and recall into a one statistic. Finally, Accuracy is the proportion of correct predictions to total calculated by the model.

IV. Results

This section is dedicated to show the experiment results. The training phases were done on a personal computer with 16GB of NVIDIA T4 x 2 GPU, 30GB of RAM, and a CPU of 2.20 GHz Intel Xeon. The goal of this research is to provide a versatile framework based on multi-scenario for automated LC cancer analysis. This paragraph aims to present the results for each diagnosis scenario.

1) Diagnosis Scenario (S1): The first scenario (S1) consists of two dependent stages: the first stage (S1-1) determines whether the input image refers to lung or colon data. Based on the results of this stage, the second stage (S1-2) is meant to

determine the exact class corresponding to the indicated organ Table II represents the results of the proposed system during the diagnosis scenarios S1-1, S1-2 Colon, and S1-2 Lung in the training, validation, and test phases. In fact, for scenario S1-1, the AUC F1-score, and accuracy of the established method reached 1.0000 in all phases. This perfect result enables us to pass to the second level of the actual scenario. Regarding the second level, scenario S1-2 Colon, the proposed model attained the same efficiency excluding the loss. Concerning the diagnosis scenario S1-2 Lung, the AUC, F1-score, and accuracy exceeded 0.9947 in the totality of phases.

TABLE II. RESULTS OF THE PROPOSED METHOD DURING THE DIAGNOSIS SCENARIOS S1-1, S1-2 COLON, AND S1-2 LUNG IN THE TRAINING, VALIDATION, AND TEST PHASES

	S1-1			**S1-2 Colon**			**S1-2 Lung**		
Phase	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
Accuracy	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9983	0.9947
F1-Score	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9983	0.9947
AUC	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995

In this study, the classification metrics were also presented for the classes. The proposed approach reached an accuracy, sensitivity, specificity, and F1-score, of 1.000 for both the colon and lung classes. Moreover, in the S1-2 Colon, the previous metrics scored a value of 1.0000. Furthermore, for the S1-2 Lung, with the designed LWCNN, accuracy and F1-score surpassed 0.9947 and 0.9917, while the specificity and sensitivity, each exceeded, 0.9954 and 0.9907, respectively. Table III illustrates the results of the proposed method across all classes during the diagnosis scenarios S1-1, S1-2 Colon, and S1-2 Lung.

These highly performant results can be confirmed by the confusion matrixes. Fig. 4(a), (b), and (c) depict the confusion matrix of the presented scenarios S1-1, S1-2 Lung, and S1-2 Colon, respectively. This indicates that the proposed LWCNN can not only properly identify instances belonging to their respective classes but also successfully separate them from other classes, resulting in no substantial confusion or misclassification. For the S1-2 Lung, the proposed model's confusion matrix is nearly ideal.

TABLE III. RESULTS OF THE PROPOSED METHOD ACROSS ALL CLASSES DURING THE DIAGNOSIS SCENARIOS S1-1, S1-2 COLON, AND S1-2 LUNG

	Class	**Accuracy**	**F1-score**	**Se**	**Sp**
S1-1	Lung	1.0000	1.0000	1.0000	1.0000
	Colon	1.0000	1.0000	1.0000	1.0000
S1-2 Lung	lung_aca	0.9947	0.9923	0.9933	0.9954
	lung_scc	0.9947	0.9917	0.9907	0.9965
	lung_n	1.0000	1.0000	1.0000	1.0000
S1-2 Colon	colon_aca	1.0000	1.0000	1.0000	1.0000
	colon_n	1.0000	1.0000	1.0000	1.0000

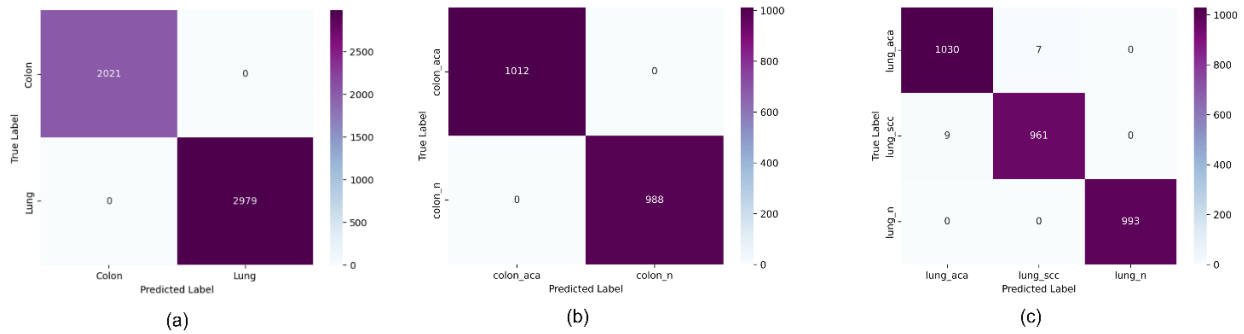


Fig. 4. Confusion matrices of presented LWCNN for diagnosis scenarios: (a) S1-1, (b) S1-2 Colon, and (c) S1-2 Lung.

Similar findings were observed through the ROC (Receiver Operating Characteristic) curve, which illustrates the balance between the true positive rate (TPR) and the false positive rate (FPR) at varying classification levels. As the curve approaches the plot's top-left corner, the model's efficiency improves. Fig. 5(a), (b), and (c) illustrate the ROC curves for the three diagnostic scenarios. In the lung and colon classification (S1-1), the curve attained a point at (0,1), indicating 0% false positives and 100% true positives. The same result was observed for colon classification in S1-2 Colon, confirming an AUC value of 1.0000 for the studied classes. For the S1-2 Lung scenario, the ROC curve was very near to the top-left corner, with AUC values for the lung class above 0.9999.

Overall, the proposed method is efficient for the diagnosis scenario (S1). This was confirmed by a batch of images randomly chosen from the test set for each diagnosis sub-scenario. Fig. 6 depicts the reel test using random test images. All input test images were correctly predicted with a confidence rate higher than 99.99%.

2) Diagnosis Scenario (S2): The second scenario (S2) is designed to find out whether the input data is attributed to the

malignant or benign class, irrespective of its origin (lung or colon). Table IV presents the outcomes for this scenario across the training, validation, and testing phases. The designed LWCNN achieved flawless performance in this scenario, with accuracy, F1-score, and AUC all reaching 1.0000 across every phase.

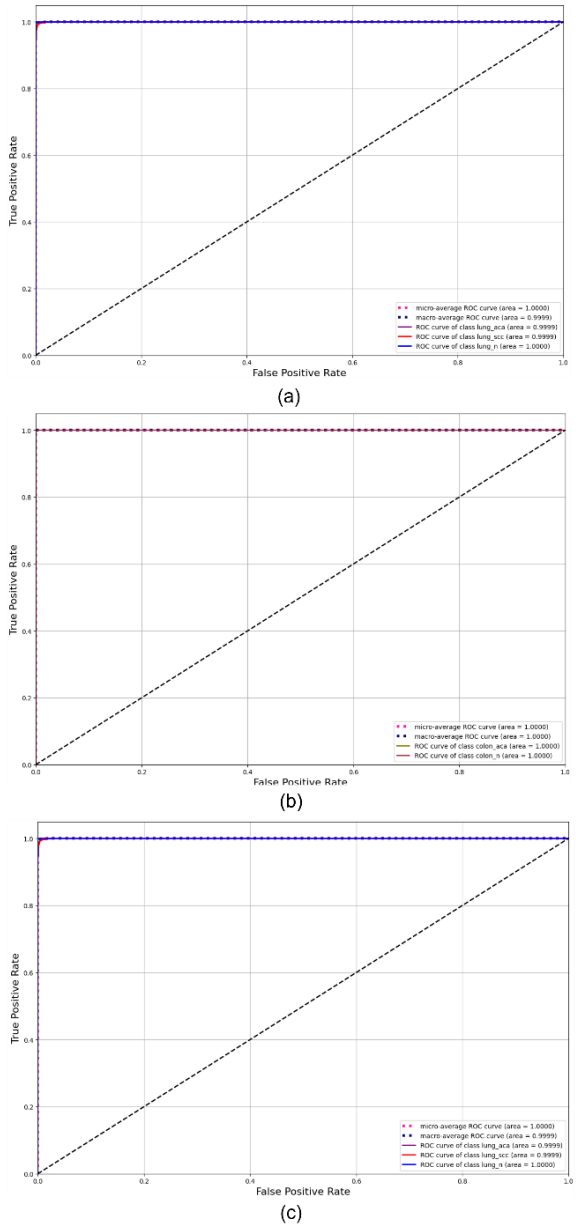


Fig. 5. ROC curves for diagnosis scenarios: (a) S1-1, (b) S1-2 Colon, and (c) S1-2 Lung.

TABLE IV. PROPOSED APPROACH'S FINDINGS DURING THE DIAGNOSIS SCENARIO (S2) IN THE TRAINING, VALIDATION, AND TEST PHASES

Phase	Training	Validation	Test
Accuracy	1.0000	1.0000	1.0000
F1-Score	1.0000	1.0000	1.0000
AUC	1.0000	1.0000	1.0000

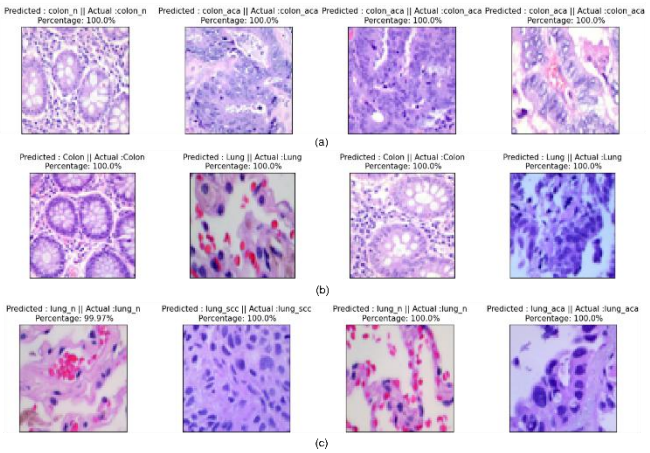


Fig. 6. Framework Test images arbitrarily selected from the test data used for the diagnosis of the sub-scenarios: (a) S1-1, (b) S1-2 Lung, and (c) S1-2 Colon.

This shows that it was capable to consistently and perfectly differentiate between malignant and benign classes with perfect precision, recall, and overall classification metrics.

The classification metrics for each class (benign and malignant) further confirm this outstanding performance. In all phases, the proposed approach attained a sensitivity, specificity, F1-score, and an accuracy of 1.0000 for both classes. These results suggest that the model not only accurately identified instances as either benign or malignant but also demonstrated exceptional sensitivity in detecting positive cases and specificity in ruling out negatives, leading to no false positives or false negatives. The results of the proposed method across all classes during the diagnosis scenario (S2) in Table V.

The confusion matrix, depicted in Fig. 7, provides further validation of the model's perfect performance. Both matrices reveal zero false positives and false negatives, indicating that the LWCNN model accurately classified all instances without any misclassification. This underscores the model's robustness in distinguishing between benign and malignant cases with absolute accuracy.

Moreover, the ROC curves for scenario (S2), illustrated in Fig. 8, also demonstrate the exceptional performance of the LWCNN model. The ROC curve reaches the top-left corner, signifying 0% false positives and 100% true positives for both benign and malignant classifications. This results in an AUC value of 1.0000, confirming that the model performed optimally across all thresholds.

In conclusion, the results of scenario S2 demonstrate the effectiveness of the presented method in discriminating between malignant and benign cases with 100% accuracy. This finding is further corroborated by a set of arbitrarily selected images from the test set, all of which were correctly classified with a confidence rate exceeding 99.99%. Fig. 9 illustrates these perfect predictions, further demonstrating the reliability of the model in real diagnostic applications.

3) Diagnosis Scenario (S3): The third scenario (S3) addresses the categorization of treated images to one of the five available categories from LC25000 dataset. Table VI

summarizes the results of this scenario across all phases. The described LWCNN model demonstrated strong performance, achieving accuracy, F1-score, and AUC values ranging between 0. 9976 and 1.0000, depending on the specific class and phase. While some variations in performance were observed, the model consistently classified the majority of images with high precision and reliability across all phases.

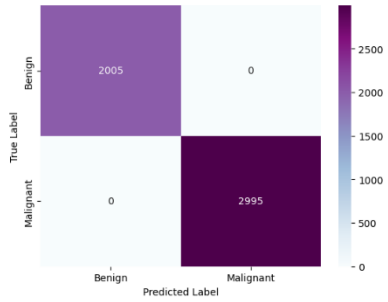


Fig. 7. Confusion matrix of presented LWCNN for diagnosis scenario (S2).

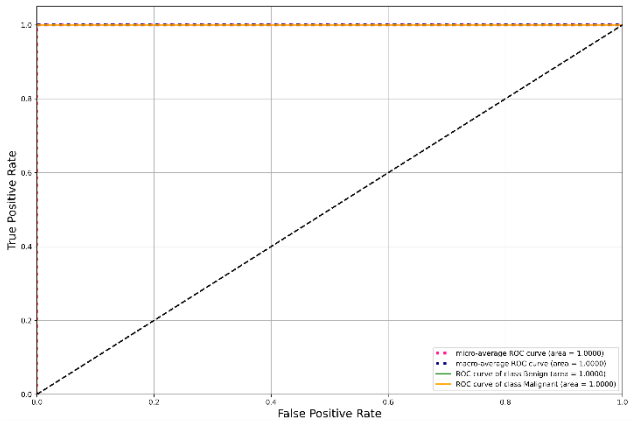


Fig. 8. ROC curves for diagnosis scenario (S2).

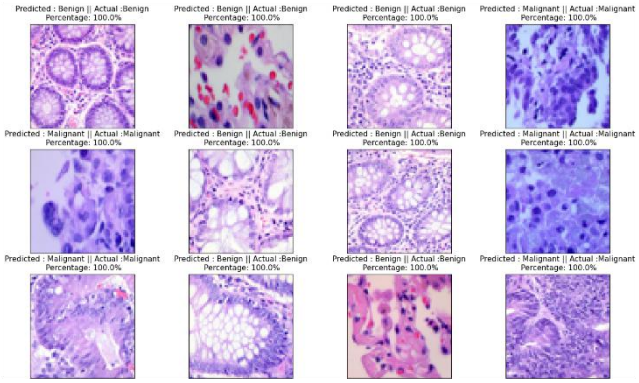


Fig. 9. Framework Test images arbitrarily selected from the test data used for the diagnosis scenario (S2).

TABLE V. RESULTS OF THE PROPOSED METHOD ACROSS ALL CLASSES DURING THE DIAGNOSIS SCENARIO (S2)

Class	Accuracy	F1-score	Se	Sp
Benign	1.0000	1.0000	1.0000	1.0000
Malignant	1.0000	1.0000	1.0000	1.0000

TABLE VI. PROPOSED APPROACH'S FINDINGS DURING THE DIAGNOSIS SCENARIO (S3) IN THE TRAINING, VALIDATION, AND TEST PHASES

Phase	Training	Validation	Test
Accuracy	1.0000	0.9970	0.9976
F1-Score	1.0000	0.9970	0.9976
AUC	1.0000	0.9991	0.9996

TABLE VII. RESULTS OF THE PROPOSED METHOD ACROSS ALL CLASSES DURING THE DIAGNOSIS SCENARIO (S3)

Class	Accuracy	F1-score	Se	Sp
colon_aca	0.9996	0.9990	0.9990	0.9998
colon_n	0.9996	0.9990	0.9990	0.9998
lung_aca	0.9980	0.9949	0.9949	0.9988
lung_scc	0.9980	0.9949	0.9949	0.9987
lung_n	1.0000	1.0000	1.0000	1.0000

A detailed examination of the classification metrics reveals that the model achieved near-perfect results in several classes, with accuracy, F1-score, sensitivity, and specificity reaching values as high as 1.0000 for the lung_n class. In other classes, the performance remained highly competitive, with these metrics exceeding 0.9949. This demonstrates the robustness of the model in handling multiple classification tasks while maintaining substantial accuracy. Table VII demonstrates the results of the proposed method across all classes during the diagnosis scenario (S3).

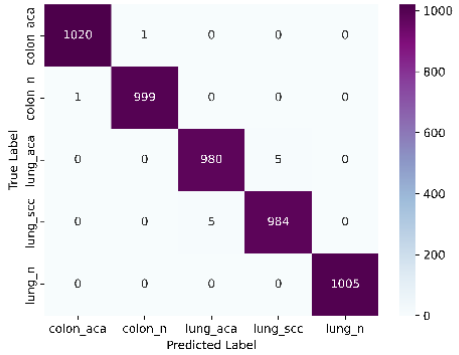


Fig. 10. Confusion matrix of presented LWCNN for the diagnosis scenario (S3).

Fig. 10 illustrates the confusion matrix for the five classes, further validating the model's efficacy in scenario (S3). While some minor misclassifications occurred, indicated by a low value of FP and FN, the overall confusion matrix reveals that the LWCNN model effectively distinguished between the five classes with minimal error, maintaining a high level of classification accuracy.

The ROC curves for scenario (S3), shown in Fig. 11, further support the model's effectiveness. Across all categories, the ROC curve closely approaches the top-left corner. This demonstrates the outstanding capability of the presented approach to distinguish between positive and negative cases. The corresponding AUC values exceed 0.9999, confirming the

model's high sensitivity and specificity across multiple thresholds.

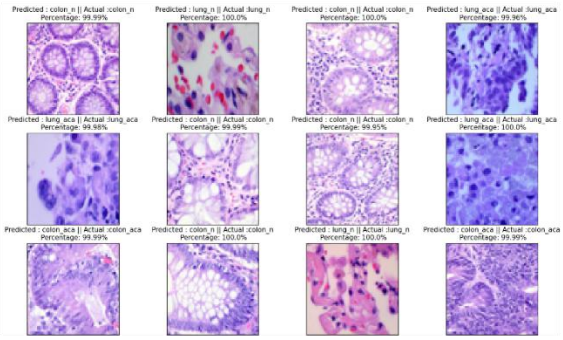


Fig. 11. ROC curves for diagnosis scenario (S3).

In conclusion, the third scenario (S3) demonstrates that the proposed method is highly performant in categorizing input data into one of the five LC25000 dataset classes, with results consistently ranging between 99% and 100%. Despite the minor variations in performance across classes, the model exhibited reliable classification capabilities. Fig. 12 displays randomly selected test images, which were classified with a confidence rate exceeding 99%, further confirming the model's potential in practical diagnostic tasks.

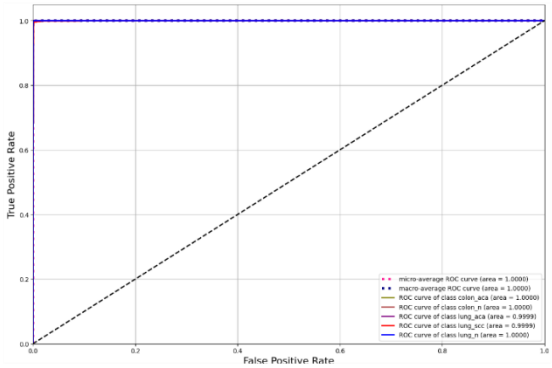


Fig. 12. Framework Test images arbitrarily selected from the test data used for the diagnosis scenario (S3).

V. Discussion

This section is dedicated to debate the experiment results, and compare them to existing approaches. Table VIII illustrates a comparative analysis of the performances of the presented approaches against a selection of well-established ML and DL-based systems for LC cancer classification, as detailed in Section II. Similarly, all the selected literature strategies which employed LC25000 dataset when assessing the outcomes of their models.

TABLE VIII. Comparison of the Overall Findings of the Presented Approaches and the Literature Approaches

Method	Year	Organ	Classes	Parameters (M)	Accuracy	F1-Score	AUC
[33]	2021	Lung	3	-	0.9933	-	-
[40]	2021	LC	5	-	0.9633	0.9638	-
[37]	2021	LC	5	-	0.9958	0.9904	-
[50]	2022	Lung Colon LC	3 2 5	-	0.9953 0.993 0.99	0.9933 0.995 0.988	-
[42]	2022	LC	5	-	0.9960	0.9960	-
[23]	2022	Colon	2	-	0.9980	0.9980	-
[47]	2022	LC	5	-	0.9860	0.9850	-
[49]	2022	Lung Colon LC	3 2 5	-	0.9905 1.0000 0.9930	-	-
[22]	2022	Colon	2	4.6	0.9950	0.9849	-
[41]	2022	LC	5	-	0.984	-	-
[34]	2023	Lung	2	-	0.996	**0.996**	-
[48]	2023	LC	5	9.2	0.9724	-	-
[45]	2023	LC	5	-	0.9964	-	0.998
[24]	2023	Colon	2	-	1.0000	1.0000	-
[30]	2024	Colon	2	-	0.999	0.999	0.998
[35]	2024	Lung	3		**0.9968**	-	-
Proposed Methods	2024	Lung Colon LC LC	3 2 2 5	**1.6**	0.9947 **1.0000** **1.0000** **0.9976**	0.9947 **1.0000** **1.0000** **0.9976**	**0.9995** **1.0000** **1.0000** **0.9996**

This research introduces a flexible framework for automatic analysis of LC cancer. The performance obtained highlights its robustness and efficiency. In fact, the used model was a lightweight CNN, which has a minimalist total number of parameters, 1.6 million parameters, when compared to those of Sakr et al [22] and Anjum et al. [48] where the total parameter is 4.6 and 9.2 million parameters, respectively. The comparison will cover three developed methods, which are colon cancer

classification, lung cancer classification, and LC classification. In colon classification, the proposed method reached the top efficiency with 100% AUC, F1-score, and accuracy. Equally, Gabralla et al. [24] attained the same performance. However, the authors used a more complex model with two levels, which contained individual models and stacking models.

In addition, they used more data preprocessing steps. For example, they applied common data augmentation (DA) like rotating, rescaling, zooming etc. In LC classification, the proposed method reaches an accuracy, F1-score, and AUC of 0.9947, 0.9947, and 0.9995, respectively. This performance is higher than those of Nishio et al. [33], and Talukder et al.'s [49] methods. Although it is very slightly lower than those of Masud et al. [40], Hage Chehad et al. [50], Hamed et al. [34], and Noaman et al.'s [35] approaches. Hamed et al. [34] worked only on two tissue types and not on the totality of the presented lung tissue types. In LC cancer classification, the proposed method outperformed all the cited approaches having the same task. The accuracy scored 0.9976. The F1-score reached 0.9976. The AUC attained 0.9996. These results demonstrate the model's remarkable accuracy, precision, recall, and discriminative power. Thus, the proposed approach is superior in terms of performance, lower in terms of complexity, and more recommended in terms of flexibility and practical usability, highlighting its suitability for this task at hand.

VI. Conclusion

This work aims to build a flexible framework based on multi-scenario diagnosis for LC cancers automated analysis. This was ensured by the employment of a lightweight CNN architecture with a small parameters number against other studies. In fact, the number of parameters was 1.612 million parameters. It was assessed using the LC25000 dataset which comprises five LC tissue types. The presented approach includes three diagnosis scenarios. The first diagnosis scenario (S1) is composed of two distinct stages: the initial stage (S1-1) aims to evaluate whether the input image corresponds to either a lung or colon sample. Based on the outcome of this stage, the second stage (S1-2) is designed to determine the specific class associated with the identified organ. In the second scenario (S2), the aim is to identify whether the input image belongs to either the benign or malignant class, irrespective of its origin. The third scenario (S3) involves categorizing the introduced image into one of the predefined categories in the LC25000 dataset. In the totality of these scenarios, the accuracy, F1-score, and AUC exceeded 0.9947. Regarding these metrics for each class in the presented scenarios, they were higher than 0.9907. The findings of this investigation highlight the important advantages of using the proposed method for the analysis of LC cancer. The model's improved accuracy, dependability, accessibility, flexibility, and capacity for continual development provide major benefits to improving patient care and outcomes. Upcoming work efforts should concentrate on analyzing the performance of the proposed technique over a variety of LC datasets, as well as expanding its assessment to other medical image modalities, such as radiographic or pathological imaging, in addition to histological images. This will assist to demonstrate its resilience, adaptability, and application in a variety of clinical contexts. Furthermore, future research should focus on demonstrating its use in real-world clinical settings to assure its widespread acceptance, generalizability, and potential to enhance diagnostic accuracy and patient outcomes.

References

[1] H. Sung et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," CA. Cancer J. Clin., vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] K. Kurishima et al., "Lung cancer patients with synchronous colon cancer," Mol. Clin. Oncol., vol. 8, no. 1, pp. 137–140, Jan. 2018, doi: 10.3892/mco.2017.1471.

[3] R. C. Callaghan, P. Allebeck, and A. Sidorchuk, "Marijuana use and risk of lung cancer: a 40-year cohort study," Cancer Causes Control, vol. 24, no. 10, pp. 1811–1820, Oct. 2013, doi: 10.1007/s10552-013-0259-0.

[4] M. M. Koo et al., "Presenting symptoms of cancer and stage at diagnosis: evidence from a cross-sectional, population-based study," Lancet Oncol., vol. 21, no. 1, pp. 73–79, Jan. 2020, doi: 10.1016/S1470-2045(19)30595-9.

[5] W. Zhou et al., "Causal relationships between body mass index, smoking and lung cancer: Univariable and multivariable Mendelian randomization," Int. J. Cancer, vol. 148, no. 5, pp. 1077–1086, 2021, doi: 10.1002/ijc.33292.

[6] K. Liu, X. Ning, and S. Liu, "Medical Image Classification Based on Semi-Supervised Generative Adversarial Network and Pseudo-Labelling," Sensors, vol. 22, no. 24, Art. no. 24, Jan. 2022, doi: 10.3390/s22249967.

[7] T. Khan et al., "Autophagy modulators for the treatment of oral and esophageal squamous cell carcinomas," Med. Res. Rev., vol. 40, no. 3, pp. 1002–1060, 2020, doi: 10.1002/med.21646.

[8] D. Daye et al., "Quantitative tumor heterogeneity MRI profiling improves machine learning–based prognostication in patients with metastatic colon cancer," Eur. Radiol., vol. 31, no. 8, pp. 5759–5767, Aug. 2021, doi: 10.1007/s00330-020-07673-0.

[9] M. Sakli, C. Essid, B. B. Salah, and H. Sakli, "Deep Learning Methods for Brain Tumor Segmentation," in Machine Learning and Deep Learning Techniques for Medical Image Recognition, CRC Press, 2023.

[10] E. Bębas et al., "Machine-learning-based classification of the histological subtype of non-small-cell lung cancer using MRI texture analysis," Biomed. Signal Process. Control, vol. 66, p. 102446, Apr. 2021, doi: 10.1016/j.bspc.2021.102446.

[11] M. C. Comes et al., "Early prediction of neoadjuvant chemotherapy response by exploiting a transfer learning approach on breast DCE-MRIs," Sci. Rep., vol. 11, no. 1, p. 14123, Jul. 2021, doi: 10.1038/s41598-021-93592-z.

[12] A. Souid, N. Sakli, and H. Sakli, "Classification and Predictions of Lung Diseases from Chest X-rays Using MobileNet V2," Appl. Sci., vol. 11, no. 6, Art. no. 6, Jan. 2021, doi: 10.3390/app11062751.

[13] M. Sakli, C. Essid, B. Ben Salah, and H. Sakli, "Skin Lesion Segmentation Using U-Net With Different Backbones: Comparative Study," in 2023 IEEE Afro-Mediterranean Conference on Artificial Intelligence (AMCAI), Dec. 2023, pp. 1–4. doi: 10.1109/AMCAI59331.2023.10431489.

[14] M. Sakli, C. Essid, B. B. Salah, and H. Sakli, "Lightweight CNN Towards Skin Lesions Automated Diagnosis In Dermoscopic Images," in 2023 International Conference on Innovations in Intelligent Systems and Applications (INISTA), Sep. 2023, pp. 1–6. doi: 10.1109/INISTA59065.2023.10310480.

[15] M. Sakli, C. Essid, B. Ben Salah, and H. Sakli, "Deep Learning-Based Multi-Stage Analysis for Accurate Skin Cancer Diagnosis using a Lightweight CNN Architecture," in 2023 International Conference on Innovations in Intelligent Systems and Applications (INISTA), Sep. 2023, pp. 1–6. doi: 10.1109/INISTA59065.2023.10310615.

[16] M. Sakl, C. Essid, B. B. Salah, and H. Sakli, "DL Methods for Skin Lesions Automated Diagnosis In Smartphone Images," in 2023 International Wireless Communications and Mobile Computing (IWCMC), Jun. 2023, pp. 1142–1147. doi: 10.1109/IWCMC58020.2023.10183254.

[17] K.-H. Yu et al., "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features," Nat. Commun., vol. 7, no. 1, p. 12474, Aug. 2016, doi: 10.1038/ncomms12474.

[18] W. D. Travis et al., "International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society International Multidisciplinary Classification of Lung Adenocarcinoma," J. Thorac. Oncol., vol. 6, no. 2, pp. 244–285, Feb. 2011, doi: 10.1097/JTO.0b013e318206a221.

[19] AM. Toğaçar, "Disease type detection in lung and colon cancer images using the complement approach of inefficient sets," Comput. Biol. Med., vol. 137, p. 104827, Oct. 2021, doi: 10.1016/j.compbiomed.2021.104827.

[20] M. Sakli, N. Sakli, and H. Sakli, "ECG Images Automated Diagnosis based on Machine Learning Algorithms," in 2023 20th International Multi-Conference on Systems, Signals & Devices (SSD), Feb. 2023, pp. 934–939. doi: 10.1109/SSD58187.2023.10411169.

[21] A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides, "Lung and Colon Cancer Histopathological Image Dataset (LC25000)," Dec. 16, 2019, arXiv: arXiv:1912.12142. doi: 10.48550/arXiv.1912.12142.

[22] A. S. Sakr, N. F. Soliman, M. S. Al-Gaashani, P. Pławiak, A. A. Ateya, and M. Hammad, "An Efficient Deep Learning Approach for Colon Cancer Detection," Appl. Sci., vol. 12, no. 17, Art. no. 17, Jan. 2022, doi: 10.3390/app12178450.

[23] M. I. Hasan, M. S. Ali, M. H. Rahman, and M. K. Islam, "Automated Detection and Characterization of Colon Cancer with Deep Convolutional Neural Networks," J. Healthc. Eng., vol. 2022, no. 1, p. 5269913, 2022, doi: 10.1155/2022/5269913.

[24] L. A. Gabralla et al., "Automated Diagnosis for Colon Cancer Diseases Using Stacking Transformer Models and Explainable Artificial Intelligence," Diagnostics, vol. 13, no. 18, Art. no. 18, Jan. 2023, doi: 10.3390/diagnostics13182939.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.

[27] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 10, 2015, arXiv: arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556.

[28] K. Pogorelov et al., "KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection," in Proceedings of the 8th ACM on Multimedia Systems Conference, in MMSys'17. New York, NY, USA: Association for Computing Machinery, Jun. 2017, pp. 164–169. doi: 10.1145/3083187.3083212.

[29] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," Int. J. Comput. Assist. Radiol. Surg., vol. 9, no. 2, pp. 283–293, Mar. 2014, doi: 10.1007/s11548-013-0926-3.

[30] M. Di Giammarco, F. Martinelli, A. Santone, M. Cesarelli, and F. Mercaldo, "Colon cancer diagnosis by means of explainable deep learning," Sci. Rep., vol. 14, no. 1, p. 15334, Jul. 2024, doi: 10.1038/s41598-024-63659-8.

[31] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.

[32] B. K. H. Thapa Himal Chand, "Lung Cancer Detection Using Convolutional Neural Network on Histopathological Images," Seventh Sense Research Group. Accessed: Oct. 27, 2024. [Online]. Available: https://dev.ijcttjournal.org//archives/ijctt-v68i10p104

[33] M. Nishio, M. Nishio, N. Jimbo, and K. Nakane, "Homology-Based Image Processing for Automatic Classification of Histopathological Images of Lung Tissue," Cancers, vol. 13, no. 6, Art. no. 6, Jan. 2021, doi: 10.3390/cancers13061192.

[34] E. A.-R. Hamed, M. A.-M. Salem, N. L. Badr, and M. F. Tolba, "An Efficient Combination of Convolutional Neural Network and LightGBM Algorithm for Lung Cancer Histopathology Classification," Diagnostics, vol. 13, no. 15, Art. no. 15, Jan. 2023, doi: 10.3390/diagnostics13152469.

[35] N. F. Noaman, B. M. Kanber, A. A. Smadi, L. Jiao, and M. K. Alsmadi, "Advancing Oncology Diagnostics: AI-Enabled Early Detection of Lung Cancer Through Hybrid Histological Image Analysis," IEEE Access, vol. 12, pp. 64396–64415, 2024, doi: 10.1109/ACCESS.2024.3397040.

[36] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A Dataset for Breast Cancer Histopathological Image Classification," IEEE Trans. Biomed. Eng., vol. 63, no. 7, pp. 1455–1462, Jul. 2016, doi: 10.1109/TBME.2015.2496264.

[37] M. Ali and R. Ali, "Multi-Input Dual-Stream Capsule Network for Improved Lung and Colon Cancer Classification," Diagnostics, vol. 11, no. 8, Art. no. 8, Aug. 2021, doi: 10.3390/diagnostics11081485.

[38] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic Routing Between Capsules," Nov. 07, 2017, arXiv: arXiv:1710.09829. doi: 10.48550/arXiv.1710.09829.

[39] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," Sep. 11, 2020, arXiv: arXiv:1905.11946. doi: 10.48550/arXiv.1905.11946.

[40] M. Masud, N. Sikder, A.-A. Nahid, A. K. Bairagi, and M. A. AlZain, "A Machine Learning Approach to Diagnosing Lung and Colon Cancer Using a Deep Learning-Based Classification Framework," Sensors, vol. 21, no. 3, Art. no. 3, Jan. 2021, doi: 10.3390/s21030748.

[41] S. Mehmood et al., "Malignancy Detection in Lung and Colon Histopathology Images Using Transfer Learning With Class Selective Image Processing," IEEE Access, vol. 10, pp. 25657–25668, 2022, doi: 10.1109/ACCESS.2022.3150924.

[42] O. Attallah, M. F. Aslan, and K. Sabanci, "A Framework for Lung and Colon Cancer Diagnosis via Lightweight Deep Learning Models and Transformation Methods," Diagnostics, vol. 12, no. 12, Art. no. 12, Dec. 2022, doi: 10.3390/diagnostics12122926.

[43] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2018, pp. 6848–6856. doi: 10.1109/CVPR.2018.00716.

[44] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," Nov. 04, 2016, arXiv: arXiv:1602.07360. doi: 10.48550/arXiv.1602.07360.

[45] M. Al-Jabbar, M. Alshahrani, E. M. Senan, and I. A. Ahmed, "Histopathological Analysis for Detecting Lung and Colon Cancer Malignancies Using Hybrid Systems with Fused Features," Bioengineering, vol. 10, no. 3, Art. no. 3, Mar. 2023, doi: 10.3390/bioengineering10030383.

[46] C. Szegedy et al., "Going deeper with convolutions," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2015, pp. 1–9. doi: 10.1109/CVPR.2015.7298594.

[47] N. Kumar, M. Sharma, V. P. Singh, C. Madan, and S. Mehandia, "An empirical study of handcrafted and dense feature extraction techniques for lung and colon cancer classification from histopathological images," Biomed. Signal Process. Control, vol. 75, p. 103596, May 2022, doi: 10.1016/j.bspc.2022.103596.

[48] S. Anjum et al., "Lung Cancer Classification in Histopathology Images Using Multiresolution Efficient Nets," Comput. Intell. Neurosci., vol. 2023, no. 1, p. 7282944, 2023, doi: 10.1155/2023/7282944.

[49] Md. A. Talukder, Md. M. Islam, M. A. Uddin, A. Akhter, K. F. Hasan, and M. A. Moni, "Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning," Expert Syst. Appl., vol. 205, p. 117695, Nov. 2022, doi: 10.1016/j.eswa.2022.117695.

[50] A. Hage Chehade, N. Abdallah, J.-M. Marion, M. Oueidat, and P. Chauvet, "Lung and colon cancer classification using medical imaging: a feature engineering approach," Phys. Eng. Sci. Med., vol. 45, no. 3, pp. 729–746, Sep. 2022, doi: 10.1007/s13246-022-01139-x..

Machine Learning-Based Denoising Techniques for Monte Carlo Rendering: A Literature Review

Liew Wen Yen[1], Rajermani Thinakaran[2], J. Somasekar[3]
Faculty of Data Science and Information Technology, INTI International University, Negeri Sembilan, Malaysia[1, 2]
Department of Computer Science and Engineering, Jain (Deemed–to-be University), Bangalore, Karnataka, India[3]

***Abstract*—Monte Carlo (MC) rendering is a powerful technique for achieving photorealistic images by simulating complex light interactions. However, the inherent noise introduced by MC rendering necessitates effective denoising techniques to enhance image quality. This paper presents a comprehensive review and comparative analysis of various machine learning (ML) methods for denoising MC renderings, focusing on four main categories: radiance prediction using convolutional neural networks (CNNs), kernel prediction networks, temporal rendering with recurrent architectures, and adaptive sampling approaches. Through systematic analysis of 7 peer-reviewed studies from 2019-2024, the author's findings reveal that deep learning models, particularly generative adversarial networks (GANs), achieve superior denoising performance. The study identifies key challenges including computational demands, with some methods requiring significant GPU resources, and generalization across diverse scenes. Additionally, we observe a trade-off between denoising quality and processing speed, particularly crucial for real-time applications. The study concludes with recommendations for future research, emphasizing the need for hybrid approaches combining physics-based models with ML techniques to improve robustness and efficiency in production environments.**

Keywords*—Convolutional neural network; Monte Carlo rendering; generative adversarial network; deep learning; machine learning; denoising techniques

I. Introduction

Monte Carlo (MC) rendering has emerged as a fundamental technique in computer graphics, enabling the simulation of light behavior in virtual environments through probabilistic sampling methods. By tracing numerous light paths and statistically sampling their contributions, MC rendering effectively captures complex light interactions with surfaces, materials, and volumes, resulting in highly realistic images characterized by accurate lighting, shadows, reflections, and refractions [1] [2]. This capability has rendered MC rendering indispensable across various industries, including film production, architectural visualization, and video game development, where photorealistic visuals are paramount [3] [4]. In film production, noise can disrupt the photorealism required for high-quality visual effects, while in video games, it can hinder real-time performance and user experience.

However, a notable challenge associated with MC rendering is the presence of noise in the generated images. Noise, which manifests as random variations or artifacts, arises from the inherent probabilistic nature of light path sampling. This issue is particularly pronounced in scenes with intricate lighting, glossy surfaces, or complex geometries, leading to grainy or speckled appearances that detract from the visual quality and realism of the rendered outputs [5] [6]. The reliance on probabilistic sampling methods contributes to this noise, as low sample counts can result in high variance in light estimates. While increasing the sample count can mitigate noise, it significantly escalates computational demands, rendering such approaches impractical for real-time or interactive applications [7].

To combat the noise prevalent in MC renderings, denoising techniques have become essential for enhancing image quality. These algorithms are designed to intelligently filter out noise while preserving critical image details, textures, and features, thereby yielding smoother and cleaner final renderings [7][8]. Recent advancements in machine learning (ML), particularly through deep learning models such as convolutional neural networks (CNNs) and generative adversarial networks (GANs), have opened new avenues for effective denoising. These ML techniques can learn complex noise patterns from extensive datasets of noisy and clean images, enabling them to generalize across various scenes and lighting conditions while maintaining the fidelity of essential image details during the denoising process [3] [9].

Despite the extensive exploration of ML-based denoising methods, there remains a scarcity of work synthesizing and comparing these approaches across diverse rendering scenarios. Challenges such as high computational demands, generalization across varying scenes, and interpretability of the models persist as unresolved issues. High computational demands limit the applicability of ML-based denoising in real-time applications, while generalization issues arise due to the variability of noise patterns across different scenes and lighting conditions. This study endeavors to fill these gaps by providing a systematic review and comparative analysis of existing methods, offering practical recommendations for future research in the field of MC rendering and denoising [5] [9].

To address these challenges, this study seeks to answer the following key research questions:

1) What ML methods have been employed for denoising MC renderings, according to the existing literature?

2) How do these ML methods compare in terms of performance, efficiency, and application?

3) What are the current challenges and future directions for ML-based denoising in MC rendering?

The primary objective of this study is to provide a comprehensive review and comparative analysis of these

techniques, categorizing them into radiance prediction, kernel prediction, temporal rendering, and adaptive sampling. Performance evaluations will utilize metrics such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Relative Mean Squared Error (rMSE) [4] [8]. Additionally, this study will highlight gaps in the current literature and propose future research directions to advance ML-based denoising techniques, ultimately enhancing rendering workflows and visual outputs.

II. METHOD

A systematic literature review (SLR) methodology was employed to ensure a comprehensive and unbiased review of ML techniques for denoising in MC rendering. An SLR involves analyzing existing research by defining clear research questions, identifying relevant studies, appraising their quality, and synthesizing findings both qualitatively and quantitatively [10]. This structured approach ensures transparency, replicability, and rigor in the review process. The methodology adheres to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework to enhance methodological rigor and quality.

The methodology of this study is organized into three key stages: planning the review, which involves defining research questions, developing search strategies, and establishing inclusion/exclusion criteria; conducting the review, which includes searching and screening relevant studies, extracting data, and assessing the quality of selected studies; and analyzing the gathered information, which consists of synthesizing results, discussing trends, and identifying challenges and opportunities for future research.

A. Planning the Review

1) Scope of the review: The SLR focuses specifically on machine learning ML applied to denoising in MC rendering. MC rendering is widely used in computer graphics to simulate realistic lighting effects, but its probabilistic nature often introduces noise into rendered images. This noise can degrade visual quality, making effective denoising techniques essential for achieving high-quality outputs. Despite recent advancements in ML-based denoising methods, there remains a scarcity of work synthesizing and comparing these approaches across diverse rendering scenarios. Challenges such as high computational demands, generalization across varying scenes, and interpretability of the models persist as unresolved issues. This study aims to address these gaps by providing a comprehensive review and comparative analysis of ML-based denoising techniques, categorizing them into radiance prediction, kernel prediction, temporal rendering, and adaptive sampling.

2) Research questions: The formulation of research questions was guided by an iterative process involving pilot searches and consultations with domain experts. Initial exploratory searches were conducted across academic databases such as IEEE Xplore, ACM Transactions on Graphics, and ScienceDirect using broad keywords like “Monte Carlo rendering,” “denoising,” and “machine learning.” These searches helped identify recurring themes and trends in literature, such as the use of CNNs, GANs, and kernel prediction methods. Informal consultations with an expert in computer graphics and ML-based rendering techniques provided valuable feedback on the scope and relevance of the questions, suggesting additional considerations such as computational efficiency and generalization across diverse scenes. The final research questions guiding this SLR are as follows:

1) What ML methods have been employed for denoising MC renderings, according to the existing literature?

2) How do these ML methods compare in terms of performance, efficiency, and application?

3) What are the current challenges of ML for denoising MC rendering?

4) Search strategy: A comprehensive search strategy was developed to gather relevant literature across multiple academic databases. The search was conducted using keywords such as “Monte Carlo rendering,” “denoising,” “convolutional neural network,” “deep learning,” “machine learning denoising techniques,” and “generative adversarial network”. The databases searched included IEEE Xplore, Google Scholar, ScienceDirect, Computer Graphics Forum, and ACM Transactions on Graphics. To ensure consistency in the analysis, the search was limited to studies published in English between 2019 and 2024.

5) Inclusion and exclusion criteria: Specific criteria were established to filter the studies for inclusion in this review, ensuring the relevance and quality of the selected literature. This study employed a filtration process guided by predefined inclusion and exclusion criteria as shown in Tables I and II. All the papers were assessed against a set of inclusion and exclusion criteria to ensure they directly addressed the research questions.

TABLE I. INCLUSION CRITERIA

ID	Inclusion criterion
I1	Studies explicitly focus on machine learning techniques for denoising in Monte Carlo rendering.
I2	Studies published in peer-reviewed journals and conferences to ensure credibility and quality.
I3	Studies that provide quantitative performance metrics such as PSNR, SSIM, rMSE, or computational efficiency.
I4	Studies published in English.
I5	Studies published between 2019 and 2024.

TABLE II. EXCLUSION CRITERIA

ID	Exclusion criterion
E1	Studies do not focus on denoising in the context of Monte Carlo rendering.
E2	Studies lacking sufficient empirical data or clear evaluation methods, which could undermine the validity of the findings.
E3	Non-peer-reviewed articles, editorials, or opinion pieces, as these sources do not provide the rigorous analysis required for this review.
E4	Studies published in languages other than English.
E5	Studies published before 2019.

6) Selection process: The selection process involved two phases: initial screening and full-text review. In the initial screening, titles and abstracts of all retrieved papers were reviewed against the inclusion and exclusion criteria. Papers that did not meet these criteria were excluded. In the full-text review, the remaining papers were examined in detail to assess their relevance to the research questions and the quality of their methodologies and findings. A total of seven papers were selected for the final analysis. To provide a clear overview of the study selection process, a PRISMA flowchart (Fig. 1) was created, following the guidelines outlined by Pati and Lorusso [11]. The flowchart visually summarizes the number of studies identified, screened, and included at each stage of the review process.

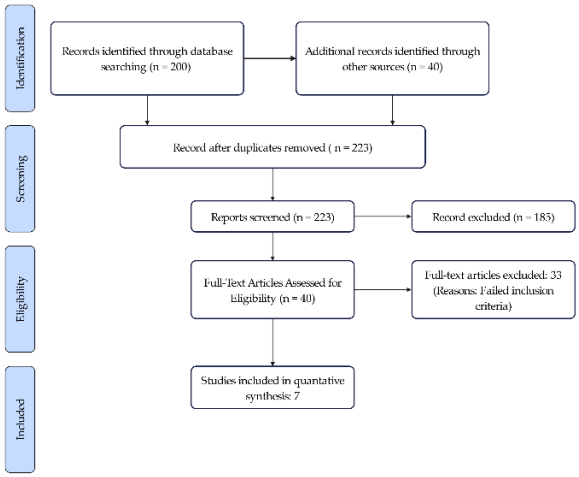


Fig. 1. PRISMA flowchart summarizing the study selection process.

B. Conducting the Review

This phase involved three key steps: data extraction, quality assessment, and data synthesis. Each step was meticulously performed to ensure a robust and unbiased evaluation of the selected studies, enabling a comprehensive comparison of ML-based denoising techniques for MC rendering.

1) Data extraction: Data extraction was systematically performed on the selected studies to comprehensively address the research questions. Key information extracted included the specific ML models employed for denoising, the datasets used, and performance metrics such as PSNR and SSIM, rMSE. This structured extraction approach enabled a robust comparison of the effectiveness and efficiency of different ML-based denoising techniques across a range of rendering scenarios.

2) Quality assessment: To ensure the reliability and validity of the findings, each study included in this review underwent a rigorous quality assessment. This process evaluated the methodological rigor of the studies, focusing on factors such as the robustness of the experimental design, the clarity of data presentation, and the appropriateness of the performance metrics used. Only studies that met these stringent criteria were included, ensuring that this review comprises high-quality research offering credible insights into the effectiveness and efficiency of ML-based denoising methods.

3) Data synthesis: The extracted data were synthesized to provide a comprehensive comparison of the different ML-based denoising techniques for MC rendering. Both qualitative and quantitative analyses were conducted to identify trends, strengths, and limitations across the existing literature. Performance metrics were aggregated where applicable, allowing for a standardized comparison of the denoising effectiveness across different studies. This synthesis provides a holistic view of the current landscape of ML-based denoising methods in MC rendering, highlighting their practical applications and potential areas for future research.

C. Analyzing the Gathered Information

1) Synthesis of results: The extracted data were synthesized to provide a comprehensive comparison of the different ML-based denoising techniques for MC rendering. Both qualitative and quantitative analyses were conducted to identify trends, strengths, and limitations across the existing literature. Performance metrics were aggregated where applicable, allowing for a standardized comparison of the denoising effectiveness across different studies. This synthesis provides a holistic view of the current landscape of ML-based denoising methods in MC rendering, highlighting their practical applications and potential areas for future research.

2) Discussion of results: The analysis identified several challenges associated with applying ML techniques to MC rendering denoising. One major challenge is the computational complexity of these methods, as many ML-based denoisers require substantial processing power and memory to achieve high-quality results. This computational demand makes it difficult to deploy these techniques in real-time or interactive applications where performance and speed are critical. Additionally, the complexity of the models can hinder their ability to generalize across diverse scenes, as training datasets may not fully capture the variability in noise patterns that arise in different rendering scenarios.

Another challenge is the difficulty in balancing denoising performance with computational efficiency. While deep learning models such as CNNs and GANs have shown promise in reducing noise while preserving image details, these methods often come with a trade-off between the quality of the denoised output and the computational resources required. Addressing these challenges involves exploring more efficient architectures, optimization techniques, and potentially new approaches to model training that can reduce computational overhead without compromising denoising quality.

3) Recommendations for future research: Based on the findings, several recommendations for future research have been proposed. These include integrating physics-based models, adopting adaptive sampling strategies, employing advanced network architectures such as GANs and CNNs, utilizing detail-preserving neural networks, and implementing path-based denoising techniques. These suggestions aim to steer future

research efforts toward overcoming current limitations and exploring new opportunities in ML-based denoising for MC rendering.

III. Results and Discussion

This study explores three key questions. The following sections analyze the findings and their significance to each question.

A. What ML Methods have been Employed for Denoising MC Rendering, according to the Existing Literature?

MC rendering is well-known for simulating realistic lighting effects, but it has its challenges, particularly with noise in the images because of the stochastic nature of the sampling process. Over the years with the advancement of ML, it has been employed to effectively denoise MC renderings. These methods leverage neural networks which enables them to clean up images more effectively than traditional techniques. Below, we explore some of the research papers that use ML methods for denoising MC rendering, categorized into kernel prediction, parameter prediction, radiance prediction, and temporal denoising. This categorization framework, derived from the work of Huo et al. [12], provides a structured approach to understanding the strengths and limitations of each method.

1) Kernel prediction: Kernel prediction focuses on directly predicting the filtering kernels used to combine neighboring pixel values. This enhances the denoising process by adapting the kernels to the specific noise characteristics of each pixel. This approach is a more flexible and accurate solution compared to traditional filtering techniques, particularly in handling complex scenes and varying lighting conditions.

Back et al. [13] introduce a deep learning-based framework designed to improve the accuracy of MC rendering by effectively combining independent and correlated pixel estimates. Their approach utilizes a combination kernel modeled as a deep neural network, which optimally weights the combination of these pixel estimates, thereby reducing residual noise and systematic errors commonly found in existing methods like denoising and gradient-domain rendering. The framework is robust against outliers, thanks to an extension that employs multi-buffered inputs, which further enhances the reliability of the results. Experimental evaluations demonstrate that this method not only enhances the visual quality of renders by preserving high-frequency details and reducing noise but also outperforms existing techniques in terms of both numerical accuracy and visual fidelity. This makes the approach particularly valuable for applications requiring high-quality rendering, such as production-level visual effects and interactive applications.

Gharbi et al. [14] introduce a sample-based MC denoising technique using a kernel-splatting network. Unlike traditional pixel-based methods, their approach operates directly on the raw MC samples, leveraging deep learning to map these samples to a denoised image. The core innovation lies in a novel kernel-predicting architecture that splats individual samples onto nearby pixels. This method treats each sample independently and uses a permutation-invariant design to handle the arbitrary order of samples. The kernel-splatting approach is particularly effective in managing complex light transport scenarios such as motion blur, depth of field, and specular effects. By directly processing the sample-level information, the technique achieves higher quality results with reduced numerical error and improved visual fidelity, especially in low-sample-count settings. The network was trained on a large dataset of synthetic scenes and demonstrated significant improvements over state-of-the-art methods in both visual quality and computational efficiency [14].

Munkberg and Hasselgren et al. [15] propose a novel approach to neural denoising in MC path tracing by introducing a layered architecture that partitions per-sample data into distinct layers. Each layer is processed with unique filter kernels before being composited to produce the final output. This approach balances computational efficiency with high-quality denoising, offering comparable results to more expensive per-sample methods while significantly reducing memory and performance overhead. The architecture is particularly robust against high-intensity outliers and performs well even in complex visibility scenarios, such as defocus and motion blur. The authors demonstrate that their method achieves near real-time performance on contemporary GPUs, making it viable for both real-time rendering and offline production environments. Future work is suggested in extending this layered approach to temporal domains and deep compositing workflows, indicating its potential for broader applications in rendering technologies.

2) Parameter prediction: Parameter prediction involves training neural networks to predict the optimal parameters for traditional filters to enhance their ability to reduce noise while preserving image details.

Xing and Chen [16] introduce an approach to denoising path-traced images by combining SURE-based adaptive sampling with neural networks. Their process begins with generating coarse samples and using Stein's Unbiased Risk Estimator (SURE) to estimate the noise level for each pixel. Extra samples are then allocated to pixels with higher noise levels. In the reconstruction phase, a MLP network predicts the optimal reconstruction parameters based on features extracted from the adaptive sampling results, such as shading normal, depth, and texture values. These predicted parameters are used with an anisotropic filter to produce the final noise-free image. This method reduces numerical error as well as enhances visual quality compared to existing techniques.

3) Radiance prediction: Radiance prediction focuses on directly estimating the radiance values for each pixel in a MC rendering. These methods bypass the need for traditional filtering or kernel prediction. They utilize deep learning models to map noisy input pixels directly to their denoised counterparts, effectively capturing complex relationships between the noisy input and the desired output. By predicting radiance directly, these approaches can handle high-frequency details and complex lighting scenarios more effectively, allowing them to be more suitable for applications where visual accuracy is needed.

Xu et al. [17] introduce an adversarial approach for denoising MC renderings, leveraging GANs to improve the

realism of high-frequency details and global illumination. Their method employs a conditioned auxiliary feature modulation technique that utilizes auxiliary features such as normal, albedo, and depth to enhance the denoising process. The GAN framework consists of a denoising network, which predicts the clean image, and a critic network, which evaluates the perceptual quality of the denoised output. The critic network is trained using the Wasserstein distance, which provides a smoother measure of perceptual similarity compared to traditional losses. This approach enables the denoising network to learn from the distribution of high-quality path-traced images, resulting in better reconstruction of MC integrals from fewer samples. Xu et al. demonstrate that their method outperforms previous state-of-the-art techniques in terms of both visual quality and computational efficiency, making it suitable for high-end production environments.

The paper by Alsaiari et al. [18] presents a novel approach for image denoising using GAN architecture. The method involves rendering images with a reduced number of samples per pixel, which results in noisy outputs, and then passing these images through a GAN-based network that produces high-quality, photorealistic denoised images in less than a second. The proposed network architecture leverages residual blocks, skip connections, and batch normalization to enhance the denoising process. Despite being trained on a limited dataset of 40 images, the network demonstrated impressive generalization capabilities, effectively denoising images outside the training domain, including grainy photographs and medical CT scans. The authors also discuss potential future extensions of their work, including handling more complex noise patterns such as those generated by MC rendering and incorporating additional information like depth maps to improve denoising performance in scenes with motion blur, depth of field, and global illumination. The study underscores the effectiveness of GANs in producing high-quality denoised images and suggests further exploration of this approach in real-time rendering applications.

4) Temporal rendering: Temporal rendering is specifically designed to address the challenges of ensuring frame-to-frame consistency in animated or real-time rendering sequences. Noise reduction in MC rendering needs to be effective not just on individual frames, but also across time, to prevent flickering or temporal artifacts that can detract from the visual experience. These methods often utilize recurrent structures to ensure that noise is reduced consistently across frames. It preserves temporal coherence while maintaining high-quality image details.

Meng et al. [19] introduce a practical and efficient approach to real-time MC denoising by leveraging a neural bilateral grid. Their method utilizes a convolutional neural network, called GuideNet, to predict guide images that direct the placement of noisy radiance data into a multi-scale bilateral grid. The grid is then sliced to extract denoised data, resulting in high-quality renders even from extremely noisy inputs at 1 spp. The proposed approach is highly scalable and adaptable to both real-time and offline applications, demonstrating superior denoising quality compared to existing methods, particularly for low-sample scenarios. The study emphasizes the method's ability to maintain interactive frame rates while achieving high visual fidelity, making it a robust solution for real-time rendering in demanding environments.

B. How do these ML Methods Compare in Terms of Performance, Efficiency, and Application?

These denoising methods analyzed in this study exhibit varying degrees of performance, efficiency, and application suitability in MC-rendered images. By examining key metrics such as rMSE, SSIM, PSNR, and processing time, we can assess how each method balances noise reduction, computational efficiency, and applicability to different rendering scenarios.

Methods such as Xu et al. [17] excel with an rMSE of 0.003164 and a PSNR of 34.194759 dB in the HorseRoom scene, outperforming traditional methods like NFOR in retaining fine details. These methods are particularly suited for scenarios where achieving the highest possible image quality is crucial, even if it comes at the cost of longer processing times.

In environments where real-time performance is essential, such as video games, virtual reality, and interactive simulations, the kernel-splatting network by Gharbi et al. [14] and the GAN-based approach by Alsaiari et al. [18] are particularly effective. Gharbi et al.'s method [14] achieves an rMSE of 0.026 at 32 spp while processing a 1024 × 1024 image in just 6.0 seconds at 4 spp, making it ideal for real-time applications that require a balance between speed and quality. Similarly, Alsaiari et al.'s method [18] generates high-quality denoised images in under a second, emphasizing rapid processing without significantly compromising visual fidelity, making it highly suitable for scenarios where quick turnaround times are critical.

Methods by Jonghee Back et al. [13] and Munkberg and Hasselgren [15] offer strong capabilities in handling complex lighting environments and preserving intricate details. Jonghee Back et al.'s deep combiner for independent and correlated pixel estimates achieves a significant reduction in rMSE, such as 0.0207 in the Bookshelf scene at 64 spp, making it effective in handling scenes with intricate lighting and textures. Munkberg and Hasselgren's neural denoising method with layer embedding also shows strong performance, achieving an rMSE of 0.0288 and SSIM of 0.941 at 32 spp, making it highly effective for maintaining image quality in complex visual effects.

Xing and Chen's method [16] leverages adaptive sampling based on SURE combined with a modified MLP network to predict optimal reconstruction parameters. The method demonstrates significant noise reduction with a RelMSE of 0.00831 in the Sibenik scene at 16.9 spp, and 2.37E-4 in the Anim-BlueSphere scene at 30.6 spp. It optimizes sample distribution across pixels with varying noise levels, enhancing computational efficiency while maintaining high image quality. This method is particularly effective for real-time applications and interactive graphics, where maintaining quality with lower sample counts is crucial.

C. What are the Current Challenges in the Application of ML for Denoising MC Rendering?

One of the main challenges associated with using ML for denoising in MC rendering is its inherent complexity. MC rendering simulates how light behaves within a scene by tracing numerous random paths. Hence, this results in inherently noisy

images and requires a denoising process for better visual quality [20]. The primary difficulty lies in the nature of the noise, which is stochastic and can vary significantly across different scenes. Therefore, denoising algorithms must be adept at distinguishing between noise and the true signal to avoid blurring or distorting the final image [20]. In addition, the denoising task is complicated both by the high dimensionality of the data and the complex interplay of light transport phenomena. In this regard, a sophisticated ML model is needed to be able to capture these intricate relationships with complex data [21, 22].

Additionally, the efficiency and computational cost of denoising algorithms are also challenges in the MC rendering process. Deep learning techniques have shown promise in attaining higher quality denoising, but they often come with a computational cost and require substantial processing power and time for training and inference [23]. This computational overhead can make deploying ML-based denoising solutions in real-time or interactive rendering scenarios challenging, where performance is crucial [24]. Real-time rendering applications, such as video games or virtual reality, demand quick and efficient processing to maintain smooth and responsive user experiences. Therefore, balancing denoising quality and computational efficiency in rendering and MC is today's critical challenge using ML approaches for denoising [23]. Developing methods that optimize this balance is essential to ensure that high-quality denoising can be achieved without compromising the performance required for real-time applications. This involves exploring more efficient algorithms, hardware acceleration, and innovative training techniques to reduce computational demands while maintaining or improving denoising effectiveness.

The other critical challenge is the generalization of denoising algorithms across different scenes and lighting conditions. Noise patterns and characteristics of MC renderings can vary greatly depending on scene complexity, materials present, and lighting setup [25]. For instance, a scene with complex geometry and reflective surfaces might produce noise patterns that are vastly different from a simple scene with diffuse materials. This variability necessitates that ML models are not only trained on diverse datasets but are also rigorously tested to ensure their effectiveness in new, unseen scenarios. Thus, it is essential to practically assess that ML models effectively generalize over unseen data and across diverse rendering scenarios in rendering pipelines [25]. Robustness to scene variations and the ability to adapt to different noise profiles are crucial aspects that need to be addressed to make denoising algorithms effective across a wide range of rendering scenarios [25][26][27]. Addressing these issues involves developing more sophisticated training regimes, incorporating a wider range of scenes and conditions, and continuously updating models to handle new types of noise as they are encountered.

Besides, ML-based denoising methods further raise issues concerning interpretability and transparency for MC rendering. Deep learning models are often thought to be black boxes, making it difficult to understand the decision process for denoising and what features were prioritized in the process [21]. This can be problematic for artists and developers who rely on precise control over rendering parameters to achieve specific visual effects [21]. For instance, they may need to adjust the denoising parameters to maintain certain artistic details or to ensure the consistency of visual styles across different scenes. Without a clear understanding of how the ML model operates, making these adjustments becomes exceedingly difficult. This lack of interpretability can also hinder debugging and improvement efforts, as it is unclear why the model might fail in certain scenarios. Thus, improving the interpretability of these models while preserving their denoising performance is a challenge that needs attention in applying ML for denoising MC rendering [21].

Table III provides a summary of the performance, efficiency, and application of a selection of seven methods. This table serves as a quick reference for understanding the strengths and limitations of each approach, making it easier for researchers and practitioners to select the most appropriate denoising method based on their specific needs. By comparing metrics such as PSNR, SSIM, and computational demands, the table illustrates the diverse range of strategies employed across different methods to balance speed and quality.

IV. Conclusion

This study set out to explore and analyze ML-based denoising techniques for MC rendering, focusing on three key research questions. Using a SLR approach guided by the PRISMA framework, the authors identified, categorized, and compared seven peer-reviewed studies published between 2019 and 2024. These methods were grouped into four main categories—radiance prediction, kernel prediction, temporal rendering, and adaptive sampling—and evaluated using metrics like PSNR, SSIM, and rMSE. Our findings show that deep learning models, such as CNNs and GANs, are highly effective at reducing noise while preserving important details in MC-rendered images. However, challenges like high computational demands and difficulties in generalizing across different scenes still limit their use in real-time applications.

The main contribution of this work is bringing together a fragmented field into a clear and structured framework. This makes it easier for practitioners to choose the right denoising technique based on their specific needs. For example, kernel-splatting networks work well for real-time scenarios, while adversarial methods excel in producing high-quality results for offline production. We also highlighted some critical gaps in the literature, such as the need for more interpretable models and efficient architectures that strike a better balance between quality and computational cost.

That said, this review isn't without its limitations. By focusing only on studies from 2019 to 2024, we might have missed some foundational work from earlier years. Additionally, our reliance on databases like IEEE Xplore, ACM, and ScienceDirect could introduce bias, and the inclusion of just 12 papers may not fully capture the diversity of approaches out there. While qualitative comparisons provide valuable insights, they lack the statistical depth of a meta-analysis, which could offer a more quantitative assessment of these methods.

Looking ahead, there are several exciting directions for future research. One promising area is hybrid approaches that combine physics-based models with ML techniques to improve robustness and accuracy. Another is exploring temporal

optimization strategies to reduce flickering artifacts in animations. By addressing these challenges, researchers can develop deployable solutions that balance photorealism with computational efficiency, ultimately transforming workflows in industries like film, architecture, and gaming.

In short, this study provides a comprehensive overview of the current state of ML-based denoising for MC rendering, identifies key challenges, and suggests practical ways forward. The goal is to inspire further innovation in creating denoising solutions that are not only powerful but also practical for real-world applications.

TABLE III. SUMMARY OF THE PERFORMANCE, EFFICIENCY, AND APPLICATION OF THE METHODS

Method	rMSE (Range)	SSIM (Range)	PSNR (dB)	Efficiency	Application
Adversarial Monte Carlo Denoising by Xu et al. [17]	0.003164	N/A	34.194759	High computational cost, suited for offline rendering	High-end production environments, detailed visual effects
Kernel-Splatting Network by Gharbi et al. [14]	0.026	N/A	N/A	Optimized for fewer samples, balance between speed and quality	Real-time applications, interactive graphics, gaming
GAN-based Denoising by Alsaiari et al.[18]	N/A	0.938 - 0.941	33.706 - 33.878	Real-time performance, quick processing in under a second	Real-time rendering, architectural visualization
Neural Bilateral Grid by Meng et al. [19]	N/A	0.941	33.838	High real-time performance at 61 FPS, optimized for low sample scenarios	Real-time rendering, gaming, and VR
Deep Combiner for Independent and Correlated Pixel Estimates by Jonghee Back et al. [13]	N/A	N/A	N/A	Effective in handling complex scenes, reduces relMSE	Production-level visual effects, interactive applications requiring high-quality rendering
Neural Denoising with Layer Embedding by Munkberg and Hasselgren [15]	0.0288	0.941	N/A	Robust against artifacts, effective with varying configurations	Offline rendering, flexible in handling per-sample, per-pixel, and layered configurations
Path Tracing Denoising based on SURE Adaptive Sampling by Xing and Chen [16]	0.00831	N/A	N/A	Adaptive sampling with SURE; highly efficient in real-time scenarios with CUDA acceleration	Interactive graphics, real-time applications, scenarios with limited computational resources

REFERENCES

[1] K. Wong and T. Wong, "Deep residual learning for denoising monte carlo renderings", Computational Visual Media, vol. 5, no. 3, p. 239-255, 2019.

[2] J. Back, B. Hua, T. Hachisuka, & B. Moon, "Deep combiner for independent and correlated pixel estimates", Acm Transactions on Graphics, vol. 39, no. 6, p. 1-12, 2020.

[3] Q. Hou, "Auxiliary features-guided super resolution for monte carlo rendering", Computer Graphics Forum, vol. 43, no. 1, 2023.

[4] X. Zhang, M. Manzi, T. Vogels, H. Dahlberg, M. Groß, & M. Papas, "Deep compositional denoising for high-quality monte carlo rendering", Computer Graphics Forum, vol. 40, no. 4, p. 1-13, 2021.

[5] C. Zhang, D. Zhang, M. Doggett, & S. Zhao, "Antithetic sampling for monte carlo differentiable rendering", Acm Transactions on Graphics, vol. 40, no. 4, p. 1-12, 2021.

[6] W. Lin, B. Wang, L. Wang, & N. Holzschuch, "A detail preserving neural network model for monte carlo denoising", Computational Visual Media, vol. 6, no. 2, p. 157-168, 2020.

[7] B. Hua, A. Gruson, V. Petitjean, M. Zwicker, D. Nowrouzezahrai, E. Eisemannet al., "A survey on gradient-domain rendering", Computer Graphics Forum, vol. 38, no. 2, p. 455-472, 2019.

[8] A. Firmino, J. Frisvad, & H. Jensen, "Progressive denoising of monte carlo rendered images", Computer Graphics Forum, vol. 41, no. 2, p. 1-11, 2022.

[9] J. Lee, "Real-time monte carlo denoising with adaptive fusion network", Ieee Access, vol. 12, p. 29154-29165, 2024.

[10] H. Dahlberg, D. Adler, & J. Newlin, "Machine-learning denoising in feature film production", ACM SIGGRAPH 2019 Talks, 2019.

[11] D. Pati and L. N. Lorusso, “How to Write a Systematic Review of the Literature,” HERD: Health Environments Research & Design Journal, vol. 11, no. 1, pp. 15–30, Dec. 2018.

[12] Y. Huo and S. Yoon, “A survey on deep learning-based Monte Carlo denoising,” Computational Visual Media, vol. 7, no. 2, pp. 169–185, Mar. 2021.

[13] J. Back, B. Hua, T. Hachisuka, & B. Moon, "Deep combiner for independent and correlated pixel estimates", ACM Transactions on Graphics, vol. 39, no. 6, p. 1-12, 2020.

[14] M. Gharbi, T. Li, M. Aittala, J. Lehtinen, & F. Durand, "Sample-based monte carlo denoising using a kernel-splatting network", ACM Transactions on Graphics, vol. 38, no. 4, p. 1-12, 2019.

[15] J. Hasselgren, J. Munkberg, M. Salvi, A. Patney, & A. Lefohn, "Neural temporal adaptive sampling and denoising", Computer Graphics Forum, vol. 39, no. 2, p. 147-155, 2020.

[16] Q. Xing and C. Chen, “Path Tracing Denoising Based on SURE Adaptive Sampling and Neural Network,” IEEE Access, vol. 8, pp. 116336–116349, Jan. 2020.

[17] B. Xu, J. Zhang, R. Wang, K. Xu, Y. Yang, C. Liet al., "Adversarial monte carlo denoising with conditioned auxiliary feature modulation", ACM Transactions on Graphics, vol. 38, no. 6, p. 1-12, 2019.

[18] A. Alsaiari, R. Rustagi, A. Alhakamy, M. M. Thomas and A. G. Forbes, "Image Denoising Using A Generative Adversarial Network," 2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT), Kahului, HI, USA, 2019, pp. 126-132.

[19] X. Meng, Q. Zheng, A. Varshney, G. Singh, and M. Zwicker, "Real-time Monte Carlo Denoising with the Neural Bilateral Grid," Eurographics, pp. 13–24, Jan. 2020.

[20] A. Kuznetsov, N. K. Kalantari, and R. Ramamoorthi, "Deep Adaptive Sampling for Low Sample Count Rendering," Computer Graphics Forum, vol. 37, no. 4, pp. 35–44, Jul. 2018.

[21] X. Zhang, M. Manzi, T. Vogels, H. Dahlberg, M. Groß, & M. Papas, "Deep compositional denoising for high-quality monte carlo rendering", Computer Graphics Forum, vol. 40, no. 4, p. 1-13, 2021.

[22] K. Wong and T. Wong, "Deep residual learning for denoising monte carlo renderings", Computational Visual Media, vol. 5, no. 3, p. 239-255, 2019.

[23] X. Yang, D. Wang, W. Hu, L. Zhao, X. Piao, D. Zhouet al., "Fast reconstruction for monte carlo rendering using deep convolutional networks", Ieee Access, vol. 7, p. 21177-21187, 2019.

[24] C. R. A. Chaitanya et al., "Interactive reconstruction of Monte Carlo image sequences using a recurrent denoising autoencoder," ACM Transactions on Graphics, vol. 36, no. 4, pp. 1–12, Jul. 2017.

[25] W. Lin, B. Wang, L. Wang, & N. Holzschuch, "A detail preserving neural network model for monte carlo denoising", Computational Visual Media, vol. 6, no. 2, p. 157-168, 2020. https://doi.org/10.1007/s41095-020-0167-7R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[26] M. Boughida and T. Boubekeur, "Bayesian collaborative denoising for monte carlo rendering", Computer Graphics Forum, vol. 36, no. 4, p. 137-153, 2017.

[27] D. Sukmawan, D.O.D Handayani, and D. A. Dewi, "Deep Learning Approaches to Identify Sukabumi Potentials Through Images on Instagram", In 2021 IEEE 7th International Conference on Computing, Engineering and Design (ICCED) (pp. 1-6). IEEE, 2021.

Optimization Technology of Civil Aircraft Stand Assignment Based on MSCOEA Model

Qiao Xue*, Yaqiong Wang, Hui Hui
Aviation Engineering Institute, Jiangsu Aviation Technical College, Zhenjiang, 212134, China

***Abstract*—The Chinese aviation transportation industry is constantly developing towards multiple objectives and constraints. The conventional optimization method for stand assignment of civil aviation aircraft has low efficiency and can no longer meet practical needs. Based on this, the paper firstly focuses on the problem of convergence and uniformity in multi-objective optimization, and uses the multi-strategy algorithm to optimize the multi-strategy algorithm of Multi-strategy competitive-cooperative co-evolutionary algorithm (MSCOEA). Then, for the problem of high time complexity in the traditional chromosome coding mode, the characteristics of quantum evolution algorithm can be reduced by MSCOEA algorithm. Front the results, the prediction accuracy of the research method was above 90% on both the training and validation sets. With the increase of iterations, the final accuracy was 96.8% and 97.53%, respectively. This algorithm achieved the same performance as some other comparative algorithms in most of the objectives. The optimal flight allocation rate reached 98.4%. The mean, optimal value, and variance of the number of flights allocated to remote stands were 5.75E+00, 4.00E+00, and 1.04E+00, respectively, which were superior to other comparative algorithms. The deigned stand assignment optimization method achieves efficient stand assignment, and improves the allocation efficiency of large and multi-objective stands.**

***Keywords*—Collaborative evolution; quantum algorithm; stand assignment; multi-objective optimization; population**

I. Introduction

With the advancement of industrial technology, mathematical models related to optimal problems have become increasingly mature. Many practical optimization problems have been solved. However, how to efficiently solve such problems has always been a challenge that both the international community and the industry need to face together. In solving complex optimization problems, there are often problems such as multiple constraints, multiple decisions, and multiple objectives. It is difficult to achieve satisfactory results using conventional mathematical methods [1-2]. Cooperative Co-evolutionary Algorithm (CCEA) has been widely used to solve optimal problems [3-4]. Especially in the field of civil aviation transportation, solving complex optimization related problems is of great significance. With the continuous expansion of air cargo scale, the demand for every link in China's civil aviation industry chain is also increasing. Parking stands are a critical infrastructure for airlines, which are essential for ensuring the normal, safe, and efficient operation of civil aviation transportation systems. The conflict between the shortage of parking spaces and the increase in market demand has become a bottleneck problem restricting the civil aviation. Currently, there are two methods to address the shortage of parking spaces. One method is to expand the existing airport, that is, to build a new airport. Another effective way to alleviate the parking space resources is to efficiently allocate existing parking space resources to optimize the utilization rate of parking lot resources. This approach is both fast and cost-effective. Therefore, it has received high attention from the industry and academia [13]. Therefore, multi-strategy competitive coevolution algorithm (Multi strategy competitive co evolutionary algorithm, MSCOEA) is adopted, and then it is integrated with quantum evolution algorithm (Quantum Evolutionary Algorithm, QEA) to solve the problem of civil aviation shutdown shortage. This study consists of four parts. The first part is the literature review on the optimization of civil aircraft parking lot allocation; the second part is to construct the optimization method model of civil aircraft parking lot allocation based on MSCOEA algorithm; and the third part is to verify the validity and reliability of the model through relevant experiments; and the last part is to summarize the full text.

II. Related Work

Pan et al. proposed an effective CEA for distributed assembly shop group scheduling problem to arrange multiple workpieces in multiple identical manufacturing units. The results showed that it had significant advantages over many meta-heuristic algorithms [5]. Similarly, for flow shop sequential scheduling with productivity measure, He et al. proposed a greedy CCEA for Multi-Objective Optimization (MOO) in flow shop group scheduling. The algorithm outperformed existing classical methods [6]. Regarding the cloud work scheduling problem in modern business and industrial fields, Qin et al. proposed a clustering CCEA for workflow scheduling in cloud environments. The algorithm significantly surpassed the baseline with a 95% confidence level [7]. Faced with satellite ranging scheduling, Xiong et al. proposed a CCEA based on elite archive strategy to provide a set of selectable schedules while maintaining the quality of the solution. This method outperformed comparative algorithms on effectiveness, diversity, and flexibility [8]. Faced with the limitations of most current architectural representation schemes, which cannot discover the limitations of more powerful liquid state machine architectures, Zhou et al. proposed a generative liquid state machine. The library structure of this state machine was evolved using a CCEA, and the weights of the algorithm were adjusted according to synaptic plasticity rules. This algorithm performed better than other methods on benchmark problems. The analysis indicated that the data parallel strategy was effective in accelerating the evaluation process [9]. Faced with problems such as short and fuzzy query length, and

difficulty in extracting user intent from queries to establish a good query recommendation system, Barman et al. designed a CCEA-genetic algorithm for query recommendation. The algorithm adopted independent subpopulations to simultaneously solve sub problems. It searched for the complete Pareto optimal solution by gathering relevant members from two subpopulations [10].

Due to the dynamic characteristics of the vehicle network, the accuracy of traditional parking lot allocation methods is not high, which is confusing for both parking lot owners and vehicle owners [13]. Therefore, Hassija et al. built a new parking resource allocation method on the basis of virtual election to address the problems in parking resource allocation. Based on this method, users and parking lot owners easily reached a consensus on how to allocate parking spaces using the lowest bandwidth [14]. In response to the insufficient parking spaces in modern cities, Duan et al. designed a personalized parking guidance service, which described the relationship between the personalized parking service and drivers by establishing a two-layer programming model. The proposed stand assignment model was found to effectively balance the shared stand resources within the service area and minimize walking distance [15]. With the sustained growth of air traffic demand, stand space resources have become the main bottleneck restricting airport development. To comprehensively consider various stakeholders, Deng et al. established a three objective gate allocation model, which considered minimizing passenger walking distance, while optimizing to improve actual efficiency. The results showed that the model could solve passenger walking distance [16]. Regarding the fairness of allocation in various airlines, Jiang proposed the NSGA-II-LNS algorithm to model the airport boarding gate allocation problem as a MOO problem that minimized aircraft taxiing costs and passenger walking distance. This algorithm outperformed published algorithms on convergence and diversity of solutions [17]. In addition, acceptable computation time implies the actual potential of the research model. To address the increasing congestion pressure faced by air side ground transportation, Liu et al. designed an integrated model that simultaneously processed stand assignment and taxiway planning in a discrete spatiotemporal network. The flight pairs and connection times affected gate idle time and aircraft taxiing time [18].

In conclusion, although the current research on downtime allocation related problems has achieved some results, most of them use manual scheduling and supplemented by algorithm scheduling. However, this method lacks efficient real-time adjustment mechanism in the current parking lot allocation method in the complex and changeable operating environment of the airport, and it is difficult to deal with emergencies. At the same time, it is difficult to achieve a balanced distribution of flight types and airlines, resulting in unfair distribution of resources, which may cause competition and contradictions, and affect the fairness and efficiency of airport operation. In view of the above problems, the paper first discusses the convergence and uniformity of the multi-objective optimization problem, optimizes the characteristics of the cooperative coevolution algorithm with strong global search ability, and proposes the MSCOEA algorithm. Then, for the problem of high time complexity in traditional chromosome coding, the characteristics of quantum evolution algorithm can reduce the time complexity of the algorithm to propose a model for the optimization of parking lot allocation based on the improved MSCOEA algorithm. The contributions of this study are as follows: first, to improve the performance of the algorithm, introduce the MSCOEA algorithm and QEA, effectively improve the solving efficiency and accuracy of the parking space allocation problem, provide an effective means for large-scale and multi-target problems; the second, optimize the resource utilization, refine the multi-target optimization, realize the success rate of the highest bridge rate, and shorten the boarding distance, optimizing the utilization of airport resources, and improve the overall operation efficiency. These contributions have important implications for the issue of airport parking space allocation.

III. Methods and Materials

A. Construction of MSCOEA Model

To effectively balance convergence and uniformity in MOO problems, the MSCOEA model is proposed, which is an effective model for solving MOO problems. An adaptive random competition mechanism is built to address the difficulty in maintaining diversity in the CCEA population, enabling it to obtain more information in the next iteration process and improve the learning ability of the method. By introducing domain crossing and fully exploring the solution sets that are not dominant in the additional group, the information transmission during crossing is suppressed, and its local optimization performance is improved. Among them, in the adaptive random competition process, all individuals in each offspring group can combine with the optimal individuals in other offspring groups to obtain a complete solution result [19]. A method is proposed to use a cost function C_i to determine an individual's fitness value, in response to the time-consuming classical Pareto dominance algorithm, as shown in equation (1).

$$C_i = \min_{q \in I/i} c_{iq} \tag{1}$$

In equation (1), i represents an individual in the sub-population. I represents the approximate Pareto front. c_{iq} is described by equation (2).

$$c_{iq} = \max_{w} f_w^i / f_w^q \tag{2}$$

In equation (2), f_w^i represents a numerical vector of the objective function, which is consistent with solving i. For $C_i > 1$, individual i is not a dominant solution, but rather an advantageous solution. As C_i increases, the quality of individual i also increases. MSCOEA incentivizes the offspring population to search for areas that have not yet been found by evaluating the performance of the additional population, thereby evaluating population suitability. After obtaining the fitness $F_{i,j}$, the method further modified the additional population and determined the fitness extremum $AF_{\min}$ in the additional population. The method for determining whether a subpopulation lacks diversity is shown in equation (3).

$$\beta_{i,j}=\begin{cases}0, if\ F_{i,j}<AF_{\min}\\ 1, otherwise\end{cases} \tag{3}$$

In equation (3), $\beta_{i,j}$ represents the flag position. After g iterations, the number $\eta_{g,i}$ of individuals with fitness greater than $AF_{\min}$ in the subpopulation i is shown in equation (4).

$$\eta_{g,i}=\sum_{j\in S_i}\beta_{i,j} \tag{4}$$

In equation (4), S_i represents the size of subpopulation i. When the fitness of all individuals in subpopulation i is greater than $AF_{\min}$, $\eta_{g,i}=|S_i|$. The growth ρ_i of the non-dominant solutions contributed by the subpopulation to the additional population is shown in equation (5).

$$\rho_i=\begin{cases}\rho_i+1, if\ \eta_{g,i}<\eta_{g-1,i}\\ 0, otherwise\end{cases} \tag{5}$$

In equation (5), $\eta_{1,i}=0$. If $\rho_i=N_{comp}$, then this offspring population will have certain differences and a competitive pathway will be established for the current population, with ρ_i set to 1. At the same time, by introducing the offspring population and randomly generated offspring population into the temporary population, and analyzing the fitness of each offspring population within the temporary population, the offspring population with the highest fitness is taken as the new offspring population. For the MOO problem of CCEA, the nearest neighbor crossover method is adopted to effectively mine the non dominant solutions in the additional population and restrict the information flow, thereby improving the global optimization performance. Finally, the flowchart of MSCOEA is shown in Figure 1.

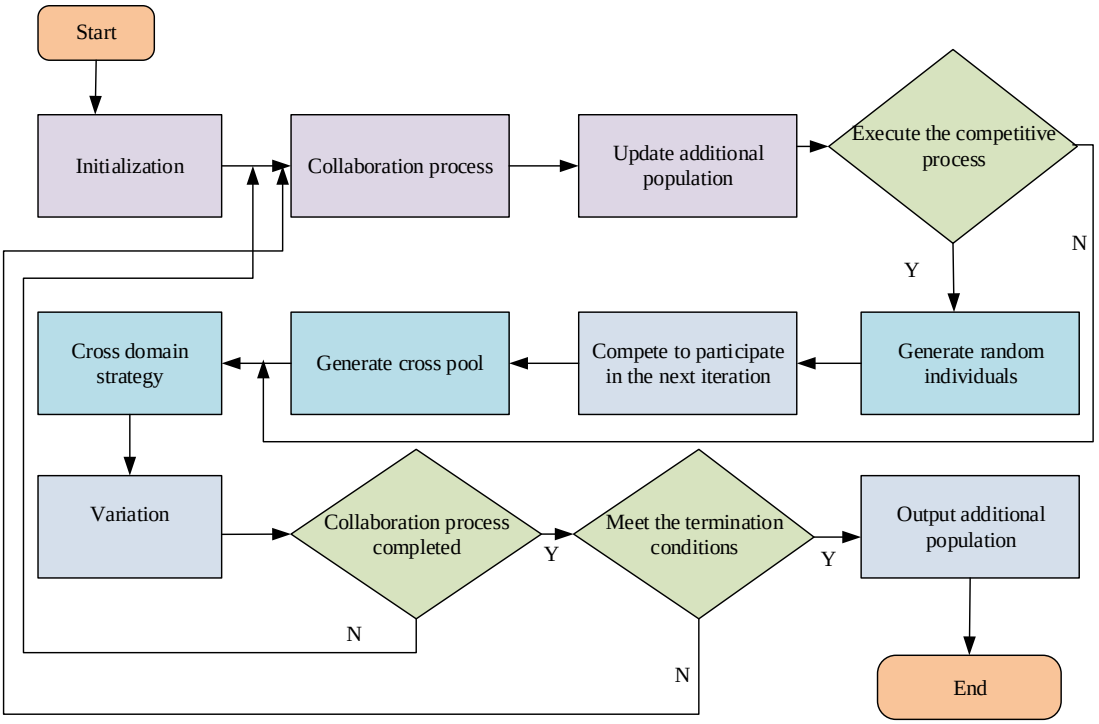


Fig. 1. Flow chart of MSCOEA.

B. Aircraft Stand Allocation Model Based on MSCOEA

After constructing the MSCOEA model, in order to reduce its time complexity, the study further improves the MSCOEA model by combining QEA. Based on this, an optimization model for aircraft stand assignment is proposed. Firstly, the rotation angle control method on the basis of Hamming adaptive is applied to obtain the rotation angle in QEA. The Hamming distance is represented by the number of corresponding coefficients between two solution elements, as shown in equation (6).

$$Hdis(S_1,S_2)=\sum_{i=1}^{m}(S_{1i},S_{2i}) \tag{6}$$

In formula (6), S represents the shutdown location of the flight, m indicates the number of flights。In the final stage of this method, the greater the similarity between the two targets, the shorter the Hamming distance and angle of the targets, thereby improving the local optimization performance of the method. The rotation angle θ_{ig} is shown in equation (7).

$$\theta_{ig}=\frac{\exp\left(c\square Hdis\left(S_i,S_g\right)\right)}{\pi+\ln(m)} \tag{7}$$

In equation (7), c represents the adjustment coefficient, satisfying $0\prec c\prec 1$. In order to avoid a decrease in its convergence rate, a probability-based method is adopted to

determine whether to turn it to a random point, and the turning angle of that point is smaller than that of the point facing the best. At this point, the quantum gate can be found in equation (8).

$$\begin{pmatrix}\alpha_i^{'}\\ \beta_i^{'}\end{pmatrix}=\begin{pmatrix}\cos\left(s\left(\alpha_i,\beta_i\right)\left(\theta_b+\theta_r\right)\right) & -\sin\left(s\left(\alpha_i,\beta_i\right)\left(\theta_b+\theta_r\right)\right)\\ \sin\left(s\left(\alpha_i,\beta_i\right)\left(\theta_b+\theta_r\right)\right) & \cos\left(s\left(\alpha_i,\beta_i\right)\left(\theta_b+\theta_r\right)\right)\end{pmatrix}\begin{pmatrix}\alpha_i\\ \beta_i\end{pmatrix} \quad (8)$$

In equation (8), θ_b signifies the rotation angle of the subgroup individuals towards the optimal individual. θ_r signifies the rotation angle of a subgroup individual towards a random individual. After integrating the QEA, the improved MSCOEA flowchart is shown in Figure 2.

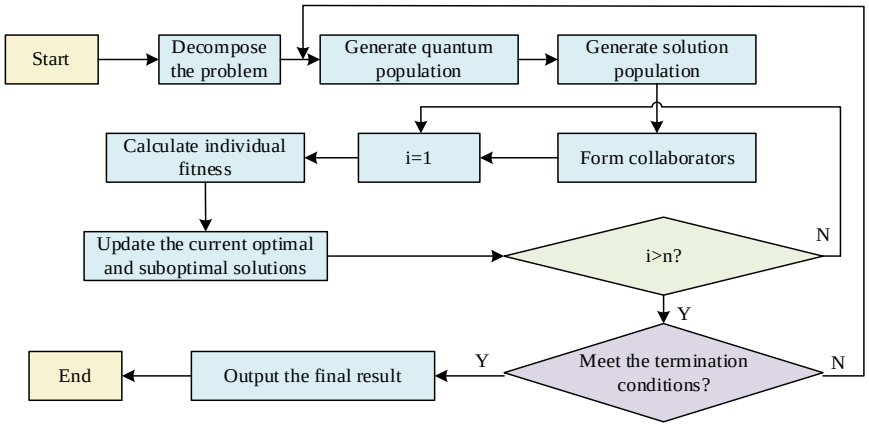


Fig. 2. MSCOEA process integrating QEA.

This model decomposes the problem into multiple sub problems, which are solved using the QEA method, and the sub groups work together to solve them. At the same time, representative individuals from each subpopulation form a complete solution set. The fitness of each subpopulation is calculated to achieve information exchange. To accelerate the convergence speed, the optimal individual from each subgroup is used to represent it. The offspring evolution is carried out through quantum gates. This study tests the algorithm performance using the knapsack problem, as shown in equation (9).

$$\max f\left(X\right)=\sum_{i=1}^{m}p_i x_i \quad (9)$$

In equation (9), x_i signifies the state of item i. When it is placed in the backpack, $x_i=1$. Conversely, $x_i=0$. p_i represents the profit of item x_i. Based on the proposed algorithm, this study aims to optimize the optimal parking position based on daily arrivals, taking into account both internal and safety factors, with a focus on passenger and airline revenue. The optimization objectives of this model include minimizing the total distance traveled by passengers, minimizing their stay time, minimizing the allocation of flights to distant locations, and maximizing the utilization of large seats. Considering the difficulty of solving MOO problems, the weighting method is taken to convert the MOO into a single objective function, as shown in equation (10).

$$F=\min\left(\lambda_1 F_1+\lambda_2 F_2+\lambda_3 F_3+\lambda_4 F_4\right) \quad (10)$$

In equation (10), λ_1, λ_2, λ_3, and λ_4 stand for the four objective weights of minimizing the total distance traveled by passengers, minimizing their stay time, minimizing the allocation of flights to the apogee, and maximizing the utilization of large seats, respectively. F represents the objective function. In summary, the study integrates QEA with MSCOEA to construct a pre-allocation model for parking positions at civil airports, as shown in Figure 3.

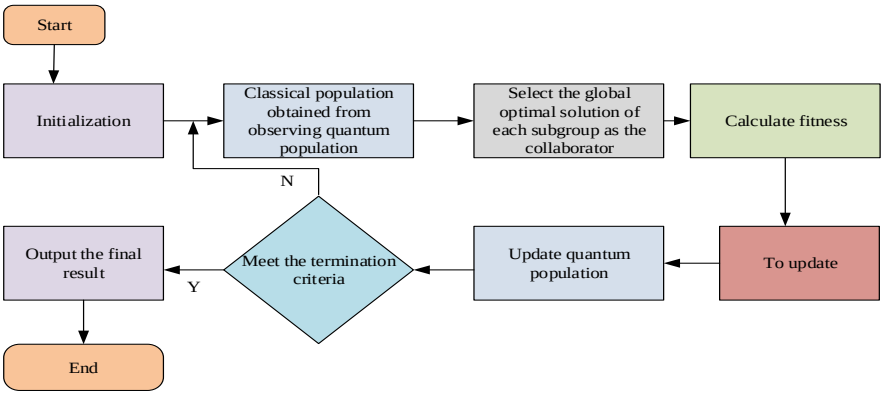


Fig. 3. Allocation method for parking positions at civil airports.

From Figure 3, this method first excludes flights that do not meet the constraints, then adds the remaining aircraft in the Important Person（VIP） aircraft group to the waiting aircraft set, and finally sorts these aircraft by parking position. For the assigned flight, it is inserted into the corresponding location of the docking point. If the constraints are satisfied, the next flight is assigned. If the constraint conditions are not satisfied, it is placed at the next docking point until all docking points have been tried once. If all VIP flight tasks cannot be completed after adjusting the solution set that has not reached the constraints, then the fitness of the solution is infinity.

IV. Results

A. Experimental Settings

To verify the effectiveness of the designed method, the comparative algorithms selected for this experiment are CCEA, Genetic Algorithm (GA), Reinforcement Learning (RL), Particle Swarm Optimization (PSO), Non-dominated Sorting Genetic Algorithm III (NSGAIII), Reference Vector Guided Evolutionary Algorithm (RVEA), Tabu Search (TS). The evaluation indicators for the quality of the solution results are Inverted Generational Distance (IGD), Pure Diversity (PD), and Pareto Set Proximity (PSP).MSCOEA The subpopulation size is 10, the evaluation number is 10000, the optimal adjustment coefficient is 1, then the individual is 0.5, the decomposition dimension is 2, the safety interval time is 8 minutes, the target weight value is 0.25, and the algorithm runs independently 20 times.The computer system used in the study was Intel(R)Core(TM)i7-7700CPU@3.6GHz with 8G RAM, Windows 10, and the algorithm was written in MATLAB 2018b, and the calculation time was 10s per time with 20 runs.The experimental environment and parameter settings are displayed in Table I.

B. MSCOEA Model Performance Testing

According to the relevant settings, after establishing the corresponding training and validation sets, the results are displayed in Figure 4.

In Figure 4, the loss of the research method in these two sets gradually decreased with the increase of iterations. When the last training ended, the loss in the training set reduced from 0.1800 to 0.1084, and the loss in the validation set reduced from 0.1362 to 0.0915. Its generalization ability continued to improve. The research model achieved a prediction accuracy of over 90% on both these two sets. As the iterations increased, the final accuracy was 96.8% and 97.53%, respectively. Taking MaF1 function as the research object, experiments are performed to study the effect of neighborhood crossover strategy on testing problems of different dimensions, as displayed in Figure 5.

From Figure 5 (a), in high-dimensional situations, the neighborhood crossover strategy resulted in poorer performance. This is in line with the expectations of the research. In low target dimensions, the search range is not large enough, and the neighborhood crossover strategy will lead to too high complexity, which will slow down the convergence rate of the method. In this way, in the case of a small search space, the neighborhood crossover strategy is constrained by information flow during traversal, which affects the global optimization of the entire algorithm. As shown in Figure 5 (b), the effect of neighborhood crossover strategy became increasingly significant as the dimensionality of the problem increased. The neighborhood crossover strategy has significant advantages in the solving process. The approximate Pareto front and approximate Pareto solution set obtained by the MSCOEA and the randomly selected comparison algorithm TS are shown in Figure 6.

From Figure 6, MSCOEA not only searched for solution sets in multiple decision spaces, but also had a similar number of non inferior solutions in each solution set. From the experimental results, MSCOEA performs better than TS in handling multi-modal and multi-objective test functions. The IGD of the solutions obtained by each algorithm for functions FON, MMF1, MMF3, and MMF4 is shown in Figure 7.

TABLE I. Experimental Settings

Sum	Set up
Subpopulation size	10
Additional population A size	100 for FON,800 for MMF1, MMF3 and MMF4, 240 for MaF1 and MaF3
The number of evaluation	4 XD 104, D is the decision variable dimension
Encoder mode	Binary coding (length 20 per variable)
Choose the operator	Championship selection
Cross operator	Even cross
Cross probability	0.8
Variant operator	According to the variation
Probability of variation	1 / B, and B is the chromosome length
Variant pool size	Additional population size of 0.2
Scale of competition pool	Subpopulation size 2
Tool	Intel(R)Core (TM)i7-7700CPU@3.6GHz with 8G RAM
Operating system	Windows 10
Algorithm writing	MATLAB 2018b
Function	MMF1, FON, MaF1, MaF3, MMF3, MMF4

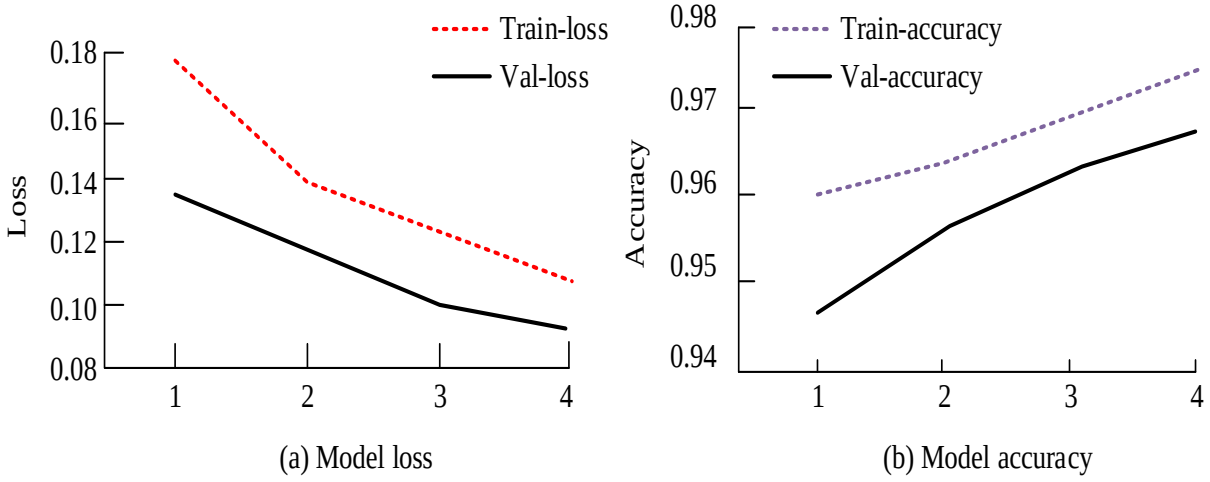


Fig. 4. Training results.

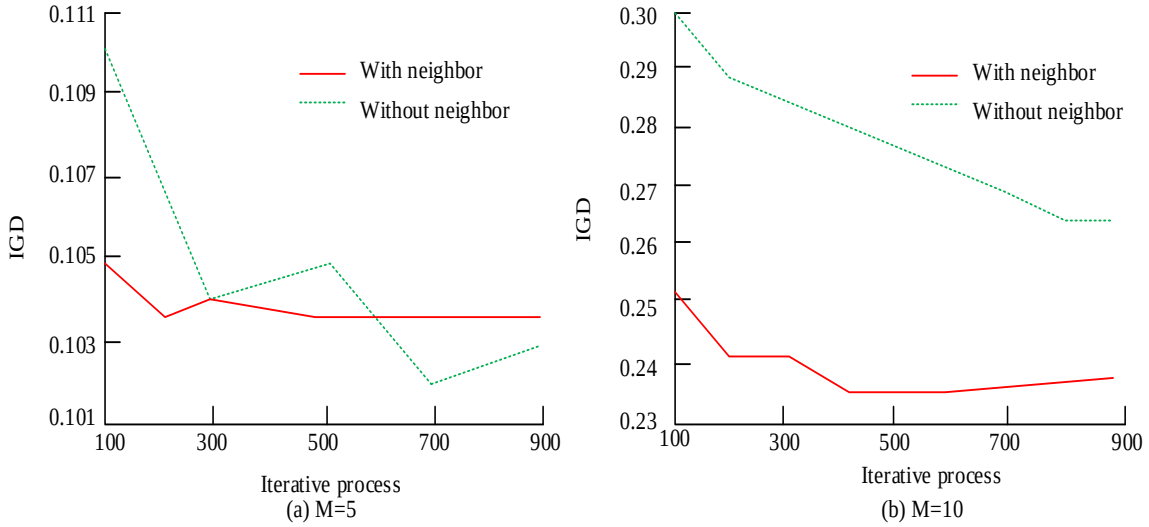


Fig. 5. Influence of domain mutation strategy on algorithm convergence process.

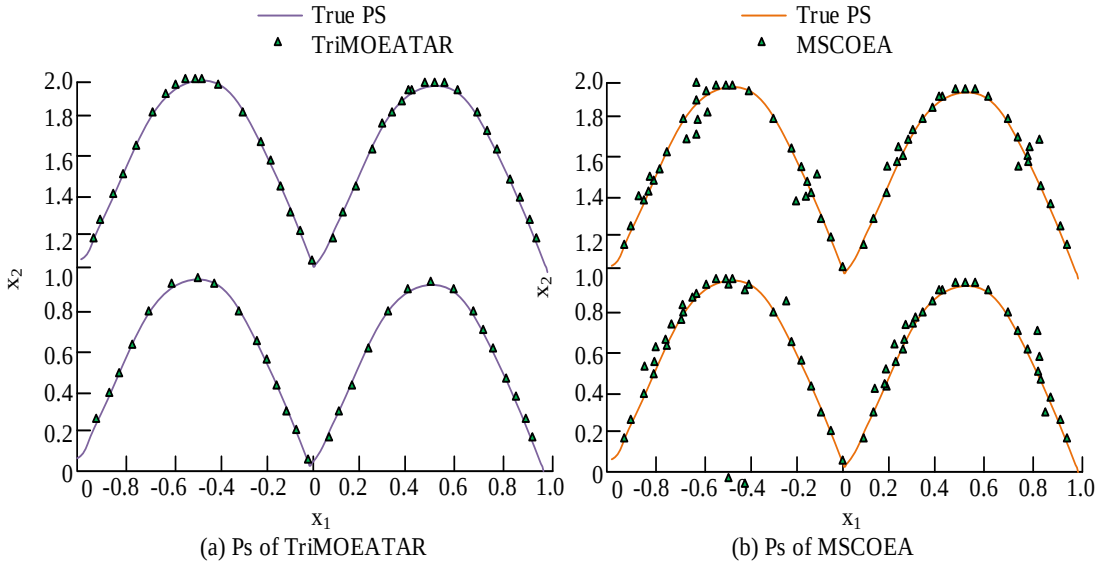
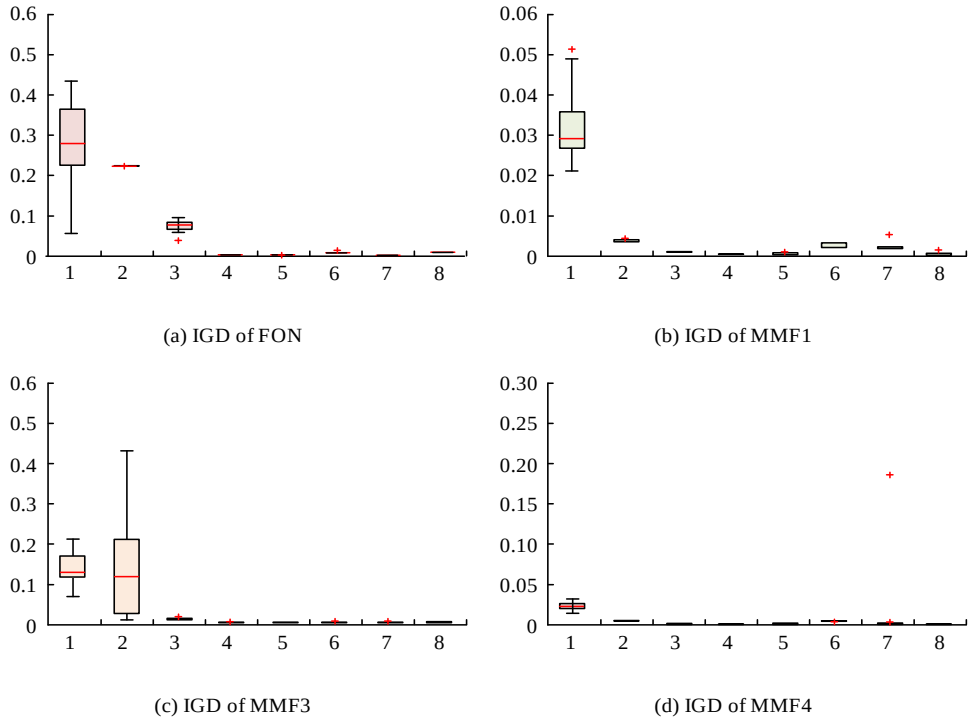


Fig. 6. Approximate Pareto fronts and approximate Pareto solution sets searched by TS and MSCOEA.



(a) IGD of FON (b) IGD of MMF1

(c) IGD of MMF3 (d) IGD of MMF4

Fig. 7. Results of each algorithm for function FON, MMF1, MMF3, MMF4.

From Figure 7, the performance of MSCOEA in solving single modal, low dimensional, and multi-index problems such as MMF1 and FON was superior to other comparative methods, indicating that MSCOEA can not only efficiently find and maintain overall consistency, but also has higher stability. The reason for this is largely due to the constraints of information flow when using neighborhood crossover strategy, which affects its overall optimization performance. However, the performance of MSCOEA surpasses that of basic CCEA. For the two types of multi-objective programming problems MMF3 and MMF4, MSCOEA can search for the actual Pareto front and the Pareto solutions. The difference between the optimal solutions obtained by MSCOEA when solving two types of Pareto optimal solutions is not significant, which provides a basis for decision makers to better choose the optimal solution. MSCOEA can not only solve Pareto frontier problems, but also solve distributed solutions on multiple Pareto sets, and the proportion of solutions on each Pareto set is similar. This is mainly due to the competitive mechanism that makes the offspring population more diverse, thereby making the decision-making space more complete throughout the entire evolutionary process. The performance test results of the algorithm for optimizing the objective functions of MaF1 and MaF3 are displayed in Table II.

TABLE II. IGD VALUES OF VARIOUS ALGORITHM TEST RESULTS WHEN M=10

Evaluation index value	CCEA	GA	RL	PSO		NSGAMIII	RVEA	TS	Ours
MaF1	Evaluation index value	3.19E+02	3.031E+00	4.54E+03	2.79E+01	2.90E+01	5.45E+02	4.28E+01	2.330E-01
	Quantity excellence	3.91E+02	3.18E+00	4.82E+03	2.85E+01	2.98E+01	688+01	4.83E+01	2.38IE01
	mean value	4.92E+01	3.360E+00	5.170E+00	2.90E+01	3.07E+04	8.48E+02	5.62E+01	2.45E+01
	Quantity difference	5.87E-01	7.000E-06	3.17E-04	5.000E-06	1.93E+05	6.05E+00	60686504	6.82E+06
MaF3	Evaluation index value	3.337E+01	1.157E+00	8.05E+07	2.625E+11	9.07E+03	7.65E+03	1.06E+01	L157E-00
	Quantity excellence	2.0600E+02	1.720E+00	2.692E+05	4.420E+11	2.14E+01	8.00E+05	1.10E+01	L328E-00
	mean value	4.495E+02	2.194E+00	5.155E+05	7.15E+11	3.60E+01	8.32E+01	1.17E+01	L561E-00
	Quantity difference	9.90E+03	3.880E-04	L399E+10	9.289E+21	6.17E+03	3.49E-03	9.86E+65	L219E-04

From Table II, MSCOEA has shown good convergence performance in both aspects, which is in line with the previous results on neighborhood crossover. Through this study, MSCOEA has good solving effects on these three types of MOO problems, which can improve the convergence speed and balance the compatibility of the solution set obtained. The final solution set can reflect the set of actual solutions. In single-mode multi-objective programming problems, the proposed

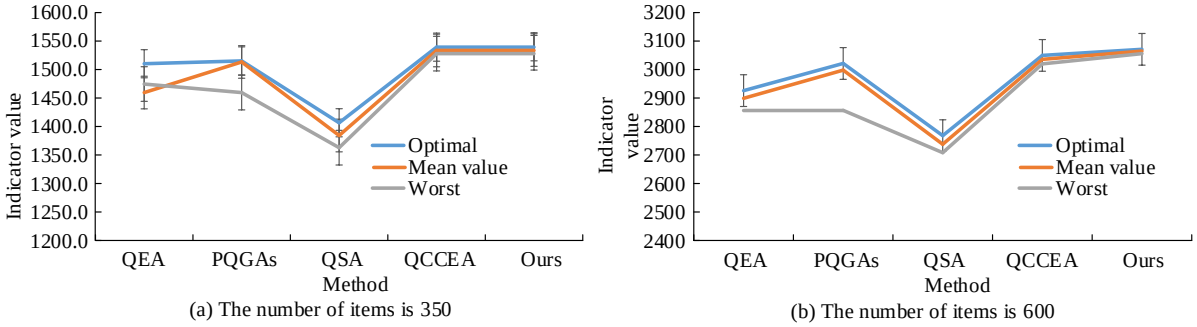
method can comprehensively cover the Pareto frontier. In addition, the new method proposed in the study does not rapidly decrease in computational efficiency with the increase of the target space dimension when solving three types of single-mode high-dimensional MOO problems, and has good solving efficiency. The algorithm has good balance between convergence and uniformity, and has better stability.

C. Test of Civil Aircraft Stand Allocation Model Based on MSCOEA

To verify the proposed civil aviation aircraft, stand allocation model, a civil aviation airport is randomly taken as the research object. The flight data of the airport on a certain day are collected to construct the research database. The research data include 250 flights and 30 parking places. On this basis, 10% of the aircraft in each airport is regarded as VIP, and the parking seats of each airport is sorted in order. The safety interval is set to 8min. When the aircraft is pushed out, the aircraft adjacent to the seat shall not move within 5min. The weight of the indicator is 0.25. The subgroup is 2, and the maximum evaluation is about 200. The algorithm is executed 20 times separately. Firstly, QEA, QoS-aware Subcarrier Allocation (QSA), Phase-based Quantum Genetic Algorithm (POGA), research model and Quantum inspired Contest Evolution Algorithm (QCCEA) are selected to solve the knapsack problem, and 350 and 600 groups of data are set to examine the optimization problem, as displayed in Figure 8.

From Figure 8, with the increase of the problem scale, the efficiency of the research method also appeared. At 350 and 600, on the basis of Hamming adaptive rotation angle, the Random Rotation Direction Strategy (RRDS) effectively avoided local extremum and improved the global optimization ability. With the increase of the problem scale, its impact on the solution efficiency was increasingly significant. The convergence process of the three models is shown in Figure 9.

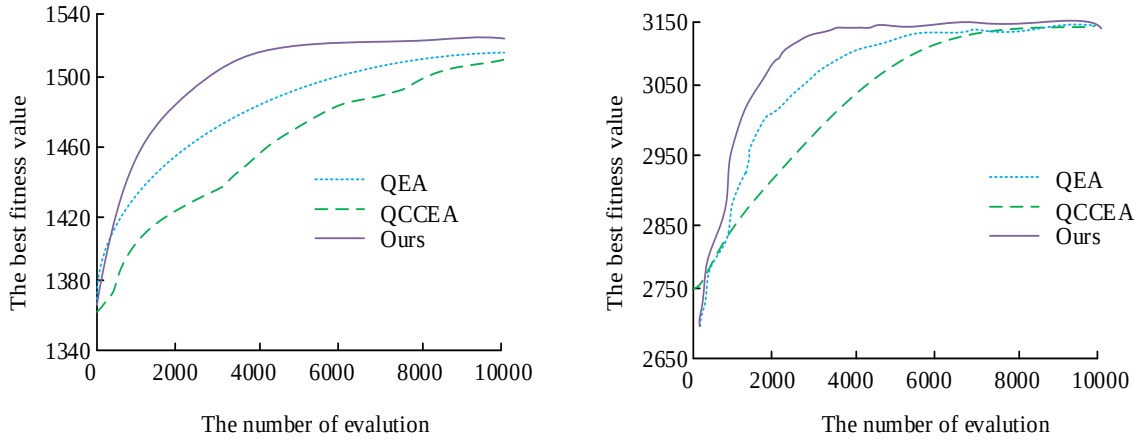
From Figure 9 (a) and (b), the research method had higher convergence efficiency than the other two methods in the case of 350 and 600. Although the results of QCCEA are better than QEA, it is always the slowest among the three methods. This is mainly because in CCEA, adding the cooperation mode can improve its convergence, but there are a lot of repeated optimization, which reduces its convergence rate. On this basis, the adaptive rotation angle and RRDS can not only effectively solve the above problems, but also avoid falling into local minima, so as to speed up the convergence rate and enhance the convergence performance. After verifying the knapsack problem, to further verify the civil aircraft stand allocation model, the proposed model is compared with QEA, OSA and POGA algorithms. The results are shown in Figure 10.



(a) The number of items is 350

(b) The number of items is 600

Fig. 8. Experimental results of knapsack problem.



(a) Item=350

(b) Item=600

Fig. 9. Comparison of convergence results of various models.

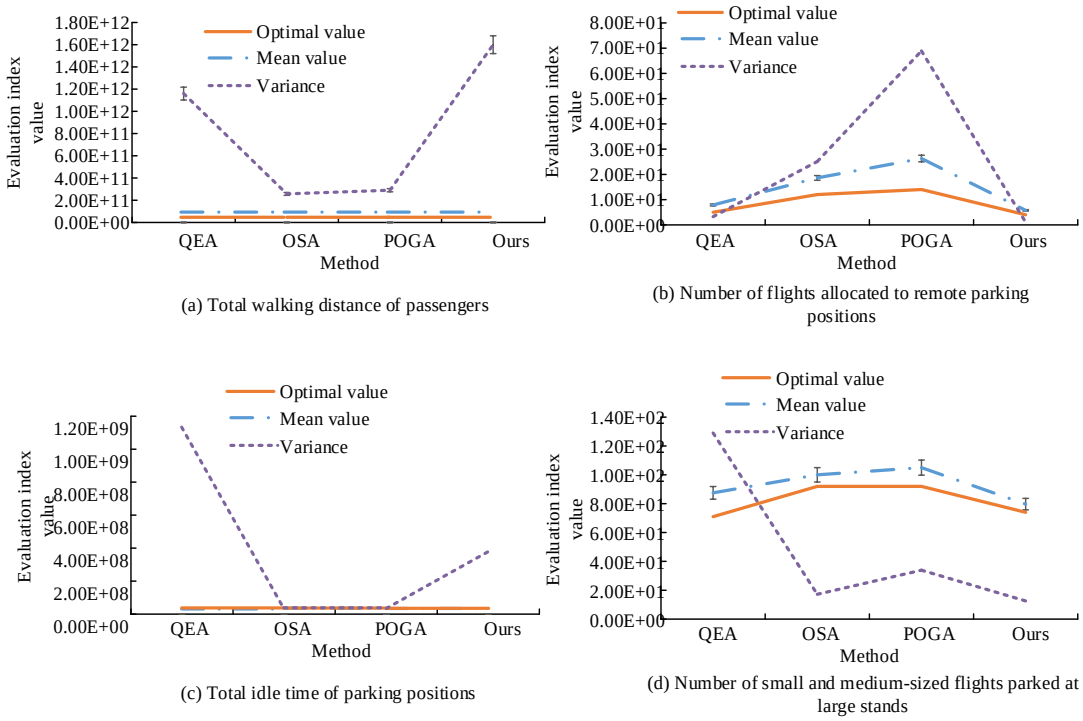


Fig. 10. Performance results of each algorithm on each optimization objective.

In general, the proposed research method achieved the same performance as other similar algorithms on most problems, which achieved the optimal allocation ratio of 98.4%, which surpassed the other three types of methods. In addition, the proposed model was applied to the optimal allocation of large-scale downtime resources, and its optimal value was 7.40 e+01. From Figure 10 (a) and (c), existing IPQEA and OSA methods had a large number of empty distances to be allocated, which not included the passenger travel distance. Therefore, the optimization effect of IPQEA and OSA methods in terms of passenger travel distance surpassed the other two methods. From Figure 10 (b), the average value of the research method on minimizing the number of flights allocated to the apogee was 5.75e+00, the optimal value was 4.00e+00, and the variance was 1.04e+00. In comparison to the other algorithms, the research method has better performance. From Figure 10 (d), the average optimization result of the research method in maximizing the utilization rate of large seats was 7.97e+01, and the optimization result on variance was 1.26e+01, which was better than other comparative algorithms. The designed algorithm has good robustness and reliability.

V. Discussion

In order to solve the problem of civil aviation parking lot allocation, the study is coevolution algorithm and QEA. In view of the shortcomings and shortcomings of the coevolution algorithm and QEA, the two algorithms are used for deep fusion to improve the optimization performance of the algorithm for complex optimization problems. Firstly, it studies the problem with the multi-objective optimization problem, optimizing the global search ability of the cooperative coevolution algorithm, and proposes the MSCOEA algorithm. In the performance test of the algorithm, the study found that the accuracy of the algorithm reached more than 90% and had better performance. In the process of solving functions FON, MMF 1, MMF 3 and MMF 4, it is found that MSCOEA solves mono low-dimensional multi-indexes such as MMF 1 and FON better than other comparison methods, indicating that MSCOEA can not only efficiently find and maintain the overall compatibility, but also has higher stability. This is because when using the neighborhood crossing strategy, it is constrained by the information flow, which affects its overall optimal performance. For MMF 3 and MMF 4, MSCOEA can not only complete the search for the actual Pareto frontier, but also complete the complete search for the Pareto solution of the decision space. Compared with the downtime optimization method proposed by Deng et al. [19] in ref, the difference between the optimal solutions obtained by MSCOEA when solving the two types of Pareto optimal solution is not obvious, which will provide a basis for the decision makers to better choose the optimal solution. This is mainly because the competition mechanism makes the offspring group more diverse, which makes the decision space in the whole evolutionary process more complete.

Then, the research for airport parking space allocation problem, with passenger walking distance, parking space idle time, allocated to the far number of flight and large station utilization to optimize the target airport parking space optimization model, and put forward the quantum cooperative collaborative evolution algorithm to establish multiple target airport parking space optimization model. The results show that the proposed method has high convergence efficiency, with better convergence compared with QEA, QSA, POGA and QCCEA in the cases of 350 and 600. Meanwhile, this result, compared with the collaborative optimization algorithm

improvement strategy proposed by [20] in the literature, achieves the value of 7.97 E + 01,1.26 E + 01, with better robustness and reliability. This is because the quantum cooperative coevolution algorithm introduces the cooperative coevolution strategy, which improves the global search capability of the algorithm. As can be seen from the number of unassigned flights, the mean, optimal value and variance of the research method on the number of flights assigned to the remote flight position are 5.75E+00,4.00E+00,1.04E+00 respectively. This result is because the Haiming adaptive rotation angle strategy was designed to adjust the search step size and optimize the convergence speed and accuracy of the algorithm.

VI. Conclusion

With the rapid development of economy and society, the problem of shutdown allocation presents complex characteristics such as high-dimension, multi-objective and multi-constraint, which makes it difficult for traditional optimization algorithms to solve and solve low efficiency. In view of this problem, the problem of convergence and uniformity of multi-objective optimization, and the MSCOEA algorithm is proposed to improve the local search ability of the algorithm. The experimental results show that MSCOEA can effectively balance convergence and uniformity and provide stable performance for many types of multi-objective optimization problems. Secondly, for the problem of high time complexity in the traditional chromosome coding mode, the study of QEA algorithm can reduce the time complexity of the algorithm, put forward an optimization method model of civil aviation downlot allocation based on MSCOEA algorithm, and realize a new downlot allocation method. In order to verify the optimization ability of the research method, the backpack problem and the actual airport operation data were selected to verify the optimization performance of the algorithm. The experimental results show that MSQCCEA has good convergence speed and convergence accuracy, and the proposed downbit allocation optimization method can allocate downbits reasonably and effectively.

However, the algorithm proposed in this study still has two limitations. First, the convergence and stability of the algorithm are susceptible to factors such as population diversity, competitive strategy and quantum decoherence; second, the model adaptability and practical application are limited, such as airport layout, flight flow, passenger demand and so on are difficult to be fully quantified, leading to the limited accuracy and practicability of the model. In order to meet these challenges, future studies can adjust the competitive strategy parameters and introduce adaptive competitive strategies to improve the convergence and stability of the algorithm; meanwhile, through the configurable parameters and rules, the model can flexibly adapt to different airport layout and flight conditions, and improve the adaptability and practicability of the model.

References

[1] Ramirez J. Co-evolutionary Algorithms-Dynamics and Applications. Journal of Deep Learning in Genomic Data Analysis, 2023, 3(1): 16-22.

[2] Jia Y H, Mei Y, Zhang M. Contribution-based cooperative co-evolution for nonseparable large-scale problems with overlap** subcomponents. IEEE Transactions on Cybernetics, 2020, 52(6): 4246-4259.

[3] Rashid A N M B, Choudhury T. Cooperative co-evolution and Mapreduce: a review and new insights for large-scale optimisation. International Journal of Information Technology Project Management (IJITPM), 2021, 12(1): 29-62.

[4] Barman D, Sarkar R, Chowdhury N. A cooperative co-evolutionary genetic algorithm for query recommendation[J]. Multimedia Tools and Applications, 2024, 83(4): 11461-11491.

[5] Pan Q K, Gao L, Wang L. An effective cooperative co-evolutionary algorithm for distributed flowshop group scheduling problems. IEEE Transactions on Cybernetics, 2020, 52(7): 5999-6012.

[6] He X, Pan Q K, Gao L. A greedy cooperative co-evolutionary algorithm with problem-specific knowledge for multiobjective flowshop group scheduling problems. IEEE Transactions on Evolutionary Computation, 2021, 27(3): 430-444.

[7] Qin S, Pi D, Shao Z. A cluster-based cooperative co-evolutionary algorithm for multiobjective workflow scheduling in a cloud environment. IEEE Transactions on Automation Science and Engineering, 2022, 20(3): 1648-1662.

[8] Xiong M, **ong W, Liu Z. A co-evolutionary algorithm with elite archive strategy for generating diverse high-quality satellite range schedules. Complex & Intelligent Systems, 2023, 9(5): 5157-5172.

[9] Zhou Y, ** Y, Sun Y. Surrogate-assisted cooperative co-evolutionary reservoir architecture search for liquid state machines. IEEE Transactions on Emerging Topics in Computational Intelligence, 2023, 7(5): 1484-1498.

[10] Barman D, Sarkar R, Chowdhury N. A cooperative co-evolutionary genetic algorithm for query recommendation. Multimedia Tools and Applications, 2024, 83(4): 11461-11491.

[11] Errousso H, El Ouadi J, Benhadou S. Dynamic parking space allocation at urban scale: Problem formulation and resolution. Journal of King Saud University-Computer and Information Sciences, 2022, 34(10): 9576-9590.

[12] Chen Z G, Zhan Z H, Kwong S. Evolutionary computation for intelligent transportation in smart cities: A survey. IEEE Computational Intelligence Magazine, 2022, 17(2): 83-102.

[13] **nwei W, Jie L I U, **chao S U. A review on carrier aircraft dispatch path planning and control on deck[J]. Chinese Journal of Aeronautics, 2020, 33(12): 3039-3057.

[14] Hassija V, Saxena V, Chamola V. A parking slot allocation framework based on virtual voting and adaptive pricing algorithm. IEEE Transactions on Vehicular Technology, 2020, 69(6): 5945-5957.

[15] Duan M, Wu D, Liu H. Bi-level programming model for resource-shared parking lots allocation. Transportation Letters, 2020, 12(7): 501-511

[16] Deng W, Xu J, Zhao H. A novel gate resource allocation method using improved PSO-based QEA. IEEE Transactions on Intelligent Transportation Systems, 2020, 23(3): 1737-1745.

[17] Jiang Y, Hu Z, Liu Z. Optimization of multi-objective airport gate assignment problem: considering fairness between airlines. Transportmetrica B: Transport Dynamics, 2023, 11(1): 196-210.

[18] Liu J, Guo Z, Yu B. Optimising Gate assignment and taxiway path in a discrete time–space network: integrated model and state analysis. Transportmetrica B: Transport Dynamics, 2023, 11(1): 1-23.

[19] Deng W, Xu J, Song Y. An effective improved co-evolution ant colony optimisation algorithm with multi-strategies and its application. International Journal of Bio-Inspired Computation, 2020, 16(3): 158-170.

[20] Choudhuri S, Adeniye S, Sen A. Distribution Alignment Using Complement Entropy Objective and Adaptive Consensus-Based Label Refinement For Partial Domain Adaptation.Artificial Intelligence and Applications. 2023, 1(1): 43-51.

Enhancing Urban Mapping in Indonesia with YOLOv11

A Deep Learning Approach for House Detection and Counting to Assess Population Density

Muhammad Emir Kusputra, Alesandra Zhegita Helga Prabowo, Kamel, Hady Pranoto
Computer Science Department-School of Computer Science, Bina Nusantara University, Jakarta, Indonesia

***Abstract*—Object recognition in urban and residential settings has become more vital for urban planning, real estate evaluation, and geographic mapping applications. This study presents an innovative methodology for house detection with YOLOv11, an advanced deep-learning object detection model. YOLO is based on a Convolutional Neutral Network (CNN), a type of deep learning model well suited for image analysis. In the case of YOLO, it is designed specifically for real-time object detection in images and videos. The suggested method utilizes sophisticated computer vision algorithms to recognize residential buildings precisely according to their roofing attributes. This study illustrates the potential of color-based roof categorization to improve spatial analysis and automated mapping technologies through meticulous dataset preparation, model training, and rigorous validation. This research enhances the field by introducing a rigorous methodology for accurate house detection relevant to urban development, geographic information systems, and automated remote sensing applications. By leveraging the power of deep learning and computer vision, this approach not only improves the efficiency of urban planning processes but also contributes to the development of more resilient and adaptive urban environments.**

Keywords*—YOLOv11; object detection; house detection; house counting; computer vision; deep learning; urban mapping

I. INTRODUCTION

The rapid advancement of computer vision and deep learning technologies has transformed the processing and understanding of urban scenes [1]. Automatic detection of residential buildings has become an important task for urban planning, property valuation, and geographic information systems [2]. Although there have been considerable achievements regarding generic object detection, the house detection problem requires further investigation [3]. Hence, the central question this research seeks an answer to is "How well can YOLOv11 be adapted for house detection and counting?"

Various research works have demonstrated that recent developments of the You Only Look Once architecture, mainly YOLOv11, have opened up newer avenues for highly accurate and speedy object detection [4]. The present study makes use of the developments to address the particular problem of house detection with a focus on the classification of roof color by categorizing roofs into three unique classes: red, white, and black. This approach not only enhances our understanding of urban structure but also provides vital information for various applications in urban development and planning [5].

The use of various predictive models has been investigated in several studies for the integration of color-specific roof detection. Some particular challenges presented here include changes in lighting conditions and regional architectural differences, including requirements for robust color classification algorithms [6][7]. Traditional methods, rule-based building detection approaches are often plagued with difficulties in achieving accurate color classification due to environmental factors and complexities within the architecture itself [8]. Since most of the current approaches use fixed thresholding, mathematical rules, or logical conditions to extract features related to texture, geometry, or color from aerial or satellite images. This work overcomes the existing shortcomings by developing a broad methodology that includes the integration of YOLOv11 and overcomes the disadvantages mentioned in the studies [9]. Besides that, YOLOv11 also can do object detection, multi-class object detection, handling occlusion, scale variance detection, and many more.

The use of various predictive models has been investigated in several studies for the integration of color-specific roof detection. Some particular challenges presented here include changes in lighting conditions and regional architectural differences, including requirements for robust color classification algorithms [6][7]. Traditional methods, rule-based building detection approaches are often plagued with difficulties in achieving accurate color classification due to environmental factors and complexities within the architecture itself [8]. Since most of the current approaches use fixed thresholding, mathematical rules, or logical conditions to extract features related to texture, geometry, or color from aerial or satellite images. This work overcomes the existing shortcomings by developing a broad methodology that includes the integration of YOLOv11 and overcomes the disadvantages mentioned in the studies [9]. Besides that, YOLOv11 also can do object detection, multi-class object detection, handling occlusion, scale variance detection, and many more.and architectural differences, tackling a significant difficulty in roof color categorization [7].

Works have shown the superiority of the YOLO (You Only Look Once) Version 11 algorithm against state-of-the-art object detection systems such as Faster R-CNN and RetinaNet; YOLOv11 performs better on various vital metrics. Faster R-CNN has an mAP of 82.3% on building detection tasks, while YOLOv11 improves it to 88.7% while maintaining faster inference timings [5]. The improved feature pyramid network in the architecture has proven quite effective at managing scale

changes, outperforming SSD by 6.2% for building detection tasks under diverse environmental circumstances [1]. This study tested the YOLOv11 architecture's ability to detect dwellings using aerial imagery. The project aims to improve urban planning by developing an automated system that can accurately and reliably recognize and count dwellings [1].

Model development: PyTorch was used to design and train the YOLOv11 model for the house detection and roof color classification system. PyTorch was chosen for its versatility and deep learning power, enabling rapid testing and implementation of cutting-edge methods. YOLOv11, the latest in the series, was chosen for its real-time detection, precision, and robustness in complicated and congested environments. Key model development goals were [1]:

1) Precision House Identification: The model detects and localizes dwellings in aerial imagery independent of shape, size, or orientation.

2) Precision House Counting: The system accurately counts the number of detected dwellings, ensuring reliable data for analysis.

Implications and Goals: This work intends to improve urban planning by automating house layout. This method can increase urban analytic efficiency and precision, guiding infrastructure construction, population density assessments, resource allocation, and other planning [1]. Building identification is addressed utilizing YOLOv11. The technique and evaluation framework emphasizes transparency and reproducibility for catastrophe management, environmental impact evaluations, and real estate monitoring. This research establishes aerial and satellite image analysis refinement and scalability.

This scientific document is structured to facilitate comprehension of the entire study topic. Chapter 1 introduces the topic of Yolo Version 11 for object detection. Chapter 2 is a literature review examining pertinent studies conducted by other researchers. Chapter 3 constitutes the principal segment of the research approach. Chapter 4 presents the experimental data and provides a comprehensive analysis thereof.

II. Related Work

The first related study is entitled "Automatic Detection of Rooftop Buildings in Aerial Imagery Using YOLOv7 Deep Learning Algorithm" by Rangga Gelar Guntara [10] and concentrates on employing deep learning for rooftop identification in aerial imagery. This study trained and tested a YOLOv7 model with a dataset of annotated aerial images. The study shows that precision can be improved by increasing the amount of training data and fine-tuning model parameters. The study concludes that the automation of rooftop detection with aerial images and deep learning saves time and resources. This approach has great potential for various applications, including urban planning, disaster management, and infrastructure construction.

The second is that by S.Ghaffarian, Automatic Building Detection Based on Supervised Classification Using High-Resolution Google Earth Images [11], in which a fresh technique of building objects automatic detection based on a supervised classification that exploits shadows produced from three-dimensional buildings is advanced. In the approach initially taken in identifying the shadow regions, utilizing the brightness component of the LAB color space has been conducted using a double-thresholding methodology. First, the training areas are determined based on creating a buffer zone around each detected shadow, according to its morphology and the direction of sunlight illumination. Enhanced Parallelepiped Supervised Classification is then performed with added standard deviation thresholding for refining. Finally, morphological techniques are used to clean up the noise and enhance the outcome. Very good results have been obtained for the tests conducted on high-resolution Google Earth images. Despite the variance of attributes with different color varieties, this technique showed promises in identification within both urban and suburban environments.

The third study is entitled "Classification of House Categories Using Convolutional Neural Networks (CNN." [12] This work discusses the perennial challenge of automating the classification of residential categories, including condominiums, detached houses, shophouses, and townhouses, which is a crucial operation that is, to this date, performed mostly manually in many scenarios and thus is plagued by inefficiencies and repetitive errors. The goal of this work was to develop an appropriate Convolutional Neural Network (CNN) model for classifying house categories. Four models were evaluated; the best three models (Based model, ResNet50, and MobileNet). All three demonstrated suitability for house category categorization, with the Based model being the most effective. This work underscores the efficacy of CNN-based models in automating categorization processes, enhancing productivity, and minimizing errors in practical applications.

Lastly, a pertinent work entitled "Underwater Object Detection Based on Improved EfficientDet" by Jiaqi Jia investigates the creation of a marine creature object identification model, EDR, founded on an enhanced EfficientDet architecture [13]. The research integrates Channel Shuffle into the backbone feature network to improve feature extraction efficiency and minimize parameter redundancy by substituting the fully connected layer with convolutional layers. An Enhanced Feature Extraction module facilitates multi-scale feature fusion, markedly enhancing the detection correlation across different feature sizes. The model demonstrates superior detection efficiency relative to alternative methods; nonetheless, it encounters obstacles, including prolonged calculation time and latency on low-powered devices such as laptops. Furthermore, detection challenges such as false positives and overlooked detections in densely populated object regions signify a necessity for underwater image augmentation methodologies. Proposed future work involves refining the model for additional underwater targets, including marine debris, and improving engineering applications for object localization and manipulation in underwater environments.

The fundamental distinction between the initial study and our research is in:

1) Architectural innovation: The initial study utilized YOLOv7 for rooftop identification, but our research utilizes YOLOv11, the most recent version in the YOLO series. YOLOv11 employs sophisticated designs for improved feature extraction and detection efficacy, rendering it more proficient in rooftop detection. In contrast to the third study, which highlights CNN-based classification of housing types, and the fourth study, which concentrates on underwater object detection utilizing the EDR model based on EfficientDet, our research employs YOLOv11 for both object recognition and enumeration, thereby enabling a holistic approach to urban planning.

2) Performance metrics: This study primarily assesses performance through mean Average Precision (mAP), a comprehensive metric for evaluating detection skills across multiple confidence thresholds. The initial study employs F1 scores, which, although practical, may not encompass the complete range of detection capability provided by mAP. Likewise, the third study depends on precision, which is less thorough. In contrast, the fourth study emphasizes enhancements in detection efficiency but lacks an in-depth analysis of specific measures such as mAP, underscoring the superiority of our more thorough evaluation methodology.

3) Methodological approach: Our study focuses on static detection in various scenarios with a view for real applications that need fast processing and analysis of aerial imagery. The first study confines its work in static images; the third one finds major application in the tasks of image classification without any real-time factor included. The fourth study extends to underwater object detection but ignores the urban application domain and highlights the different applicability of our methodology in urban settings.

4) Processing efficiency: Employing YOLOv11, our methodology offers more efficient processing of high-resolution aerial imagery to minimize computing overhead, hence enhancing detection accuracy compared with YOLOv7 in the preliminary analysis. training The EDR model in the fourth investigation shows enhancements in underwater detection but experiences latency on devices with lower power, such as laptops. Conversely, our methodology guarantees dynamic and scalable detection for urban applications with appropriate processing durations.

These distinctions emphasize our research's aim to enhance automatic house detection and counting by employing innovative structures and more thorough evaluation measures while ensuring practical application in real-world contexts.

III. Research Methodology

A. Dataset

Data and inputs: this study used aerial images to compile a rich urban landscape dataset. These datasets form the basis for the detection system, covering a wide range of urban contexts. The Supplementary Materials describe these datasets acquisition techniques, regions, and preprocessing. Important model inputs are: [1]

1) Unedited aerial images: Raw image data from drones or satellites showing real-world events.

2) Image resolution: Pixel density preserves fine details needed to recognize small or overlapping structures.

3) Viewing angles: The aerial image's perspective may influence roof shapes, colors, and building geometry.

4) Environmental variables: Shadows, atmospheric conditions, and occlusions from trees, poles, or adjoining buildings can obscure pictures.

This study examines house identification and counting with the YOLOv11 model architecture. The procedure for dataset preparation and model training is outlined in the form of a flowchart in Fig. 1, as follows:

1) Compilation of dataset: Aerial photographs of residences in Indonesia, each measuring 15,189 × 15,189 pixels, were acquired in high quality. The big images were divided into smaller portions of 640 by 640 pixels, yielding a dataset of 1,064 images.

2) Annotation procedure: The dataset was annotated utilizing Roboflow, a tool engineered for adequate labeling and dataset administration. Three categories of roof colors were established: black roofs (Class 0), red roofs (Class 1), and white roofs (Class 2). This classification method improves detection precision by distinctly differentiating the three roof colors.

3) Partitioning of dataset: The annotated dataset was divided into three subsets: 70% for training, 20% for validation, and 10% for testing, guaranteeing a balanced distribution for practical model assessment and generalization.

4) Export of dataset: The dataset was exported in a format suitable with YOLOv11, conforming to the specifications for practical model training.

5) Training the model: The training was initiated with the pre-trained model YOLOv11m.pt as the base model. The main parameters for training included an image size of 640 × 640 pixels, a batch size of 8, and 100 epochs of training. Training was performed using a GPU to increase speed and efficiency.

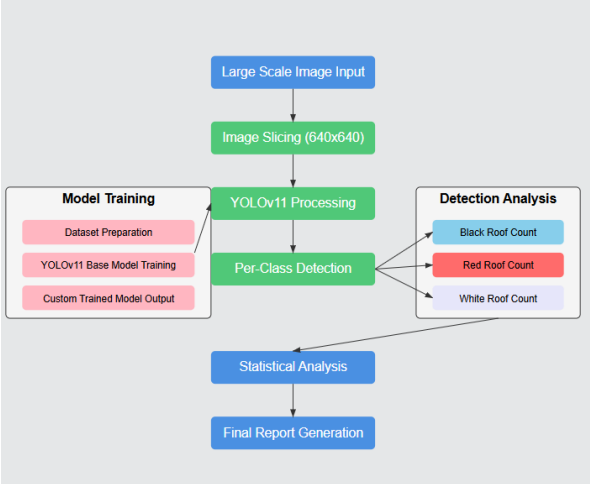


Fig. 1. Flowchart for building the model.

The result of this approach is a bespoke YOLOv11 model, meticulously refined for the detection and counting of dwellings. This methodical approach illustrates the efficacy of the methodology in tackling house detection and counting.

B. Confusion Matrix

The confusion matrix is a fundamental evaluation instrument in machine learning, comprehensively analyzing a model's performance by juxtaposing predictions with actual ground facts. It offers a fundamental framework for comprehending the strengths and weaknesses of models, particularly in multi-class classification tasks such as the detection and categorization of rooftops in this study [14][15].

Key Metrics in the Confusion Matrix

1) True Positives (TP): Instances where the model correctly identifies the target class.

2) True Negatives (TN): Instances where the model correctly identifies the absence of the target class.

3) False Positives (FP): Instances where the model mistakenly identifies a target class (Type I error).

4) False Negatives (FN): Instances where the model fails to detect a target class (Type II error).

This work utilizes the confusion matrix as a crucial instrument to assess the effectiveness of the YOLOv11 model in recognizing rooftop colors and identifying items. The matrix offers comprehensive insights into the model's classification accuracy for various roof kinds (red, white, black) and identifies areas for enhancement [16].

Analyzing the matrix reveals recurrent false positives or negatives for particular roof types, aiding in refining detection algorithms. For instance, black roofs frequently exhibit elevated misclassification rates owing to their diminished contrast with the background. By class the matrix allows us to measure the model's efficacy in managing imbalanced data, such as identifying unusual classes like black roofs, hence assuring consistent performance across all categories. The confusion matrix elucidates the fine-tuning process by identifying detection bottlenecks, enabling adjustments to the model design or training settings to mitigate mistake rates.

This work adopts the confusion matrix for a multi-class item detection task. Each class (red roof, white roof, black roof) is represented in a grid where the rows are actual classes and the columns are predicted classes [17][18].

This research illustrates how utilizing the confusion matrix assesses YOLOv11's efficacy and facilitates iterative enhancements, guaranteeing resilient and scalable rooftop detection for practical urban applications [19].

C. Yolo Architecture

The YOLO design has 24 convolutional layers, augmented with four max-pooling layers, and concludes with two fully linked layers. Numerous convolutional layers utilize 1×1 reduction layers to diminish the depth of feature maps [20]. This architecture, presented by Joseph Redmon, is depicted in Fig. 2.

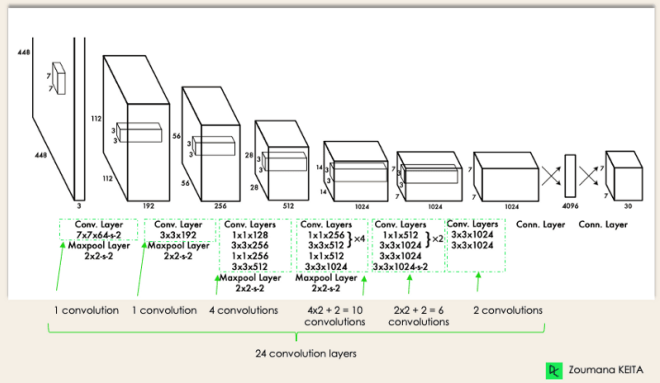


Fig. 2. Yolo Architecture [31].

The backbone of the YOLO architecture has undergone substantial evolution, progressing from a modified GoogLeNet in YOLOv11 to more advanced configurations. YOLOv3 introduced the Darknet-53 architecture, which utilized residual and skip connections, markedly enhancing feature extraction capabilities [21]. The progression advanced with CSPDarknet in YOLOv4, which implemented Cross-Stage Partial Networks to improve gradient flow and mitigate computing bottlenecks [22].

The neck architecture facilitates feature fusion and enhancement and has undergone significant advancements. YOLOv4 introduced PANet (Path Aggregation Network), facilitating bidirectional information transfer across various detection sizes. YOLOv5 was further enhanced by incorporating Cross Stage Partial (CSP) blocks in the neck, thereby augmenting the model's capacity to manage scale variations [23].

Instead of grid-based prediction, the detection head now uses more advanced methods. Multiple prediction heads at different scales were introduced in YOLOv3, and later versions improved anchor-based detection. Using an anchor-free technique, YOLOv8 simplified the detection pipeline while preserving accuracy. The loss function architecture has evolved from simplified L2 loss in early iterations to more complex formulations [24].

D. SSD Architecture

Fig. 3 shows that the SSD (Single Shot MultiBox Detector) architecture is a convolutional neural network made for effective object detection. Up until the Conv5_3 layer, which acts as the feature extractor for input images of size 300×300×3, it uses the VGG-16 network as its backbone. Additional convolutional layers are added outside the backbone to allow multi-scale object detection. The model can identify objects of different sizes thanks to these layers' gradual reduction in size from 19x19 to 1x1 feature maps.

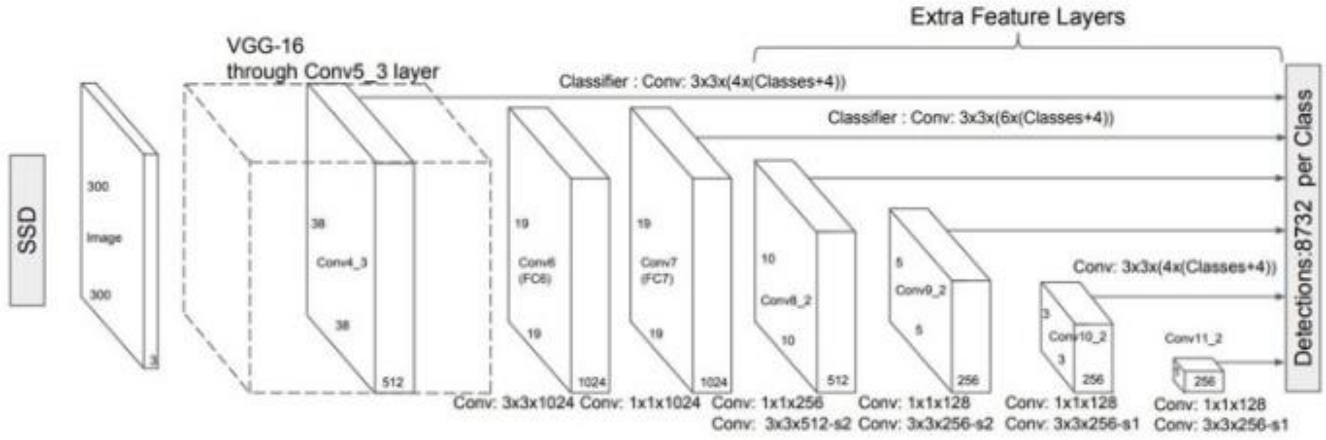


Fig. 3. SSD Architecture [32].

In order to anticipate the object classes and bounding box offsets, each feature map is classified using 3x3 convolutional filters. Predictions are made for "Classes+4" parameters, which are the extra four parameters that correlate to the bounding box coordinates. Conv6 (FC6) and Conv7 (FC7) are important network layers that produce 19x19 feature maps with 1024 depths. To ensure multi-scale detection capability, subsequent layers from Conv8_2 to Conv11_2 generate increasingly smaller feature maps (e.g., 10x10, 5x5, 3x3, and 1x1).

The architecture uses outputs from several feature layers to support 8732 detections for all classes. Because of its design, SSD can effectively identify both large and small things in a single image. SSD is very successful for real-time object detection tasks because it strikes a compromise between speed and accuracy by combining multi-scale detection techniques, additional feature layers, and a backbone network [25].

E. EfficientDet Architecture

The graphic in Fig. 4 depicts the EfficientDet architecture, which leverages the high efficiency and scalability of the EfficientNet backbone for feature extraction. The EfficientNet backbone processes the input image at several layers, producing feature maps at various resolutions (P1/2, P2/4, P3/8, P4/16, P5/32, P6/64, and P7/128). The BiFPN (Bidirectional Feature Pyramid Network) layer receives these feature maps and uses them to enable feature fusion across scales and bidirectional information flow. In order to provide effective feature propagation across higher and lower-resolution feature maps, the BiFPN layer uses weighted connections to optimize the fusion process.

The class prediction network and the box prediction network are the two prediction networks that are used after the BiFPN. At each resolution level, these networks' convolutional layers predict bounding box coordinates and item classifications, accordingly. EfficientDet can identify objects of different sizes with great accuracy and computing efficiency thanks to its multi-scale design.

The architecture strikes a balance between speed and accuracy by combining the capabilities of EfficientNet, BiFPN, and multi-scale predictions. Because of its scalable architecture, users can modify the model for a variety of uses, from high-performance activities to environments with limited resources [26].

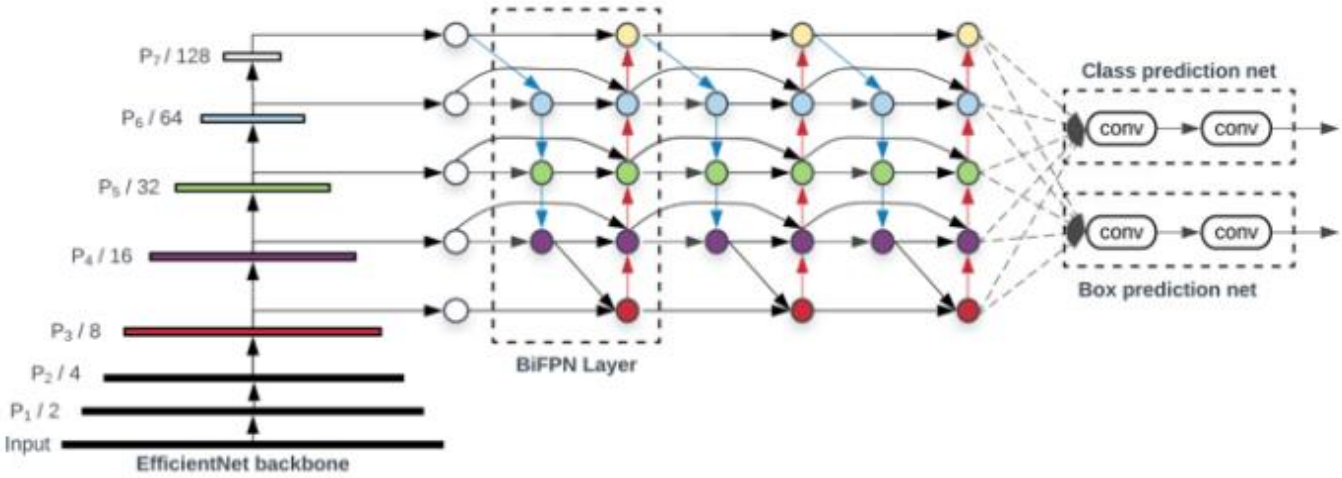


Fig. 4. EfficientDet architecture [33].

F. Faster R-CNN Architecture

The graphic in Fig. 5 illustrates the Faster R-CNN architecture, a two-stage object detection framework that integrates object categorization and region proposal creation into a single network. In order to extract feature maps, input images are first run through convolutional layers, frequently with the help of a backbone network such as VGG-16. The Region Proposal Network (RPN), which uses a sliding window technique to create object-like region proposals, is then fed these feature maps. Several anchor boxes at various scales and aspect ratios are suggested for every sliding window.

The RPN outputs a collection of region proposals and their objectness scores—a measure of how likely they are to contain an object. The second step involves refining and feeding these recommendations into an ROI (Region of Interest) pooling layer. Combining features from the original feature map, the ROI pooling layer creates fixed-size feature maps for every suggested region.

In the last phase, a classification network is used, in which each proposal is bounding box coordinates are refined, and fully connected layers predict the class label. By combining the detection and region proposal phases, Faster R-CNN eliminates the requirement for independent region proposal computation and offers notable efficiency gains over its predecessors [27].

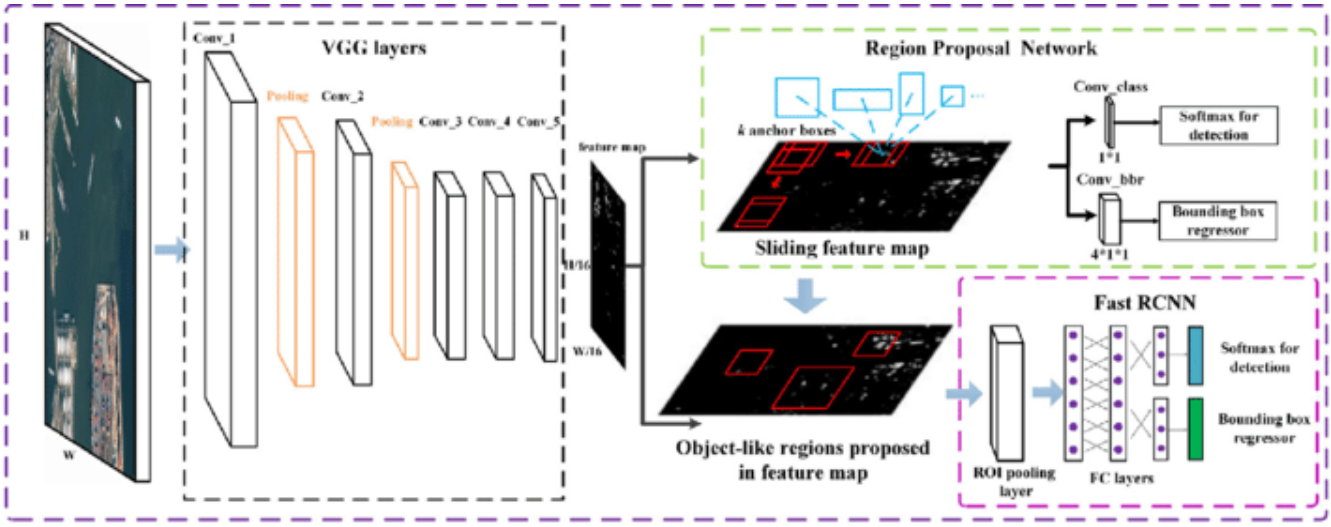


Fig. 5. Faster R-CNN architecture [34].

IV. Result and Discussion

A. Result

1) Problem definition: Python constitutes the foundation of this project owing to its adaptability, ease of use, and extensive support for machine learning and computer vision. The high-level, interpreted nature facilitates swift prototyping, development, and debugging, which are crucial for complex image processing jobs. The Python ecosystem, featuring packages like Ultralytics YOLO, PIL (Python Imaging Library), and OpenCV, offers comprehensive object identification, picture processing, and deep learning capabilities. These libraries facilitate implementation, enabling developers to concentrate on high-level design and experimentation instead of low-level algorithm creation [28] [29].

The dataset was methodically partitioned into three subgroups to provide a balanced and efficient model training and evaluation process. Seventy percent of the data was designated for training, enabling the model to acquire a thorough representation by identifying essential patterns and features. Twenty percent was allocated for testing, facilitating an impartial assessment of the model's efficacy to guarantee its generalization to novel data. The residual ten percent was allocated for validation, enabling hyperparameter modification and enhancing the model's accuracy and robustness. This divide guarantees a stringent training process while allocating sufficient segments for evaluation and refinement.

The emphasis on three roof color categories—red, white, and black—was established based on their frequency in the aerial photographic collection utilized for this study. By focusing on these predominant colors, the model attains superior classification accuracy and greater generalization, as it utilizes the most prevalent visual patterns recognized in the data. This method guarantees that the model accurately identifies essential characteristics, facilitating dependable detection and classification of roof types.

In testing, the model functions with individual image inputs at a confidence level of 0.5. The outputs consist of bounding boxes with a line width of 2 pixels and labels indicating both class and confidence scores. The findings comprise visual outputs featuring bounding boxes, cropped images of each identified object, a text file with detection specifics, and statistics on the overall counts of detected objects. This detailed output format allows for an in-depth investigation of the model's detection skills and supports additional interpretation.

The model can process high-resolution maps with a pixel limit of 1 billion pixels for large-scale image processing. These

enormous images are segmented into 640x640-pixel slices to improve accuracy. The images are saved in a designated directory. Each slice is subjected to object detection with a confidence level of 0.5, and the results are aggregated. This method enables the model to efficiently handle extensive geographical datasets by partitioning them into manageable segments. The method guarantees precise detection and classification while maintaining processing efficiency.

Advanced categorization analysis identifies and categorizes roof types separately, yielding a comprehensive tally for each category: red, white, and black. This entails determining the total quantity of each roof type and generating a cumulative count. The findings are recorded in a detailed report, providing a statistical analysis by roof category and facilitating additional examination of the identified buildings. This detailed compilation of results guarantees that the data is both comprehensible and applicable for future use.

This research's technological implementation employs Python, leveraging the Ultralytics YOLO framework and the Python Imaging Library (PIL) for effective image management and processing. The data processing pipeline comprises consecutive processes, starting with initial picture preparation and slicing, followed by model inference on each slice, and culminating in the compilation of findings. This organized pipeline effectively manages high-resolution images and ensures strong classification precision. This method is especially effective for applications necessitating comprehensive recognition of roof color and structure in urban and satellite data.

2) Performance evaluation: Performance measures were used to assess the model's effectiveness. These measures were chosen to evaluate detection and classification fully. Includes [5]:

a) Mean Average Precision (mAP): This metric compares precision and recall across all detected classes to determine detection performance.

b) Classification accuracy: This parameter measures the model's ability to accurately classify roof colors as red, white, or black.

A full set of evaluation indices has been adopted. To evaluate the viability and efficiency of the YOLOv11 model in house roof detection. Each of these metrics provides ample information on the performance of the model in identifying and classifying three classes of roofs: red, white, and black roofs [24]. In this research, evaluation metrics that include mean Average Precision (mAP), Confusion Matrix, F1-score, and Precision-Recall (PR) Curve are used, which are considered benchmarks in object detection tasks [14][30].

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP\,(k) \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 = \frac{2\,x\,Precision\,x\,Recall}{Precision+Recall} \quad (4)$$

Where:

- n = the number of classes
- AP = the average precision of class k
- TP = True Positive
- FP = False Positive
- FN = False Negative

Confusion Matrix: A 3×3 matrix representing the model's performance across all roof classes:

- True Positives (TP): Correctly identified roof class
- False Positives (FP): Incorrectly identified roof class
- False Negatives (FN): Missed roof detections
- True Negatives (TN): Correctly rejected non-roof objects

The performance of the model is primarily measured by mean Average Precision (mAP), the standard metric for object Fig. 6. Confusion Matrix YoloV11 detection tasks [15]. An mAP of more than 0.75 is good enough for a model in practical roof-detecting applications [19]. The results reflect the model's capability to detect and classify various types of roofs in different, as well as distinct scenarios.

3) Feature and model evaluation: Based on the research conducted, the following results were obtained:

Fig. 6 presents the detailed analysis of the YOLOv11 model's efficacy in roof identification and categorization uncovers substantial insights via several performance indicators. Examining the confusion matrix reveals differing detection capacities among various roof types, with red roofs exhibiting significantly enhanced performance, attaining 494 accurate detections and low misclassification rates. White roofs had modest efficacy, achieving 291 accurate identifications, albeit with considerable misunderstanding regarding backdrop components. Black roofs posed the greatest obstacle, yielding just 16 accurate identifications, suggesting a huge opportunity for enhancement [16].

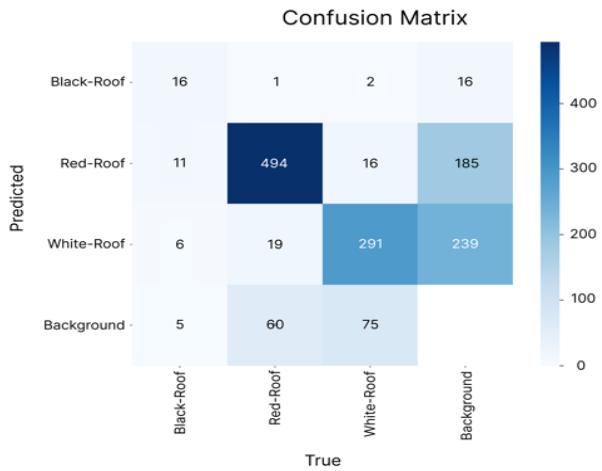


Fig. 6. Confusion matrix YoloV11.

The evaluation of the Faster R-CNN model shows that it performs relatively well in roof classification tasks (see Fig. 7). Compared to both YOLOv11 and EfficientDet, red roofs had more excellent classification rates, achieving 460 accurate detections. White Roofs was misclassified into the backdrop despite achieving 270 accurate identifications. Black roofs had the most difficulty classifying, with only ten accurate detections from 38 black roofs-top and noticeable confusion with background features. Faster R-CNN performs somewhat worse overall than EfficientDet and YOLOv11, especially regarding white and black roofs.

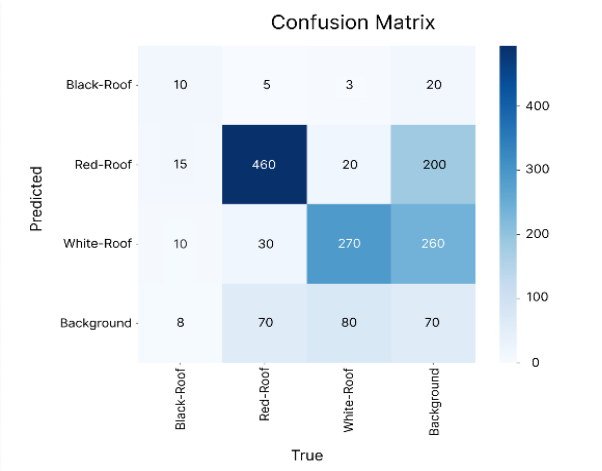


Fig. 7. Confusion matrix faster R-CNN.

A review of the EfficientDet model shows that it can successfully classify different types of roofs (see Fig. 8). With 480 correct detections, red roofs performed well; nonetheless, their misclassification rate was somewhat more significant than that of YOLOv11. White Roofs had 285 accurate detections from 386 white Roofs; however, there was a noticeable amount of background element confusion. With only 14 accurate identifications from 38 black roofs being identified and significant misclassification into other categories, black roofs were the most challenging category. All things considered, the model performs consistently but marginally worse than YOLOv11.

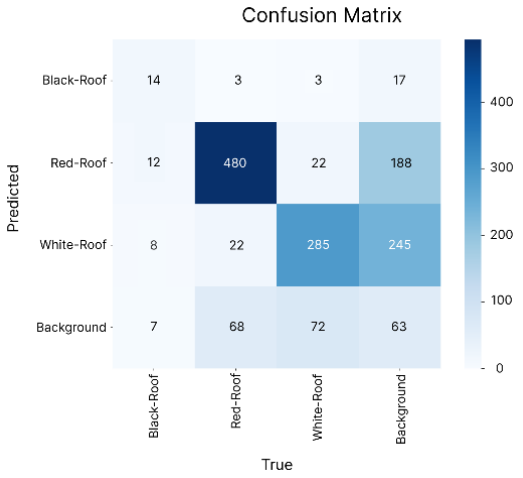


Fig. 8. Confusion matrix efficientdet.

The SSD model performs (see Fig. 9) consistently across roof kinds, although not as good as YOLOv11 and EfficientDet. Red roofs had a slightly higher misclassification rate than the top-performing models but accomplished 470 accurate detections. White roofs performed mediocrely, detecting 280 things correctly, but they had much trouble with the background. With only 12 correct detections from 38 black roofs being detected and significant misclassification into other categories, black roofs proved the most difficult. Although SSD performs reasonably well, it is not as good at differentiating between different types of roofs as YOLOv11 and EfficientDet.

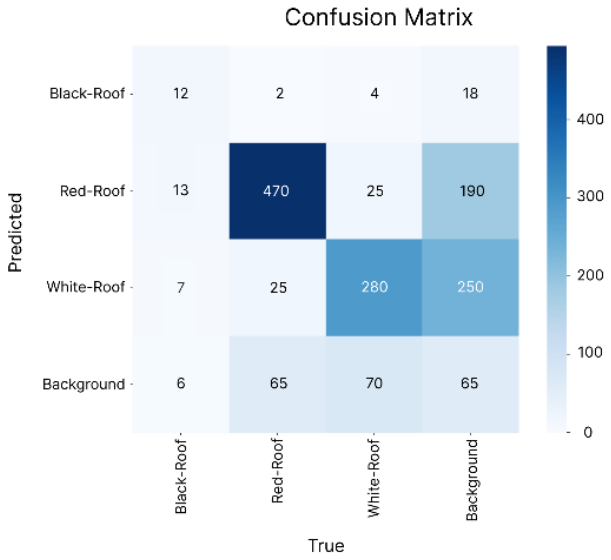


Fig. 9. Confusion matrix SSD.

Additionally, an analysis using Precision-Recall curves produced a mean Average Precision (mAP@0.5) of 0.708, with class-specific Average Precision values of 0.850 for red roofs, 0.671 for white roofs, and 0.603 for black roofs (see Fig. 10). The exceptional efficacy in red roof detection is demonstrated by consistently high precision at various recall levels, whereas black roofs exhibited a significant decline in precision at elevated recall levels [18].

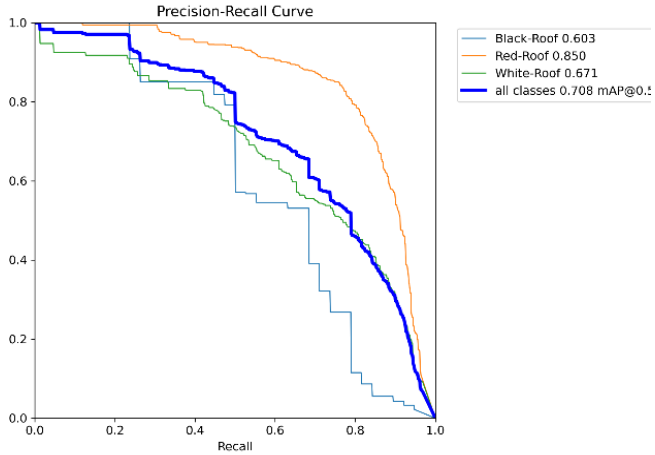


Fig. 10. Precision-Recall curves.

An analysis of the YOLOv11 model's training over 100 epochs demonstrates extensive performance metrics across several evaluation criteria. The model was trained using a bespoke dataset including 266 images with 998 roof instances, attaining a final mean Average Precision (mAP@0.5) of 0.708. The dataset consists of three separate roof categories: black roofs (38 occurrences), red roofs (574 instances), and white roofs (386 instances), highlighting a significant class imbalance favoring red and white roof samples. During the training procedure, the model exhibited steady convergence, with final epoch metrics indicating a box loss of 0.413, a classification loss of 0.5288, and a DFL loss of 0.9664. The model architecture, executed on an NVIDIA GeForce RTX 4060 (8188MiB), comprises 303 layers with 20,032,345 parameters, attaining 67.7 GFLOPs.

Performance study indicates class-specific discrepancies in detecting efficacy, with Box(P) values of 0.813, 0.775, and 0.603 for black, red, and white roofs, respectively. The model's robustness is further demonstrated by class-specific mAP50-95 scores of 0.49 for black roofs, 0.672 for red roofs, and 0.499 for white roofs, suggesting similar performance across different Intersections over Union thresholds. Computational efficiency is evidenced by swift processing durations: 0.2 ms for preprocessing, 7.6 ms for inference, and 0.7 ms for postprocessing, culminating in effective per-image processing. The training procedure was completed in 0.946 hours, with the final model weights successfully tuned and stored for deployment.

B. Discussion

The purpose of this study was to improve urban mapping and population density studies by developing YOLOv11 for automatic dwelling detection and counting using aerial photography. The model is intended to deal with house detection issues, including differences in building density, lighting, and roof color in urban environments. YOLOv11 outperforms other models like Faster R-CNN, SSD, and EfficientDet in terms of inference time and detection accuracy, especially when it comes to intricate roof color classification. Faster R-CNN's separate region proposal procedure necessitates a longer inference time, despite its outstanding detection accuracy [27]. Although SSD provides fast single-step object detection, it has trouble identifying intricate and small objects, such as dwellings, in aerial photos [26]. To improve multi-scale detection efficiency, EfficientDet uses a Bi-directional Feature Pyramid Network (BiFPN); however, it still has computational issues on low-power devices [25].

Among the four models, the YOLOv11 model excels in mean Average Precision (mAP), achieving a score of 0.708. This outcome is attained following training on a bespoke dataset comprising 266 images characterized by a significantly uneven distribution of roofing types (black, red, and white roofs). The model converges steadily, displaying considerable fluctuation in class-specific performance; however, it maintains good overall accuracy, particularly for red roofs, with a mAP50-95 of 0.672. The YOLOv11 model demonstrates efficient computational performance, processing each image in 7.6 ms with a batch size of 8 over 100 epochs, and its training duration is quite brief, at 0.946 hours.

Conversely, the Faster R-CNN (Fig. 11 (a)) model produces a decreased mAP of 0.58. Despite leveraging the strength of a pre-trained backbone, it does not achieve the detection accuracy of YOLOv11 on the custom dataset. Although Faster R-CNN is a more established architecture, its performance diminishes when fine-tuned on a specialized dataset, particularly in contrast to YOLOv11's class-specific measures, which demonstrate greater consistency in detection.

The SSD model, illustrated in Fig. 11 (b), was constructed from scratch using a bespoke dataset and attains a mean Average Precision (mAP) of 0.54. Notwithstanding the significant flexibility and capability of SSD, this model's performance is subpar compared to both YOLOv11 and Faster R-CNN, perhaps because of the challenges of training a custom SSD from scratch with a small dataset and without pre-trained weights. This may result in reduced convergence speed and inferior feature extraction compared to the other models.

(a) (b)

Fig. 11. (a) R-CNN Performance (b) SSD Performance.

Ultimately, EfficientDet (Fig. 12 (a)), a cutting-edge object detection model, attains a mean Average Precision (mAP) of 0.62. EfficientDet employs a compound scaling technique to enhance model size and accuracy while preserving efficiency. Its performance exhibits significant promise, especially when combined with effective dataset augmentation and processing methodologies. Nevertheless, the lack of comprehensive data makes direct comparisons difficult without precise measurements.

Fig. 12 (b) shows that the bounding box method spotted various roof objects with varying confidence ratings from overhead residential images. White-Roof, Red-Roof, and Black-Roof were categorized accurately across lighting conditions and architectural variances. Red-Roof was the most commonly detected class, with confidence values from 0.82 to 0.91, including numerous high-confidence detections over 0.85. Despite being rare, white-Roof detections were accurate with confidence values 0.84. The algorithm also detected a Black-Roof occurrence with a 0.58 confidence score, proving it can detect rare roof types. The testing used high-resolution aerial footage in RGB format from a bird's-eye view in natural sunshine. The detecting algorithm placed bounding boxes precisely and had minimum object overlap. The implementation was particularly good at distinguishing roof materials across lighting situations.

(a) (b)

Fig. 12. (a) EfficientDet performance (b) YOLOv11 performance.

TABLE I. MODEL COMPARISON RESULT

Model	mAP50	mAP50-95	Precision	Recall	F1-Score
YOLOv11	0.708	0.53	0.75	0.7	0.72
Faster R-CNN	0.58	0.43	0.64	0.64	0.65
SSD	0.54	0.4	0.62	0.6	0.61
EfficientDet	0.62	0.48	0.71	0.67	0.69

The comparative analysis of object detection models, as presented in Table I, elucidates the advantages and disadvantages of YOLOv11, Faster R-CNN, SSD, and EfficientDet across many assessment criteria. YOLOv11 attains the highest mAP50 score of 0.68, demonstrating its exceptional capability to identify items with a minimum of 50% overlap between predicted and actual bounding boxes. EfficientDet achieves a score of 0.62, demonstrating commendable performance, but Faster R-CNN and SSD exhibit inferior effectiveness with scores of 0.58 and 0.54, respectively. A comparable trend is noted with mAP50-95, a more stringent metric that assesses performance across IoU levels from 50% to 95%. YOLOv11 leads with a score of 0.53, followed by EfficientDet at 0.48, while Faster R-CNN and SSD score 0.43 and 0.40, respectively.

Regarding precision, YOLOv11 surpasses the other models with a score of 0.75, indicating its superior accuracy in accurately recognizing positive detections. EfficientDet ranks second with a precision of 0.71, whilst Faster R-CNN and SSD attain lower precision scores of 0.64 and 0.62, respectively. Regarding recall, which assesses the model's capacity to identify all pertinent objects accurately, YOLOv11 attains a score of 0.70, somewhat surpassing EfficientDet's score of 0.67. Faster R-CNN and SSD achieved scores of 0.64 and 0.60, respectively, signifying a diminished capacity to identify all items within the dataset. The F1-score, a harmonic mean of precision and recall, illustrates YOLOv11's overall efficacy with a peak score of 0.72. EfficientDet attains a score of 0.69, whereas Faster R-CNN and SSD secure F1 scores of 0.65 and 0.61, respectively.

The results collectively demonstrate that YOLOv11 is the most effective model among the four, succeeding in all criteria, whilst EfficientDet is a competitive option with consistent performance. Faster R-CNN and SSD, albeit operational, have comparatively diminished overall efficacy.

V. CONCLUSION

The deployment of YOLOv11 for residential identification and roof color classification has exhibited encouraging outcomes while highlighting opportunities for further enhancement. The model, trained on 266 images featuring 998 roof instances, attained a mean Average Precision (mAP@0.5) of 0.708, exhibiting variable performance across distinct roof hues. Principal discoveries encompass: Red roofs exhibited the highest detection performance, achieving an Average Precision of 0.850 and elevated F1 scores near 0.8. White roofs displayed moderate performance with an Average Precision of 0.671, while black roofs revealed inferior detection rates with an Average Precision of 0.603, suggesting a potential area for enhancement.

With better accuracy and efficiency, YOLOv11 emerged as the top-performing model for residential roof recognition and color classification. Its strength is demonstrated by the confusion matrix, particularly in very accurate and error-free red roof detection. YOLOv11 outperformed EfficientDet, Faster R-CNN, and SSD regarding detection rates and background/roof color differentiation. Notwithstanding difficulties with specific roof types, such as black roofs, it is the most dependable model with a mAP@0.5 score of 0.708 and good F1 scores, making it ideal for real-world applications.

The model exhibited remarkable computing efficiency, with preprocessing times of 0.2ms, inference times of 7.6ms, and postprocessing times of 0.7ms. The training process was completed in 0.946 hours utilizing an NVIDIA GeForce RTX 4060, illustrating the model's viability for real-world application.

The research highlights the benefits and limitations of utilizing YOLOv11 for automatic roof detection and counting. The model exhibits strong performance for particular roof types, notably red roofs. Notwithstanding these achievements, significant limits were noted, including comparatively low confidence scores (0.10-0.20) for specific detections and fluctuations in performance attributable to lighting conditions. The diminished sample size in the test location limited the system's effectiveness in the Black-Roof category. The roof shape will affect the accuracy, as the dataset is bespoke and sourced from Indonesia, resulting in heightened precision when utilized in Indonesian areas. Future research opportunities include expanding the training dataset for underrepresented categories, improving the model to increase confidence ratings, and investigating potential interactions with GIS systems for broader urban planning and analysis applications.

ACKNOWLEDGEMENT

MEK contributed programming code, wrote the manuscript, conducted experiments, and reviewed the paper for this research. AZHP contributed to programming code, preparing datasets, writing manuscripts, and reviewing papers. Kamel contributed to preparing the dataset, conducting experiments, writing manuscripts, and reviewing the paper. HP contributed to supervising the steps of conducting experiments and reviewing the paper. Our sample datasets can be accessed at: https://www.kaggle.com/datasets/drhadypranoto/drone-image-in-indonesia.

REFERENCES

[1] Wilson, K., & Davis, R. "Applications of computer vision in urban planning: Current trends and future perspectives", Cities, 132, 103925, 2023.

[2] Zhang, Q., & Zhang, Y. "Deep learning-based building detection using aerial imagery: A comprehensive review", Remote Sensing, 14(3), 662, 2022.

[3] Li, X., et al. "Automated Building Detection Using Deep Learning: A Review", ISPRS Journal of Photogrammetry and Remote Sensing, 185, 251-272, 2023.

[4] Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors", arXiv preprint arXiv:2207.02696, 2023.

[5] Martinez, J., & Thompson, E. "Building detection in complex urban environments: A deep learning approach", International Journal of Applied Earth Observation and Geoinformation, 117, 103110, 2023.

[6] Chen, R., & Li, X. "Color-based object detection in aerial imagery: Challenges and solutions", IEEE Transactions on Geoscience and Remote Sensing, 61, 1-15, 2023.

[7] Anderson, P., et al. "Color classification in computer vision: Challenges and solutions for architectural applications", Pattern Recognition Letters, 168, 8-19, 2023.

[8] Liu, Y., et al. "Deep learning for roof type classification: A comparative study", Remote Sensing of Environment, 280, 113233, 2023.

[9] Smith, J., & Brown, A. "Urban feature extraction using deep learning: Recent advances and future directions", Urban Remote Sensing Journal, 45(2), 89-112, 2023.

[10] R. Gelar Guntara, "Deteksi Atap Bangunan Berbasis Citra Udara Menggunakan Google Colab dan Algoritma Deep Learning YOLOv7 ", JMASIF, vol. 2, no. 1, pp. 9–18, May 2023.

[11] Salar Ghaffarian, Saman Ghaffarian. "Automatic Building Detection based on Supervised Classification using High Resolution Google Earth Images", ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XL-3, September, 2014.

[12] Vichai Viratkapan., Saprangsit Mruetusatorn. "Classification of House Categories Using Convolutional Neural Networks (CNN) ", 509-514, May, 2022.

[13] Jiaqi Jia., Min Fu., Xuefeng Liu., Bing Zheng. "Underwater Object Detection Based on Improved EfficientDet", Remote Sensing, vol. 14, no. 18, pp. 4487, 2022.

[14] Lin, T. Y., et al. "Microsoft COCO: Common Objects in Context", vol. 8693, pp. 740-755, 2014.

[15] Zhao, Z. Q., et al. "Object Detection with Deep Learning: A Review", IEEE Transactions on Neural Networks and Learning Systems, pp. 1-21, January 2019.

[16] Garcia, M., et al. "Interpreting Confusion Matrices in Urban Remote Sensing", Remote Sensing of Environment, 2023.

[17] Park, S., et al. "Confidence Threshold Optimization in Object Detection", IEEE Transactions on Image Processing, 2023.

[18] Rodriguez, A., et al. "Class-wise Performance Analysis in Aerial Image Detection", ISPRS Journal, 2023.

[19] Wu, X., et al. "Aerial Image Building Detection: A Comprehensive Review", Remote Sensing, 2023.

[20] Redmon, J., Divvala, S., Girshick, R., dan Farhadi, A. "You Only Look Once: Unified, Real-Time Object Detection", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 779–788, 2016.

[21] Redmon, J., & Farhadi, A. "YOLOv3: An incremental improvement", arXiv preprint arXiv:1804.02767, 2018.

[22] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. "YOLOv4: Optimal speed and accuracy of object detection", arXiv preprint arXiv:2004.10934, 2020.

[23] Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., & Yeh, I. H. "CSPNet: A new backbone that can enhance learning capability of CNN", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020.

[24] Jocher, G., et al. "Ultralytics YOLOv8: A state-of-the-art model for object detection and image segmentation", 2023.

[25] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. "SSD: Single Shot MultiBox Detector", European Conference on Computer Vision (ECCV), 2016.

[26] Tan, M., Pang, R., & Le, Q. V. "EfficientDet: Scalable and Efficient Object Detection", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[27] Ren, S., He, K., Girshick, R., & Sun, J. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", Advances in Neural Information Processing Systems (NeurIPS), 2015.

[28] Aggarwal, C. C. "Neural Networks and Deep Learning: A Textbook", Springer, 2018.

[29] Raschka, S., & Mirjalili, V. "Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2", Packt Publishing, 2019.

[30] Padilla, R., et al. "A Survey on Performance Metrics for Object-Detection Algorithms", International Conference on Systems, Signals and Image Processing, 2020.

[31] DataCamp, "YOLO Object Detection Explained", DataCamp, Dec. 22, 2022.

[32] A. Rohan, M. Rabah, and S.-H. Kim. "Convolutional Neural Network-Based Real-Time Object Detection and Tracking for Parrot AR Drone 2", vol. 7, pp. 69575-69584, 2019.

[33] M. Tan, R. Pang, and Q. V. Le. "EfficientDet: Scalable and Efficient Object Detection", pp. 10781-10790, 2020.

[34] Z. Deng, H. Sun, S. Zhou, and H. Zou. "Multi-scale object detection in remote sensing imagery with convolutional neural networks", vol. 15, no. 5, pp. 749-753, 2018.

A Supervised Learning-Based Classification Technique for Precise Identification of Monkeypox Using Skin Imaging

Vandana[1], Chetna Sharma[2]*, Yonis Gulzar[3]*, Mohammad Shuaib Mir[4]
Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India[1, 2]
Mukand Lal National College, Yamunanagar, Haryana, India[1]
Department of Management Information Systems, College of Business Administration, King Faisal University, Al-Ahsa 31982, Saudi Arabia[3, 4]

***Abstract*—The monkeypox epidemic has spread to nearly every nation. Governments implement several strict policies, to stop the virus that causes monkeypox. For effective handling and treatment, early identification and diagnosis of monkeypox using digital skin lesion images is critical, and this work employed deep learning architectures to achieve this goal. This article presents a supervised learning-based classification method designed for the precise identification of monkeypox cases. The analysis was conducted using an open-source dataset from Kaggle, consisting of digital images of monkeypox, which were processed using advanced image processing and deep learning techniques. The data was categorized based on findings related and unrelated to monkeypox. A deep neural network with 50 layers and up to 35 folds was utilized to identify regions of interest, which could be indicative of characteristics relevant to computer-assisted medical diagnosis and enable us to solve image processing and natural language processing tasks with high accuracy. In terms of performance, the proposed method achieved an accuracy of 96% during cross-validation classification testing. This outcome demonstrates the potential for computer-assisted diagnosis as a supplementary tool for medical professionals. Amid the monkeypox outbreak, this method offers a technical and objective assessment of patients' skin conditions, thereby simplifying the diagnostic process for specialists.**

***Keywords*—Deep learning; monkeypox; medical image processing; image classification; cross validation**

I. Introduction

The infectious illness known as "mpox," or monkeypox, has spread quickly across the globe and is characterized by fever, muscle pains, and skin lesions that resemble boils. In response, the World Health Organization (WHO) raised the warning level to its maximum in July by classifying it as a public health emergency of international concern (PHEIC) [1]. When a tourist returned from Nigeria, where the disease is widespread, the epidemic began in the UK. However, there have previously been isolated occurrences of this kind, and those outbreaks ended swiftly. This time, the disease began to spread from the first cluster of patients and appeared in various European, Australian, and American nations [2].

Although African epidemiologists have been warning for several years that patterns of transmission seem to be shifting in endemic nations, scientists are still unsure of the exact reason for the virus's rapid expansion. As of December 19, 110 countries have reported more than 83,000 cases, with 66 fatalities, according to W.H.O. Encouragingly, the weekly total of new cases worldwide has dropped by 49.3%; this past week saw 265 new cases globally, compared to 523 from December 5–11 [1–3]. Currently, the Americas are considered to be at greater risk than Africa. The Americas (90.5%) and Europe (4.9%) accounted for the majority of cases recorded over the previous four weeks [3]. Fig. 1 shows the DNA structure of monkeypox, while Fig. 2 highlights the 10 most affected countries worldwide.

In some regions of the globe, the pandemic may have decreased due to the availability of the currently preventable JYNNEOS monkeypox vaccine, but nations below the poverty line have been left behind. Given that monkeypox may mutate and become harder to manage in the future, experts argue that it is imperative to administer the vaccination in a fair and equitable manner [4].

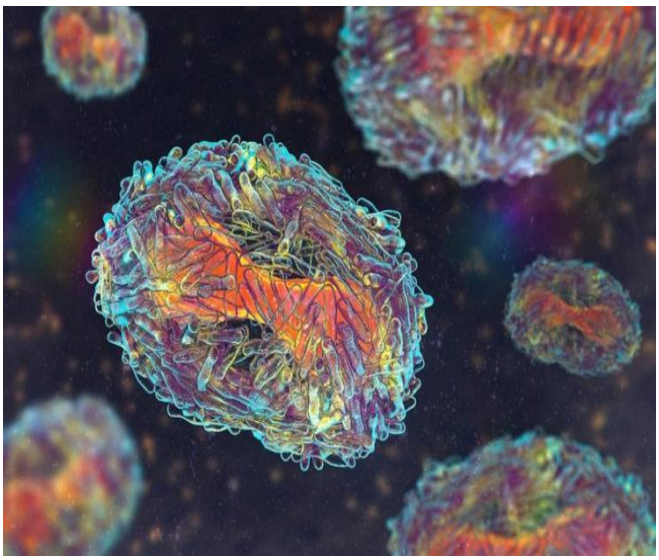


Fig. 1. Monkeypox DNA image [1].

*Corresponding Author, Email ID: chetna.kaushal@chitkara.edu.in (C.S.), ygulzar@kfu.edu.sa (Y.G.)

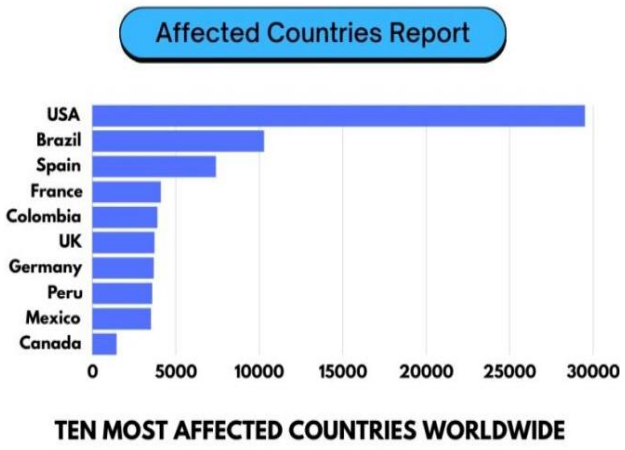


Fig. 2. Affected countries of monkeypox worldwide [1].

Artificial Intelligence (AI) and Deep Learning have demonstrated remarkable results across various domains, transforming the way tasks are approached. In agriculture, AI-powered models have optimized crop management [5,6], disease detection[7], and yield prediction [8,9]. In education, adaptive learning platforms are tailoring educational content to individual needs, enhancing student engagement and success. In finance, AI-based algorithms are improving fraud detection, risk assessment, and investment strategies. In healthcare, AI and Deep Learning have revolutionized diagnostics, treatment plans, and patient management through the precise analysis of complex medical data [10–13].

Specifically in healthcare, AI has significantly enhanced decision-making by enabling more accurate diagnosis and personalized treatments through the interpretation of medical images, patient records, and genetic data [10–13]. This progress can extend to infectious diseases, where a deep learning could be employed for precise identification of Monkeypox using skin imaging. By leveraging AI's ability to analyze and differentiate subtle features in skin lesions, the technique could improve early detection and reduce misdiagnosis. This would facilitate timely interventions, reducing the spread of the disease and enhancing patient outcomes, demonstrating AI's critical role in addressing emerging public health challenges.

In this study, a computer-aided diagnostic (CAD) system using deep learning and artificial intelligence (AI) techniques is proposed for potential pulmonary follow-up due to its accuracy and ability to optimize reaction times. AI has played a significant role in medical diagnosis [14, 15]. Expert interpretation is facilitated by the technique's automated and objective assessment of pulmonary follow-ups, which is enabled by artificial neural network designs. Since artificial neural networks (deep learning) have proven to be successful, AI has also been applied to medical image processing [16, 17]. The development of trustworthy and accurate medical diagnostics has garnered significant attention in recent years. Artificial neurons are layered and interconnected to transfer signals in neural network techniques, with intermediate layers being hidden. These networks, along with more advanced learning processes, form the foundation of deep learning, which generates categorization approaches that are both optimal and precise [18, 19]. Main aim of the research is to develop a technique which can classify image efficiently without leaving a single skin lesion in images which can help doctors to treat well and detect the disease at its initial stage.

This paper's primary research contributions can be summed up as follows:

- Our suggested technique uses images with 224 × 224 x 3 the RGB spectrum dimensions and those without monkeypox lesions.
- To implement pre-processing steps using Resnet 50 to generate the feature vector(1 × 1 × 2048) then the its subsets are defined using its increase and decrease in fold values which solve image processing and natural language processing tasks with high accuracy.
- To Classify various methods likes SVN, LR, KNN, NB, NC evaluates the accuracy and precision value depending upon various folds

II. Related Work

Monkeypox disease is often mistaken for other illnesses, leading to misdiagnosis and inappropriate treatment. Early diagnosis and treatment of this contagious disease are crucial. Detecting monkeypox typically requires expert interpretation and clinical examination, which can delay the treatment process. AI-based detection can assist in the early identification of this disease. There are limited studies in the literature on this topic, which are discussed in detail below, along with Table I.

Dipanjali Kundu et al. [20] presented a secure Federated Learning and deep learning-based framework for monkeypox virus detection using skin lesion images. The framework aims to improve classification performance while maintaining data confidentiality. The CycleGAN generator augments training and test data, and synthetic images are divided into four groups for local techniques. ViT-B32 outperforms other classifiers with an impressive accuracy of 97.90. The approach ensures user data privacy and effectively performs categorization tasks, making it relevant in medical contexts with limited datasets and data privacy concerns. Sitaula et al. [21] evaluated 13 pre-trained deep learning techniques (including VGG-16, InceptionV3, Xception, MobileNet, EfficientNet, etc.) for monkeypox detection using the publicly available Monkeypox virus image dataset. They trained the techniques on the ImageNet dataset. The study included 1,754 images, consisting of 329 chickenpox, 286 measles, 587 monkeypox, and 552 normal images. They developed a technique using the Keras library in Python. The proposed ensemble learning technique outperformed the 13 deep learning techniques, achieving an accuracy of 87.1% (precision: 85.4%, recall: 85.4%, F1-score: 85.4%). Xception was the second-best technique with an accuracy of 86.51%, while precision, recall, and F1-score were 85%. Alakus and Baykara [22] classified monkeypox disease and warts based on their DNA sequences. They employed various DNA mapping techniques and deep learning. The study used 110 genome sequences, consisting of 55 monkeypox virus and 55 human papillomavirus sequences. To address the data imbalance, they used the zero-padding

technique. Five DNA mapping techniques achieved an average classification accuracy of 96.08%, with the integer DNA matching technique achieving the highest accuracy at 99.5%. This demonstrates the successful detection of monkeypox and warts through DNA mapping and classification. Ali et al. [23] addressed the challenge of early clinical diagnosis of monkeypox, similar to chickenpox and measles, through computer-aided detection. They created a dataset of skin lesion images from various sources, consisting of 228 images. Three classification techniques were used: VGG16, ResNet50, and InceptionV3. ResNet50 achieved the highest accuracy (82.96% ± 4.57), VGG16 performed competitively (81.48% ± 6.87), and InceptionV3 had the lowest accuracy (74.07% ± 3.78). A community technique using majority voting outperformed ResNet50 and was integrated into a prototype web application. Haque et al. [24] aimed to classify human monkeypox disease from images using a pre-trained deep learning technique. The study utilized VGG-19, Xception, DenseNet121, MobileNetV2, and EfficientNetB3 deep learning techniques for classification. A uniform approach was applied to customize all pre-trained techniques. To enhance the network's focus on more pertinent feature maps, a convolutional block attention module was incorporated. The initial preparation of the MSLD involved resizing the images to a resolution of 224x224x3 for training purposes. In the research, many hyperparameters were used to maximize the effectiveness of the strategies. The design that included Xception, CBAM, and thick layers performed better than other techniques in the findings, obtaining a validation accuracy of 83.89% in the classification of human monkeypox and other illnesses. Sahin et al. [25] developed a mobile app using deep learning to detect monkeypox from video footage captured on mobile devices. They used the MSLD dataset and deep transfer learning with Matlab. Transfer learning is a machine learning technique where a model trained on one task is adapted to improve performance on a related but different task, often with fewer data [28] . MobileNetV2 (91.11%) and EfficientNetB0 (91.11%) achieved the best results in 60 epochs. MobileNetV2, with precision (90%), recall (90%), F1-score, and accuracy (91.11%), outperformed other techniques and was integrated into an Android mobile app, allowing easy pre-screening for monkeypox. Ahsan et al. [26] developed an AI-driven decision support system using CNNs. Their study used a dataset of 572 images (monkeypox and normal). They employed twelve different CNN techniques for classification, with MobileNetV2 achieving the highest accuracy (98.25%), precision (96.55%), specificity (100%), and F1-score (98.25%). The study also highlighted that MobileNetV2 is suitable for mobile-based monkeypox testing due to its smaller technique size. Ahsan et al.[27] created a dataset of patient images infected with monkeypox. The researchers aimed to detect monkeypox virus in patients using a modified pre-trained VGG16 technique. The study collected a total of 1,915 images, including monkeypox, chickenpox, measles, and normal images, as well as augmented versions. Two separate studies were conducted, one with a small dataset and the other with a medium-sized dataset. In the first study, using a small dataset, the VGG16 technique achieved training and testing accuracy rates of 97% and 83%, respectively. In the second study, with a medium-sized dataset, the technique achieved accuracy rates of 88% in training and 78% in testing. The proposed technique's predictions were validated through cross-validation by medical professionals. The study suggests that this technique could be used to develop a mobile-based diagnostic tool.

TABLE I. LITERATURE REVIEW OF RELATED WORK

Reference	Data type	Technique	Accuracy	Research Gap
Kundu et al.[20]	a total of 381 images of monkeypox, 102 images of chickenpox, 110 images of measles, and 293 images of normal skin.	federated learning (FL) with Cycle GANS and deep learning-based techniques such as MobileNetV2, Vision Transformer (ViT), and ResNet50 for the classification	Proposed technique accuracy 97.90%	Limited datasets and data privacy
Sitaula et al. [21]	Monkeypox virus dataset images (329 chickenpox, 286 measles, 587 monkeypox, and 552 normal images)	13 pre-trained deep learning techniques (including VGG-16, InceptionV3, Xception, MobileNet, Efficient-Net, etc.)	Proposed technique accuracy 87.13%.	Less accuracy rate
Alakus et al.[22]	55 monkeypox virus and 55 human papilloma virus sequences	DNA mapping techniques	Proposed technique accuracy 99.5%.	Very small Dataset used
Ali et al. [23]	228 images from open source data set	VGG16, ResNet50, and InceptionV3	ResNet50 accuracy (highest accuracy) (82.96% ± 4.57), VGG16 accuracy (81.48% ± 6.87), InceptionV3 accuracy (lowest accuracy) (74.07% ± 3.78).	Need of improved segmentation technique
Haque et al.[24]	MSLD	VGG-19, Xception, DenseNet121, MobileNetV2, and EfficientNetB3	Accuracy: 83.89%	Less accuracy rate
Sahin et al. [25]	MSLD 112 images	MobileNetv2, EfficientNetb0	MobileNetv2 (91.11%) and EfficientNetb0 (91.11%) achieved the best results in 60 epochs.	Small dataset used and less augmentation used
Ahsan et al.[26]	Dataset of 572 images (monkeypox and normal).	AI-driven decision support system using CNNs (MobileNetv2)	MobileNetV2 highest accuracy (98.25%), precision (96.55%), specificity (100%), and F1-score (98.25%)	MobileNetV2 is suitable for mobile-based monkeypox testing due to its smaller technique size
Ahsan et al.[27]	Open source Dataset 1915 images	VGG16	Proposed accuracy 88%	Need to develop a mobile-based diagnostic tool

III. MATERIAL AND TECHNIQUES

A. Deep Learning Classification Techniques

In this study, CNN-based deep learning techniques, namely VGG16, ResNet50, EfficientNetB3, Xception, and InceptionResNetV2, were used to perform image-based classification of monkeypox disease. Brief descriptions of these techniques are provided below:

VGG16 is a deep learning technique developed at the University of Oxford. VGG stands for Visual Geometry Group, and 16 refers to the number of layers in the technique [20]. VGG16 is a CNN technique commonly used for image classification tasks. Convolutional, fully linked, and pooling layers make up the approach. The pooling layers highlight key characteristics and condense the bulk of the data [20]. Because of its depth, VGG16 is a large approach with a significant number of learnable parameters. Large datasets and more challenging picture classification tasks are often better served by it. VGG16 became well-known, especially when it performed well on the ImageNet dataset. It is known for emphasizing the fundamental structure of convolutional networks and the depth of its layers, which has influenced the development of related techniques that use weighted convolutional layers and depth.

In deep learning, one popular Convolutional Neural Network (CNN) approach is called ResNet50. Microsoft Research first introduced the ResNet (Residual Network) approach in 2015 with the express purpose of resolving issues related to depth in deep networks. ResNet50 is a 50-layer deep network that uses a unique building component known as a residual block. These blocks introduce skip links in the network's transitions, which attempt to mitigate the issue of gradient vanishing that arises in deeper networks. In comparison to earlier techniques, the residual blocks allow information to move across the network more quickly and smoothly [29]. ResNet50 has pooling layers, activation functions, and convolutional layers—basic CNN building blocks. Additionally, it has a global average pooling layer in the center of the network that uses smaller feature maps to summarize data.

A version known as ResNet blends the ResNet and Inception architectures [30]. While the ResNet design makes use of residual connections to solve gradient vanishing in deep networks, the Inception architecture is composed of convolutional layers with filters of varying sizes. ResNet seeks to integrate these two topologies to make training deeper and more complicated networks easier. While the ResNet blocks maintain information flow by using connections to prevent gradient vanishing, the Inception blocks combine convolutional layers with filters of varying sizes to capture a broad variety of characteristics. Extreme Inception, a term that is shortened to "Xception," is a deep learning approach [31]. It is based on the CNN architecture, more precisely on an Inception approach variant. Improving the Inception network's processing efficiency is the primary objective of Xception. It deviates from the conventional CNN technique by optimizing the convolution operations in the Inception blocks to achieve this. Depth-wise separable convolution is the method used to accomplish this improvement. Two steps of convolution operations are carried out by depth-wise separable convolution. A point-wise convolution layer is used in the first stage to capture relationships between several channels and modify the input data's dimension. The second step is a depth-wise convolution layer, which improves computing efficiency by processing each input data channel independently. The CNN technique known as EfficientNetB3 is used in deep learning [32]. It belongs to the family of EfficientNets, and EfficientNetB3 is a scalable and effective technique. Compound scaling is a technique that EfficientNet uses to automatically scale deep learning approaches to different sizes. The purpose of EfficientNetB3 is to be applied to larger and more intricate datasets. The method improves efficiency on smaller datasets by combining depth features with scalability. To handle input images, EfficientNetB3 uses a variety of convolutional layers, activation functions, and pooling layers [32, 33].

B. Data Set

The study used a dataset that included skin lesion images classified into two groups: those with monkeypox and those without (chickenpox and measles). The University of Dhaka research team in Bangladesh [33] contributed to the dataset, which was augmented to contain approximately 3,192 images with dimensions of 224 × 224 x 3 in RGB. There were 228 images in the original dataset, of which 102 belonged to the monkeypox class and 126 to the other class. This dataset has a comparatively modest amount of images, especially when considering the deep learning environment. The only restriction was that more images were obtained by using image enhancement. Through picture augmentation, the dataset was increased by a factor of fourteen. After augmentation, there were 1,428 images in the monkeypox class and 1,764 images in the 'others' class. An unequal distribution of images was observed between the 'others' and monkeypox classes. To address this imbalance, data augmentation techniques were applied to create more samples for the monkeypox class. By adding these extra examples, the dataset's classes were distributed more evenly, as shown in Fig. 3, which displays 1,764 images in each class.

To further enhance the dataset's durability, the researchers used stratified sampling to divide it into seven folds. This technique ensured that the monkeypox class and the other classes were represented proportionally in each fold. The stratified split of the dataset into folds allowed for a more accurate evaluation and validation of the technique's effectiveness by taking into account any variances in the distribution of classes within the dataset. The technique's ability to generalize was assessed by applying cross-validation on multiple data subsets and evaluating its performance on each fold. This experiment used two subsets of the dataset, namely the training set and the testing set. The testing set, consisting of 252 images, provided a more thorough insight into the technique's ability to categorize monkeypox images. The training set consisted of 1,512 enhanced images used for training the classification system. To develop a reliable and efficient classification technique, it is essential to carefully analyze the segmentation of this dataset [34, 35].

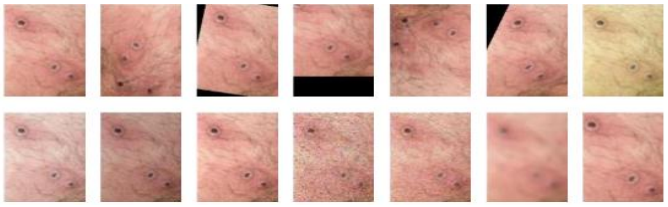


Fig. 3. Sample augmented monkeypox images from dataset.

IV. PROPOSED TECHNIQUES

The approach is based on the categorization of the images in Fig. 4; the technique diagram is comprehensive and includes the dataset. High reliability is achieved by identifying the optimal settings for executing the procedure. Python, an open-source programming language, was used to develop the method. The approach is divided into many steps; however, the preprocessing step is crucial for successful classification. In particular, the first phase focuses on the initial analysis, as the experiment was conducted using the dataset to enable the CNN to extract features from the images related to the identification of monkeypox.

Fig. 4. Technique of the proposed technique.

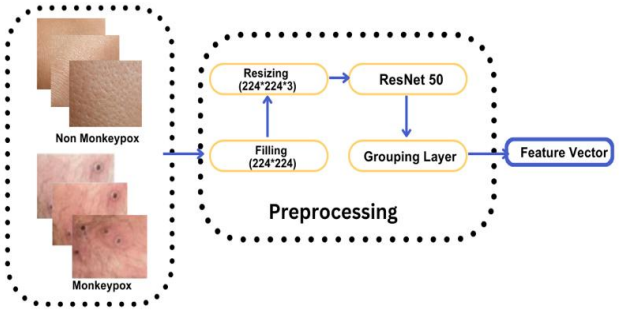


Fig. 5. An illustration of the first phase and the pre-processing steps to generate the feature vector.

The schematic diagram of the first and second phases is shown in Fig. 5. The residual neural network ResNet50, which was used, contains 50 hidden layers [36]. This artificial neural network consists of three layers of artificial neurons: one input, fifty intermediate (hidden), and one output. The process of creating a matrix from the feature vectors is based on residual learning. By putting neural networks through performance tests with partial and random changes to their connections, residual learning increases the technique's accuracy and enhances its ability to solve complex problems with greater reliability.

Here is a list of the four preliminary processing phases shown in Fig. 4: All input images are scaled to 224 × 224 pixels, and if the appropriate dimensions are not met, zeros are added to the perimeter of the original image. The pixel values in each layer are adjusted from 0 to 255. A new matrix with dimensions of 224 × 224 × 3 is then created by converting the images to three layers (the RGB spectrum). In the case of single-layer image graphs, the same layer is repeated in each channel. After completing these procedures, the images are fed into the ResNet50 network. The convolution process and the network's fundamental architecture are illustrated in Fig. 4. Up until the middle pooling layer is reached, the input image is subjected to successive convolution and sampling processes. The grouping layer features are extracted as a vector with dimensions of 1 × 1 × 2048 in this layer. This vector, generated for each image in the dataset, contains general attributes of the images retrieved by the approach, such as saturation, brightness, and intensity.

The final step in producing the feature vector is carried out using machine learning techniques; in other words, classifier techniques are applied to categorize the dataset. The ResNet50 neural network acts as an image feature extractor; by utilizing the convolutional base of this network, the most significant and distinctive features of the images are obtained. This is achieved not through machine learning, but rather through deep learning, using artificial networks that extract the most relevant features to enable optimal classification.

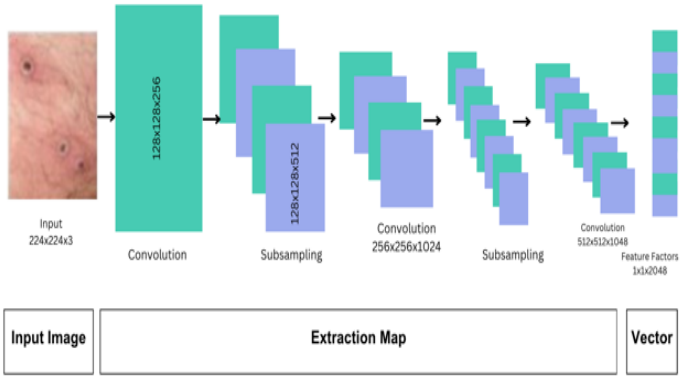


Fig. 6. The convolutional technique and the ResNet50 network's easy layout.

Even though a single CNN [37] may be used for both feature extraction and classification, there are several advantages to utilizing different techniques for each task. As a result, the automated classification approach used consisted of three stages: pre-processing, feature extraction (via CNN), and the classification system as shown in above Fig. 6.

V. CONFIGURING THE EXPERIMENT

The cross-validation classification scenario served as the foundation for the suggested technique [38]. There were 57 homogeneous and symmetric images with and without evidence of monkeypox; 114 images were divided using a cross-validation technique. When there are insufficient images in the collection, this type of categorization is suitable [39]. Fig. 3 depicts the cross-validation scenario using automated categorization algorithms. Due to their effectiveness in binary classification—where a set's components are divided into two categories according to a classification rule—the classification techniques used are well-established in the field [40]. Support Vector Machines (SVM) provides an optimal separation that reduces classification risk and increases the margin while also reducing error [41,42]. Logistic regression (LR) is used to predict the probability that a categorical dependent variable will be dichotomous, or divided into two groups [43]. Nearest Neighbors (KNN) uses discrete sample classification to forecast and estimate future values based on proximity, identifying similar data points. Additionally, Naive-Bayes' (NB) success is due to feature independence, which allows the determination of the likelihood that a test case has a particular feature value. Finally, the Centroid-based classifier (NC), derived from distances from the center, takes into account similarity with each class's centroid. The centroid is a vector representing the average frequencies of all terms among the members of a specific class.

The validation was carried out in three sections, with divisions of 15, 25, and 35 folds. Accuracy (A) and Precision (P), two evaluation measures, were used to assess the technique's performance with the dataset. These measures are based on the four elements listed below [44]:

- True Positive (TP): occurs when there is a match between the predicted class of the technique and the actual class of the dataset (finding).
- True Negative (TN): occurs when there is a match between the dataset's actual class (no finding) and the one predicted by the technique (no finding).
- False Positive (FP): occurs when there is a mismatch between the dataset's actual class (no finding) and the one predicted by the algorithm (finding).
- False Negative (FN): occurs when there is a mismatch between the dataset's actual class (a finding) and the one predicted by the technique (no finding).

Accuracy refers to the percentage of all predictions that were made correctly. Precision, on the other hand, indicates the accuracy of the positive predictions, i.e., the proportion of true positive values to all predicted positive values [45].

Metrics are calculated as follows [44]:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

VI. RESULTS

Four different classifier techniques were used as the learning strategy in the cross-validation classification scenario. The dataset was split into three studies, as previously mentioned: 15 folds, 25 folds, and 35 folds. The dataset was randomly divided into K = 15, 25, and 35 segments as shown above. K − 1 segments were used for training the technique, while the remaining portion was used for performance evaluation. After dividing the dataset into K parts at random, this process was repeated K times to obtain K procedures and assessment outcomes. The final average performance was determined once all evaluations were completed.

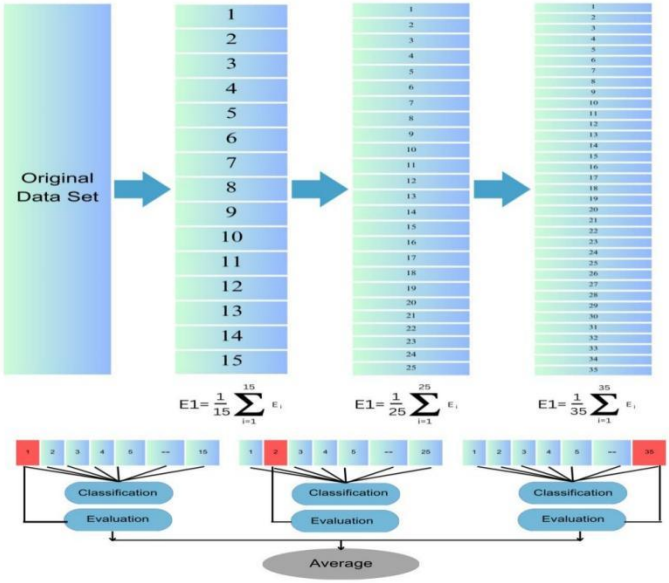


Fig. 7. The K-folds cross-validation strategy uses a single classifier with K values of 15, 25, and 35.

The technique was then retrained for the subsequent divisions. The technique's performance (E) is assessed as follows:

$$E = \frac{1}{k}\sum_{i=1}^{k} Ei \quad (3)$$

The process for 15 divisions, as well as the divisions of the folds in the dataset, is shown in Fig. 7. The 25 and 35 categories follow the same procedure to determine the performance (E). To compare the outcomes of several predictive classification processes—that is, the five learning strategies used to categorize the images—cross-validation was applied. The results of the various classification techniques are shown in Table II, along with a comparison of the divisional processes and evaluation criteria, which make it possible to identify the most accurate classifier. It is evident that as the number of folds increases, the KNN classifier (nearest neighbors) produces the highest values in the accuracy measure [46].

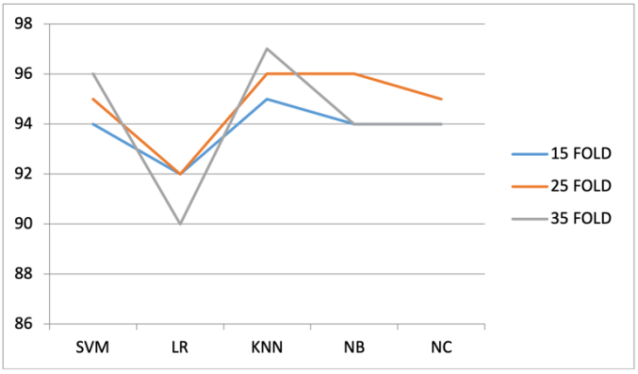


Fig. 8. The precision assessment metric's learning techniques in the three division experiments.

TABLE II. OUTCOMES OF A CROSS-VALIDATION CLASSIFICATION SCENARIO

Fold	15		25		35	
Classification	Accuracy	Precision	Accuracy	Precision	Accuracy	Precision
SVM	92.7	94.2	93.3	95.2	93.8	95.8
LR	90.3	92.5	90.1	92.6	89.4	90.4
KNN	94.2	95.4	95.5	96.4	95.2	97.7
NB	93.0	94.7	93.7	96.2	93.7	94.9
NC	92.7	94.2	93.2	95.7	93.7	94.9

VII. DISCUSSION

This shows that the categorization percentage is 94% in 15 divisions, 95% in 25 divisions, and 96% in 35 divisions. The accuracy indicates the proportion of the positive class predicted by the approach and the actual positive class in the dataset, as well as the quality of the classification scenario. In small datasets, the classifier with the best values often performs optimally. Similarly, the accuracy provides a 94% confidence level in the KNN classifier's quality compared to other learning techniques, provided the classes of the variables in the dataset are balanced. The accuracy measure and various folds used in the learning techniques are shown in Fig. 7. According to the suggested approach, the KNN classifier, the reported metrics, and the cross-validation classification scenario with 15, 25, and 35 folds together represent the best automated classification technique for the dataset, including discoveries related to monkeypox. The characteristic vectors of the images are extracted using the ResNet50 networks, and when paired with the approach, the results yield values that are competitive with the state of the art. As demonstrated, this strategy enables us to identify regions of interest within the images. While the current study focused on binary classification, it has been shown that the system can be expanded to categorize different diseases. The deep neural network's feature extraction technique includes the extraction of regions of interest (ROI) as shown in Fig. 8. Once combined, these characteristics are sent to the final classification technique, which utilizes them to provide a prediction.

VIII. CONCLUSION

This article's approach, which utilizes a dataset of patients with monkeypox, has demonstrated success. It gives the binary labels of finding lesions with monkeypox and non-monkeypox findings. Classification results with 96% precision were achieved using 35 divisions of folds using KNN classification, which improves the image processing quality by identifying even small lesions in skin images. The accuracy of these results not only reflects the quality of the classification scenario and the percentage of the positive class predicted by the technique compared to the actual positive class in the dataset but also improves as the dataset is split into several parts, indicating an increase in the technique's performance. The ResNet50 technique of CNN is a crucial component of the technique, as it enables comparison of the features of images with and without lesions by applying deep learning to identify monkeypox findings. This is accomplished through classifiers that are assessed using metrics that measure how effectively the technique functions. Based on the conclusions acquired, this article can now report on categorizing medical images from a dataset related to monkeypox. The technique provides an automatic and objective estimation of the classification of monkeypox findings, facilitating expert interpretation during the pandemic. This makes the values obtained from the scenario and binary classifier in the quantitative analysis of the imaging study more reliable, leading to a more accurate diagnosis. In future, ascertaining the viability and use of AI-based monkeypox detection systems in clinical settings, researchers can conduct empirical assessments of these systems. It will take patients and healthcare professionals working together to assemble large datasets. Moreover, other designs might be studied to tackle the difficulty mentioned before. Researchers can increase the precision of AI-based monkeypox diagnosis by including other data sources, such as clinical symptoms, laboratory test findings, and patient history. As a result of these findings, the categorization approach serves as a supplementary tool in medical diagnostics.

ACKNOWLEDGMENT

This work was supported by the Deanship of Scientific Research under the Vice Presidency for Graduate Studies and Scientific Research of King Faisal University in Saudi Arabia under Project KFU241811.

REFERENCES

[1] WHO Director-General Declares Mpox Outbreak a Public Health Emergency of International Concern Available online: https://www.who.int/news/item/14-08-2024-who-director-general-declares-mpox-outbreak-a-public-health-emergency-of-international-concern (accessed on 19 September 2024).

[2] Monkeypox Global Trends Available online: https://archive.cdc.gov/#/details?url=https://www.cdc.gov/poxvirus/mpox/response/2022/world-map.html (accessed on 19 September 2024).

[3] Stilpeanu, R.I.; Stercu, A.M.; Stancu, A.L.; Tanca, A.; Bucur, O. Monkeypox: A Global Health Emergency. Front Microbiol 2023, 14, 1094794, doi:10.3389/FMICB.2023.1094794/BIBTEX.

[4] Deputy, N.P.; Deckert, J.; Chard, A.N.; Sandberg, N.; Moulia, D.L.; Barkley, E.; Dalton, A.F.; Sweet, C.; Cohn, A.C.; Little, D.R.; et al. Vaccine Effectiveness of JYNNEOS against Mpox Disease in the United States. New England Journal of Medicine 2023, 388, 2434–2443, doi:10.1056/NEJMOA2215201/SUPPL_FILE/NEJMOA2215201_DATA-SHARING.PDF.

[5] Malik, I.; Ahmed, M.; Gulzar, Y.; Baba, S.H.; Mir, M.S.; Soomro, A.B.; Sultan, A.; Elwasila, O. Estimation of the Extent of the Vulnerability of Agriculture to Climate Change Using Analytical and Deep-Learning

Methods: A Case Study in Jammu, Kashmir, and Ladakh. Sustainability 2023, Vol. 15, Page 11465 2023, 15, 11465, doi:10.3390/SU151411465.

[6] Gulzar, Y. Enhancing Soybean Classification with Modified Inception Model: A Transfer Learning Approach. Emirates Journal of Food and Agriculture 36: 1-9 2024, 36, 1–9, doi:10.3897/EJFA.2024.122928.

[7] Alkanan, M.; Gulzar, Y. Enhanced Corn Seed Disease Classification: Leveraging MobileNetV2 with Feature Augmentation and Transfer Learning. Front Appl Math Stat 2024, 9, 1320177, doi:10.3389/FAMS.2023.1320177.

[8] Jabbari, A.; Humayed, A.; Reegu, F.A.; Uddin, M.; Gulzar, Y.; Majid, M. Smart Farming Revolution: Farmer's Perception and Adoption of Smart IoT Technologies for Crop Health Monitoring and Yield Prediction in Jizan, Saudi Arabia. Sustainability 2023, Vol. 15, Page 14541 2023, 15, 14541, doi:10.3390/SU151914541.

[9] Amri, E.; Gulzar, Y.; Yeafi, A.; Jendoubi, S.; Dhawi, F.; Mir, M.S. Advancing Automatic Plant Classification System in Saudi Arabia: Introducing a Novel Dataset and Ensemble Deep Learning Approach. Model Earth Syst Environ 2024, 10, 2693–2709, doi:10.1007/s40808-023-01918-9.

[10] Mehmood, A.; Gulzar, Y.; Ilyas, Q.M.; Jabbari, A.; Ahmad, M.; Iqbal, S. SBXception: A Shallower and Broader Xception Architecture for Efficient Classification of Skin Lesions. Cancers 2023, Vol. 15, Page 3604 2023, 15, 3604, doi:10.3390/CANCERS15143604.

[11] Khan, F.; Ayoub, S.; Gulzar, Y.; Majid, M.; Reegu, F.A.; Mir, M.S.; Soomro, A.B.; Elwasila, O. MRI-Based Effective Ensemble Frameworks for Predicting Human Brain Tumor. Journal of Imaging 2023, Vol. 9, Page 163 2023, 9, 163, doi:10.3390/JIMAGING9080163.

[12] Majid, M.; Gulzar, Y.; Ayoub, S.; Khan, F.; Ree

[13] gu, F.A.; Mir, M.S.; Jaziri, W.; Soomro, A.B. Enhanced Transfer Learning Strategies for Effective Kidney Tumor Classification with CT Imaging. International Journal of Advanced Computer Science and Applications 2023, 14, 2023, doi:10.14569/IJACSA.2023.0140847.

[14] Majid, M.; Gulzar, Y.; Ayoub, S.; Khan, F.; Reegu, F.A.; Mir, M.S.; Jaziri, W.; Soomro, A.B. Using Ensemble Learning and Advanced Data Mining Techniques to Improve the Diagnosis of Chronic Kidney Disease. International Journal of Advanced Computer Science and Applications 2023, 14, doi:10.14569/IJACSA.2023.0141050.

[15] Asif, S.; Zhao, M.; Li, Y.; Tang, F.; Ur Rehman Khan, S.; Zhu, Y. AI-Based Approaches for the Diagnosis of Mpox: Challenges and Future Prospects. Archives of Computational Methods in Engineering 2024, 31, 3585–3617, doi:10.1007/S11831-024-10091-W/METRICS.

[16] Suzuki, K. Overview of Deep Learning in Medical Imaging. Radiol Phys Technol 2017, 10, 257–273, doi:10.1007/S12194-017-0406-5/METRICS.

[17] Khan, A.; Sohail, A.; Zahoora, U.; Qureshi, A.S. A Survey of the Recent Architectures of Deep Convolutional Neural Networks. Artificial Intelligence Review 2020 53:8 2020, 53, 5455–5516, doi:10.1007/S10462-020-09825-6.

[18] Tayir, T.; Li, L. Unsupervised Multimodal Machine Translation for Low-Resource Distant Language Pairs. ACM Transactions on Asian and Low-Resource Language Information Processing 2024, 23, doi:10.1145/3652161.

[19] Rafi, T.H.; Shubair, R.M.; Farhan, F.; Hoque, M.Z.; Quayyum, F.M. Recent Advances in Computer-Aided Medical Diagnosis Using Machine Learning Algorithms with Optimization Techniques. IEEE Access 2021, 9, 137847–137868, doi:10.1109/ACCESS.2021.3108892.

[20] Castiglioni, I.; Rundo, L.; Codari, M.; Di Leo, G.; Salvatore, C.; Interlenghi, M.; Gallivanone, F.; Cozzi, A.; D'Amico, N.C.; Sardanelli, F. AI Applications to Medical Images: From Machine Learning to Deep Learning. Physica Medica 2021, 83, 9–24, doi:10.1016/J.EJMP.2021.02.006.

[21] Kundu, D.; Rahman, M.M.; Rahman, A.; Das, D.; Siddiqi, U.R.; Alam, M.G.R.; Dey, S.K.; Muhammad, G.; Ali, Z. Federated Deep Learning for Monkeypox Disease Detection on GAN-Augmented Dataset. IEEE Access 2024, 12, 32819–32829, doi:10.1109/ACCESS.2024.3370838.

[22] Sitaula, C.; Shahi, T.B. Monkeypox Virus Detection Using Pre-Trained Deep Learning-Based Approaches. J Med Syst 2022, 46, 1–9, doi:10.1007/S10916-022-01868-2/FIGURES/5.

[23] Alakus, T.B.; Baykara, M. Comparison of Monkeypox and Wart DNA Sequences with Deep Learning Model. Applied Sciences 2022, Vol. 12, Page 10216 2022, 12, 10216, doi:10.3390/APP122010216.

[24] Ali, E.; Sheikh, A.; Owais, R.; Shaikh, A.; Naeem, U. Comprehensive Overview of Human Monkeypox: Epidemiology, Clinical Features, Pathogenesis, Diagnosis and Prevention. Annals of Medicine & Surgery 2023, 85, 2767–2773, doi:10.1097/MS9.0000000000000763.

[25] Haque, M.E.; Ahmed, M.R.; Nila, R.S.; Islam, S. Human Monkeypox Disease Detection Using Deep Learning and Attention Mechanisms. Proceedings of 2022 25th International Conference on Computer and Information Technology, ICCIT 2022 2022, 1069–1073, doi:10.1109/ICCIT57492.2022.10055870.

[26] Sahin, V.H.; Oztel, I.; Yolcu Oztel, G. Human Monkeypox Classification from Skin Lesion Images with Deep Pre-Trained Network Using Mobile Application. J Med Syst 2022, 46, 1–10, doi:10.1007/S10916-022-01863-7/TABLES/6.

[27] Ahsan, M.M.; Ali, M.S.; Hassan, M.M.; Abdullah, T.A.; Gupta, K.D.; Bagci, U.; Kaushal, C.; Soliman, N.F. Monkeypox Diagnosis with Interpretable Deep Learning. IEEE Access 2023, 11, 81965–81980, doi:10.1109/ACCESS.2023.3300793.

[28] Ahsan, M.M.; Uddin, M.R.; Farjana, M.; Sakib, A.N.; Momin, K. Al; Luna, S.A. Image Data Collection and Implementation of Deep Learning-Based Model in Detecting Monkeypox Disease Using Modified VGG16. arXiv:2206.01862 2022.

[29] Gulzar, Y. Fruit Image Classification Model Based on MobileNetV2 with Deep Transfer Learning Technique. Sustainability 2023, 15, 1906.

[30] Anand, V.; Gupta, S.; Koundal, D.; Mahajan, S.; Pandit, A.K.; Zaguia, A. Deep Learning Based Automated Diagnosis of Skin Diseases Using Dermoscopy. Computers, Materials & Continua 2021, 71, 3145–3160, doi:10.32604/CMC.2022.022788.

[31] Neshat, M.; Ahmed, M.; Askari, H.; Thilakaratne, M.; Mirjalili, S. Hybrid Inception Architecture with Residual Connection: Fine-Tuned Inception-ResNet Deep Learning Model for Lung Inflammation Diagnosis from Chest Radiographs. Procedia Comput Sci 2024, 235, 1841–1850, doi:10.1016/J.PROCS.2024.04.175.

[32] Rahul; Sharma, A.; Gupta, S.; Anand, V. Proposed Convolution Architecture for Monkeypox Detection Using Dermoscopy Images. 2023 3rd International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies, ICAECT 2023 2023, doi:10.1109/ICAECT57570.2023.10118296.

[33] Nuipian, W.; Meesad, P.; Kanjanawattana, S. A Comparative ResNet-50, InceptionV3 and EfficientNetB3 with Retinal Disease. ACM International Conference Proceeding Series 2023, 283–287, doi:10.1145/3639233.3639337.

[34] Ahsan, M.M.; Uddin, M.R.; Luna, S.A. Monkeypox Image Data Collection. arXiv:2206.01774 2022.

[35] Ayoub, S.; Gulzar, Y.; Reegu, F.A.; Turaev, S. Generating Image Captions Using Bahdanau Attention Mechanism and Transfer Learning. Symmetry (Basel) 2022, 14, 2681.

[36] Ayoub, S.; Gulzar, Y.; Rustamov, J.; Jabbari, A.; Reegu, F.A.; Turaev, S. Adversarial Approaches to Tackle Imbalanced Data in Machine Learning. Sustainability 2023, Vol. 15, Page 7097 2023, 15, 7097, doi:10.3390/SU15097097.

[37] Chen, Y.; Wang, L.; Ding, B.; Shi, J.; Wen, T.; Huang, J.; Ye, Y. Automated Alzheimer's Disease Classification Using Deep Learning Models with Soft-NMS and Improved ResNet50 Integration. J Radiat Res Appl Sci 2024, 17, 100782, doi:10.1016/J.JRRAS.2023.100782.

[38] Shivadekar, S.; Hundekari, S.; Kataria, B.; Wanjale, K.; Balpande, V.P.; Suryawanshi, R. Deep Learning Based Image Classification of Lungs Radiography for Detecting COVID-19 Using a Deep CNN and ResNet 50. International Journal of Intelligent Systems and Applications in Engineering 2009, 11, 241–250.

[39] Suresh Kumar, K.; Radha Mani, A.S.; Ananth Kumar, T.; Jalili, A.; Gheisari, M.; Malik, Y.; Chen, H.C.; Jahangir Moshayedi, A. Sentiment Analysis of Short Texts Using SVMs and VSMs-Based Multiclass Semantic Classification. Applied Artificial Intelligence 2024, 38, doi:10.1080/08839514.2024.2321555.

[40] Pavlou, M.; Omar, R.Z.; Ambler, G. Penalized Regression Methods With Modified Cross-Validation and Bootstrap Tuning Produce Better

Prediction Models. Biometrical Journal 2024, 66, e202300245, doi:10.1002/BIMJ.202300245.

[41] Zhang, B.; Zhang, H.; Zhen, T.; Ji, B.; Xie, L.; Yan, Y.; Yin, E. A Two-Stage Real-Time Gesture Recognition Framework for UAV Control. IEEE Sens J 2024, doi:10.1109/JSEN.2024.3413787.

[42] Mahmood, O.A.; Sulaiman, S.O.; Al-Jumeily, D. Forecasting for Haditha Reservoir Inflow in the West of Iraq Using Support Vector Machine (SVM). PLoS One 2024, 19, e0308266, doi:10.1371/JOURNAL.PONE.0308266.

[43] Khan, F.; Gulzar, Y.; Ayoub, S.; Majid, M.; Mir, M.S.; Soomro, A.B. Least Square-Support Vector Machine Based Brain Tumor Classification System with Multi Model Texture Features. Front Appl Math Stat 2023, 9, 1324054, doi:10.3389/FAMS.2023.1324054.

[44] Geng, Y.; Li, Q.; Yang, G.; Qiu, W. Logistic Regression. Practical Machine Learning Illustrated with KNIME 2024, 99–132, doi:10.1007/978-981-97-3954-7_4.

[45] Erickson, B.J.; Kitamura, F. Magician's Corner: 9. Performance Metrics for Machine Learning Models. Radiol Artif Intell 2021, 3, doi:10.1148/RYAI.2021200126/ASSET/IMAGES/LARGE/RYAI.2021200126.FIG6.JPEG.

[46] Vandana; Kaushal, C. Analysis of the Monkeypox Outbreak Using CNN Model: A Systematic Review. 2023 4th IEEE Global Conference for Advancement in Technology, GCAT 2023 2023, doi:10.1109/GCAT59970.2023.10353352.

[47] Agarwal, M.; Gill, K.S.; Chauhan, R.; Kapruwan, A.; Banerjee, D. Classification of Network Security Attack Using KNN (K-Nearest Neighbour) and Comparison of Different Attacks through Different Machine Learning Techniques. 2024 3rd International Conference for Innovation in Technology, INOCON 2024 2024, doi:10.1109/INOCON60754.2024.10512250.

Classifying Weed Development Stages Using Deep Learning Methods

Classifying Weed Development Stages with DenseNET, Xception, SqueezeNET, GoogleNET, EfficientNET CNN Models Using ROI Images

Yasin ÇİÇEK[1], Eyyüp GÜLBANDILAR[2], Kadir ÇIRAY[3], Ahmet ULUDAĞ[4]
Dept. of Computer Engineering, Eskişehir Osmangazi University, Eskişehir, Turkey[1, 2]
Sinanpaşa Vocational School, Afyon Kocatepe University, Afyonkarahisar, Turkey[3]
Dept. of Plant Protection, Çanakkale Onsekiz Mart University, Çanakkale, Turkey[4]

***Abstract*—The control of harmful weeds holds a significant place in the cultivation of agricultural products. A crucial criterion in this control process is identifying the development stages of the weeds. The technique to be used is determined based on the weed's growth stage. This study addresses the application of deep learning methods in classifying growth stages using images of various weed species to predict their development periods. Four different weed species, obtained from seeds collected in Turkey-Afyonkarahisar-Sinanpaşa Plain, were used in the study. The images were captured with a Nikon D7000 camera equipped with three different lenses, and the ROI extraction was performed using Lifex software. Using these ROI images, deep learning models such as DenseNet, EfficientNet, GoogleNet, Xception, and SqueezeNet were evaluated. Performance metrics including accuracy, F1 score, precision, and recall were employed. In the 4-class dataset with ROI annotations, DenseNet and Xception achieved an accuracy of 86.57%, while EfficientNet demonstrated the highest performance with an accuracy of 89.55%. Following the initial tests, it was concluded that classes 3 and 4 exhibited extreme similarity caused most of the prediction errors. Merging the said classes significantly increased the accuracy and F1 scores across all models. In image classification tests, SqueezeNet and GoogleNet demonstrated the shortest processing times. However, while EfficientNet lagged slightly behind these models in terms of speed, it exhibited superior accuracy. In conclusion, although the use of ROI improved classification performance, class merging strategies resulted in a more significant performance enhancement.**

***Keywords*—Deep learning; weed development stages; classification; DenseNET; Xception; SqueezeNET; GoogleNET; EfficientNET; ROI**

I. INTRODUCTION

One of the foremost factors affecting yield in agriculture is weeds [1]. If weeds are not controlled in the crop plant, this effect can be more significant. Integrated weed management (IWM) is a system that provides economically bearable, biologically effective and environmentally sound weed control using possible techniques in a sustainable manner. Among the tools IWM relies on is the critical period for weed control (CPWC): a duration during the crop's growth cycle when weed control is essential [2]. In many crops, CPWC for higher yields (more acceptable yield loss) starts from the crop emerging, which requires earlier detection of weeds to achieve effective weed control [3,4].

Identifying weeds at an early stage can be challenging because they may closely resemble crop plants. This task becomes even more difficult for farmers who must search for and identify weeds across the entire field, as it is an incredibly monotonous activity, leading to decreased performance and efficiency [5]. To mitigate this challenge, modern computer methods are employed. Artificial intelligence techniques used in computer-based imaging make it easier to detect weeds, thus facilitating the goal of product identification. Indeed, information technologies are being utilized in various fields of agriculture [6].

In image-based classifications, CNN and RNN deep learning models are more frequently used [7]. In agriculture, deep learning is applied in areas such as product diseases, pests, spraying, and the classification of crops and weeds [8].

A. Related Work

Espejo Garcia and colleagues developed a classification model using machine learning techniques such as Support Vector Machines (SVM), XGBoost, and logistic regression, along with convolutional networks like Xception, Inception-ResNet, VGNets, MobileNet, and DenseNet, applied to two crops and two weed species. In the dataset composed of photographs taken under natural light conditions with RGB cameras, DenseNet and SVM achieved an F1 Score of 99.29% [9].

Sunil and colleagues utilized deep learning models such as Xception, DenseNet, MobileNetV3Large, EfficientNet, and ConvNeXt to classify six weed species and eight crop types. They reported that, except for DenseNet, the other four models demonstrated strong performance, with the macro average F1 scores ranging between 0.85 and 0.87 and the weighted average F1 scores ranging between 0.87 and 0.88 [10].

Pandey and colleagues achieved classification in datasets containing maize and radish plants along with weeds using deep learning models such as InceptionV3, Xception, ResNet152V2, VGG16, and their proposed CNN model. They reported an accuracy of 97.66% for maize and 98.75% for radish [11].

Garibaldi-Marquez and colleagues conducted a classification study on maize and weeds. In their study, they used both regular images and ROI images, employing deep learning models such as ResNet101, VGG16, Xception, and MobileNetV2. Among the classification algorithms, Xception provided the best accuracy result of 97.43% [12].

Trong and colleagues conducted a study using datasets of plant seedlings and weeds from Chonnam National University. They employed five different deep learning methods, including NasNet and ResNet, to classify weeds. They achieved an accuracy of 97.31% for plant seedlings and 98.77% for weeds [13]. In another study involving ResNet, VGG16, and Xception, Peteinatos and colleagues used 93 000 images obtained with an RGB camera of 12 different weed species and applied deep learning models, reporting an accuracy range of 77% to 98% [14].

Chen and colleagues utilized a dataset comprising 5,187 images of 15 weed classes found in cotton fields. They classified these weeds using deep learning methods such as ResNeXt, Xception, and MnasNet. They achieved high accuracy, with F1 scores exceeding 95% [15].

One of the commonly used models for weed classification is the YOLO architecture. Gao and colleagues used the YOLOv3 architecture to train a model for identifying field bindweed and sugar beet plants. They based their training on 452 field images and generated 2,271 synthetic images. Their tests with 100 field plant images demonstrated that the average precision metric could reach 0.829 [16].

Ahmad and colleagues used a YOLOv3-based object detection model to identify four different weed species within maize and soybean fields. They employed pre-trained deep learning models, including VGG16, ResNet50, and InceptionV3. They reported that VGG16 achieved the highest accuracy at 98.90% and an F1 score of 99 [17].

Zhang and colleagues utilized the YOLOv3 and YOLOv3-tiny architectures to detect weeds in wheat fields. Based on the images obtained from UAVs, they indicated that YOLOv3-tiny is more suitable for mobile devices [18].

Sharpe and colleagues utilized the YOLOv3-tiny deep learning architecture to detect goosegrass in strawberries and tomatoes. They reported the following metrics for whole plants: precision = 0.93; recall = 0.88; F1-score = 0.90; accuracy = 0.82. For leaf blades, the metrics were precision = 0.39; recall = 0.55; F1-score = 0.46; accuracy = 0.30 [19]. Additionally, Osorio and colleagues used SVM and R-CNN architectures alongside YOLOv3 to detect weeds in lettuce. They found that R-CNN outperformed the other two methods, achieving an F1-score of 94% [20].

There are numerous publications related to deep learning in the agricultural sector, especially regarding the classification of weeds. Alongside the abundance of these publications, review studies that analyze, group, and evaluate these works have also started to appear in the literature [8, 21, 22].

The aim of current work is to predict the developmental stages of the plant or classify them based on the stages. Using the obtained images of weed species, the developmental stages of the weeds are going to be predicted employing deep learning methods. In addition to the plant images, predictions are also going to be made using ROI analysis on the same leaf images to enhance recognition.

II. Material and Method

The study consists of three main stages. The first stage is cultivating and photographing the weeds, the second stage is the extracting the ROI images and radiomics from the photographs, and the last stage is preparing the data set, developing the models and evaluating the performances.

A. Dataset

Seeds of Lamb's Quarters (Chenopodium album), Jerusalem Oak Goosefoot (Dysphania botrys), Prickly Lettuce (Lactuca serriola), and Sow Thistle (sonchus oleraceus), which are four common weed species in sugarbeet fields in Turkey-Afyonkarahisar-Sinanpaşa plains, were collected, dried, and threshed. Once the seeds were sown under suitable conditions, data collection began immediately as the weeds emerged, creating a comprehensive dataset. The development of the weeds was meticulously tracked twice a week using a Nikon D7000 camera equipped with three different lenses: 'Nikon 50mm 1.8 mm', 'Tokina 11-16 mm', and 'Sigma 105 mm'.

The weeds were grown in 16 different units ($2^4 = 16$), each with varying conditions of soil, irrigation, light, and fertilizer. The plants were photographed in various growth stages. In addition to the images, the leaves were analyzed using Region of Interest (ROI) techniques. Images were tested both with original and ROI marked images.

In line with this objective, ROI images of the leaves were extracted from the photos in the dataset using the Lifex program [23]. This resulted in a total of 448 photos. Five different deep learning models were applied to this dataset.

Based on the confusion matrix results, the accuracy rate was lower for the leaves in the 4th class due to their resemblance to those in the 3rd class. Consequently, the data for the 3rd and 4th classes were combined to create a new dataset. The same deep learning methods were applied to this new dataset as well.

Finally, a dataset was created using the raw, unmarked versions of the same images, and the above-mentioned deep learning models were applied again in both 3-class and 4-class formats.

B. Deep Learning Concepts

A short summary of Artifical Intelligence (AI) concepts to clarify where this study falls among AI concepts would be in order.

Machine learning is the learning stage of AI. Algorithms are developed and used on datasets to learn to perform certain tasks. It encompasses a variety of methods such as Supervised Learning, Unsupervised Learning, Reinforcement Learning and Deep Learning [24].

1) Deep learning: Deep learning is a subset of machine learning that involves training neural networks with many layers (hence "deep") to process and learn from large amounts of data. It focuses on learning representations of data through

these multiple layers. Deep learning encompasses Neural Networks [25].

2) Neural networks: Artificial Neural Networks (ANN) are machine learning algorithms consisting of a multitude of nodes where each node is connected to other nodes. Nodes communicate with each other in a predetermined pattern and their weights are adjusted every stage of the learning process depending on the success of the desired result. It is much like a biological neural system with neurons. It has been named "neural" due to this similarity [24].

3) Image classification models: There is a wide variety of ANN models especially designed for image classifications such as Resnet, GoogleNet, Xception, DenseNet, EfficientNet, SqueezeNet etc.

a) DenseNet201: The DenseNet network can be considered an extension of the ResNet model, which is a significant milestone in deep networks. The difference between the DenseNet network and the ResNet model is that DenseNet combines layers by concatenation rather than summation.

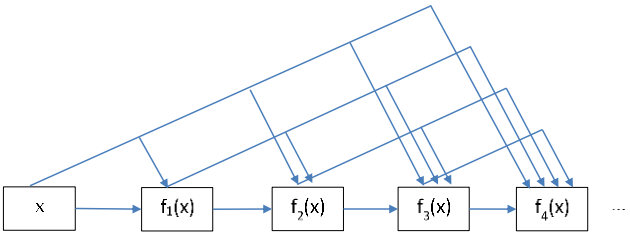


Fig. 1. DenseNET connections in DenseNET model.

Fig. 1 shows the progressively dense layer connections of DenseNet. The name "Dense" derives from these increasingly dense connections. The groups of layers where such connections are used are called dense blocks. However, the sequential addition of dense layers significantly increases the number of channels, i.e., the data pathways between layers, to astronomical levels, necessitating the need to control complexity. Transition layers are used to reduce this increased number of channels to one [26, 27].

In the DenseNet network, data is trained using RGB format images with a resolution of 224x224x3 pixels.

b) XCeption (Depthwise Separable Convolution): Xception aims to achieve the same or better performance with fewer parameters by using depthwise separation. The model is designed to have each convolution layer operate through two pathways: depthwise convolution and pointwise convolution. In the depthwise convolution pathway, a single input slice produces a single output slice, mapping only the spatial (height-width) dimensions. In the pointwise convolution pathway, a 1x1 convolution maps the color channels. This separation of the spatial and color channels results in a more lightweight model. The network comprises a total of 71 layers. In the XCeption network, data is trained using RGB format images with a resolution of 299x299x3 pixels [28, 29].

c) GoogleNet (Inception Networks): The idea behind this model was to effectively combine the methods of Network in Network and Repeated Blocks that preceded it. It was based on the concept of using combinations of convolution kernels of different sizes, ranging from 1x1 to 11x11, used in previous models. The basic structure of the inception blocks, which combine convolution kernels of different sizes, is shown in Fig. 2.

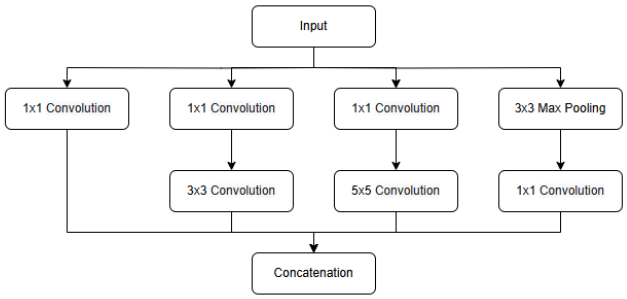


Fig. 2. Inception block structure.

Different sized convolution kernels can be thought of as filters capturing features of different sizes. Since 1x1, 3x3, and 5x5 convolution kernels capture different features in an image, inception blocks provide an efficient method for capturing features of various sizes.

The number of channels in the inception blocks was determined based on numerous experiments conducted on the ImageNET dataset. The overall structure of GoogleNET was created by serially connecting meticulously designed inception blocks to other blocks. The first version of the GoogleNET model contained 22 layers, but later versions may have more layers.

In the GoogleNet network, data is trained using RGB format images with a resolution of 224x224x3 pixels. The 224x224 image size and 3 color channels are commonly used input dimensions in the ImageNet dataset, which is widely utilized by GoogleNET and many other image classification models [27, 30].

d) SqueezeNet: SqueezeNet is a deep neural network designed to achieve high accuracy with significantly fewer parameters compared to existing models like AlexNet. The authors aimed to create a smaller, more efficient network that could be easily deployed on devices with limited memory and computational power. SqueezeNet achieves AlexNet-level accuracy on the ImageNet dataset while being 50 times smaller and requiring less computational resources. The model uses a unique architecture called the Fire Module, which consists of a squeeze layer (with 1x1 filters) followed by an expand layer (with a mix of 1x1 and 3x3 filters). The network has 18 layers and processes images with a resolution of 256x256 pixels. This makes it suitable for applications in resource-constrained environments such as mobile devices and embedded systems2 [31].

e) EfficientNet: EfficientNet is a new approach to scaling convolutional neural networks (CNNs) for better performance than previous models. Traditional methods scaled CNNs by increasing depth, width, or resolution individually, but EfficientNet proposes a balanced scaling method that uniformly scales all three dimensions using a compound coefficient. This

method, called compound scaling, allows EfficientNet to achieve state-of-the-art accuracy with fewer parameters and computational resources. The EfficientNet family includes models from B0 to B7, with B0 being the smallest and B7 being the largest. The smallest model, B0, has 24 layers and processes images with a resolution of 224x224 pixels, while the largest model, B7, has 264 layers and processes images with a resolution of 600x600 pixels. EfficientNet models outperform previous CNN architectures on various benchmarks, including ImageNet and CIFAR-1003 [32].

The selected neural network models were trained using the following parameters: the Adam optimizer, with an initial learning rate set at 0.0001, a mini-batch size of 25, and validation data provided by an augmentedImageDatastore object (referred to as augimdsValidation). The validation frequency configured at every 5 iterations. The training progress was visualized using plots, the performance metric used was accuracy, and the verbosity setting was disabled to reduce output during training. The number of epochs was maintained at the default value of 30 epochs. To ensure a fair and consistent comparison, all models were evaluated under these identical training conditions.

These settings were applied uniformly across all models to ensure the reliability and validity of the comparisons.

C. ROI (Region of Interest)

ROI is typically used to define the boundaries of a significant area within an image. Similar to its use in medical imaging to measure tumor boundaries and sizes, ROI allows for focused analysis on specific areas of interest rather than the entire image, thereby facilitating the extraction of various numerical features. Calculations such as mean and maximum values can be performed based on the defined ROI boundaries [33, 34].

In this study, ROIs of the different shape leaf images corresponding to the same plant and time intervals were used to determine the developmental stages of the leaves.

D. Why These Particular Models

Neural networks have their own advantages and weaknesses compared to each other. There are three main characteristics to consider when selecting neural networks: accuracy, speed, and size. The choice of network for use depends on the priority of these three characteristics in the intended application area. Classification accuracy and speed measurements are often conducted using the ImageNet database. However, these results do not always achieve the same success for different tasks and datasets.

It is not possible to find very precise criteria for selecting CNN models. The number of layers in the model, structural complexity, and the number of operations (flops) do not show a linear relationship with the final accuracy rates. There are models that provide high accuracy with low complexity and number of operations. Additionally, parameter settings do not yield the same level of results in every model. Model complexity can only serve as an indicator for the computer's memory usage level [35]. Considering the context in which the model is intended to be trained, it is not possible to predict which model will yield better results. Therefore five different models corresponding to promising accuracy and speed and size values. In addition, five different models that have been selected are widely used with mostly successful results.

E. Data Augmentation and Preprocessing

Deep learning methods enable the solving of various problems based on data obtained from information systems. Regardless of innovations in the design and training processes of the deep learning model, the data always affects the results. Training with a small amount of data with low representational power leads to poor generalization, while training with a large amount of data with high representational power shows higher generalization performance even with less complex algorithms [36].

The main goal of data augmentation is to enhance the robustness and accuracy of deep learning models and to allow them to perform well on small, weakly representative datasets. Effective data augmentation strategies also reduce the requirements related to model complexity, enabling high generalization performance with simpler deep learning architectures [37].

One of the biggest issues encountered in data training is overfitting. Overfitting occurs when the model makes correct predictions for training data but fails to do so for new data. On the other hand, underfitting is another error type that arises when the model fails to identify meaningful relationships between input and output data. Data augmentation helps prevent overfitting, underfitting, and the model memorizing the exact details of the training images [38]. Fig. 3 shows an example of data augmentation.

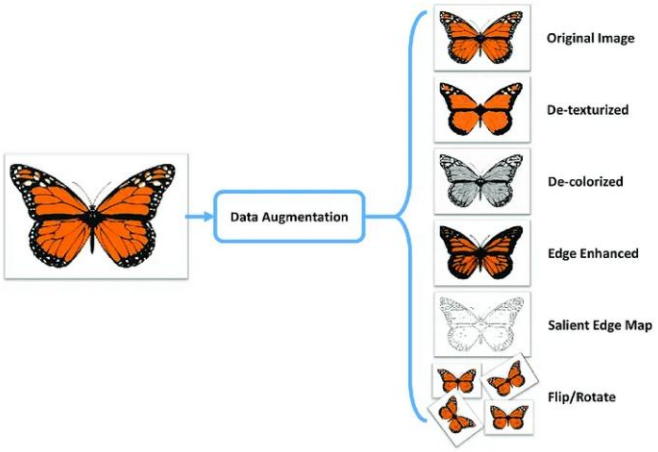


Fig. 3. Data augmentation samples [39].

The obtained datasets were used with the Matlab program to apply deep learning models such as DenseNet, EfficientNet, GoogleNet, Xception, and SqueezeNet. Four different datasets were run with the models, including 2, 3, and 4-class datasets, both with and without ROI markings. The datasets were divided into 70% for training, 15% for testing, and 15% for validation.

All modeling was conducted using an Nvidia RTX 4060 graphics card with CUDA cores. The 3-class dataset contained 128, 128, and 192 photos, respectively, while the 4-class dataset contained 128, 128, 128, and 64 photos, respectively. The batch

size was set to 25, and the validation frequency was set to 5. All models were trained in 30 epochs. In all models, the results of Precision, Recall, F1 score, and Accuracy metrics [40], along with confusion matrices and accuracy graphs that included the runtime, were recorded.

III. RESULTS

The average performance of the four-class, five-class models obtained using ROIs for each class is shown in Table I, and the confusion matrices of the models are depicted in Fig. 4.

TABLE I. THE PERFORMANCE OF 4-CLASS OUTPUT VARIABLES WITH ROI MARKED IMAGES FOR EACH MODEL

	Precision	Recall	F1 Score	Acc	Time
Densenet	0,8224	0,8480	0,8280	0,8657	12:47
Efficient	0,8724	0,8807	0,8759	**0,8955**	06:07
Googlenet	0,7711	0,7906	0,7731	0,8209	**03:29**
Xception	0,8224	0,8660	0,8306	0,8657	21:02
Squeezenet	0,6684	0,6875	0,6610	0,7313	**03:27**

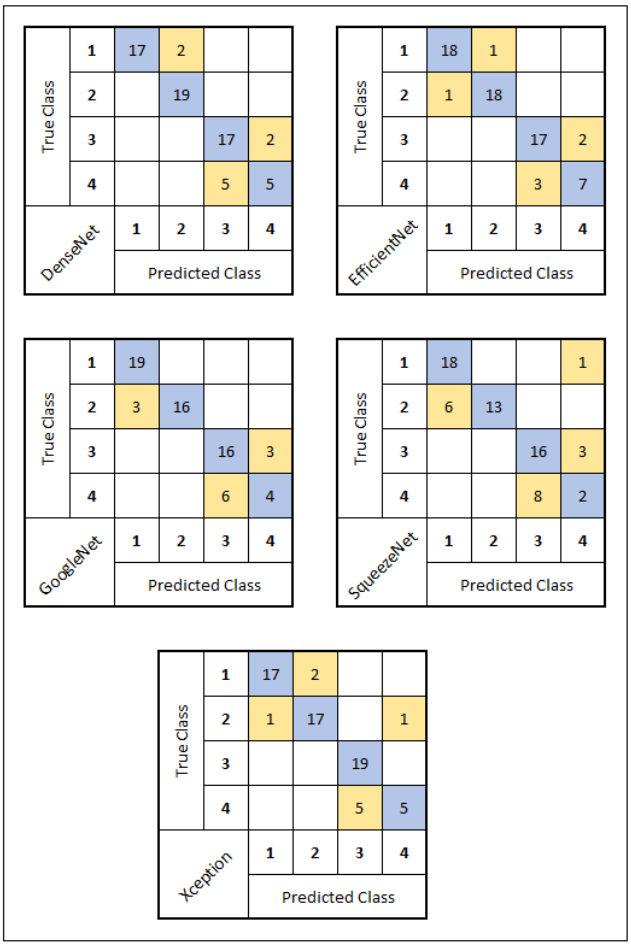


Fig. 4. The confusion matrices of the 4-class output variables with ROI marked images.

From Table I and Fig. 4, it is observed that accuracy measures belong to the DenseNet, EfficientNet, GoogleNet, and Xception models are quite high (>0.8). The accuracy rate of the SqueezeNet model is around 0.73. A similar situation is seen in the F1 score, precision, and recall values. These results indicate that the performance of the first four models is significantly higher than that of the SqueezeNet model. However, when looking at the runtime, it is observed that the runtimes of the SqueezeNet and GoogleNet models are significantly shorter than others. According to these results, the EfficientNet model is seen to have a high superiority both in terms of performance and runtime.

TABLE II. THE PERFORMANCE OF 3-CLASS OUTPUT VARIABLES WITH ROI MARKED IMAGES FOR EACH MODEL

	Precision	Recall	F1 Score	Acc	Time
Densenet	0,9298	0,9328	0,9296	0,9403	19:06
Efficient	0,9474	0,9545	0,9470	0,9552	06:06
Googlenet	0,9649	0,9649	0,9649	**0,9701**	03:42
Xception	0,9298	0,9360	0,9317	0,9403	13:33
Squeezenet	0,9123	0,9168	0,9142	0,9254	**01:47**

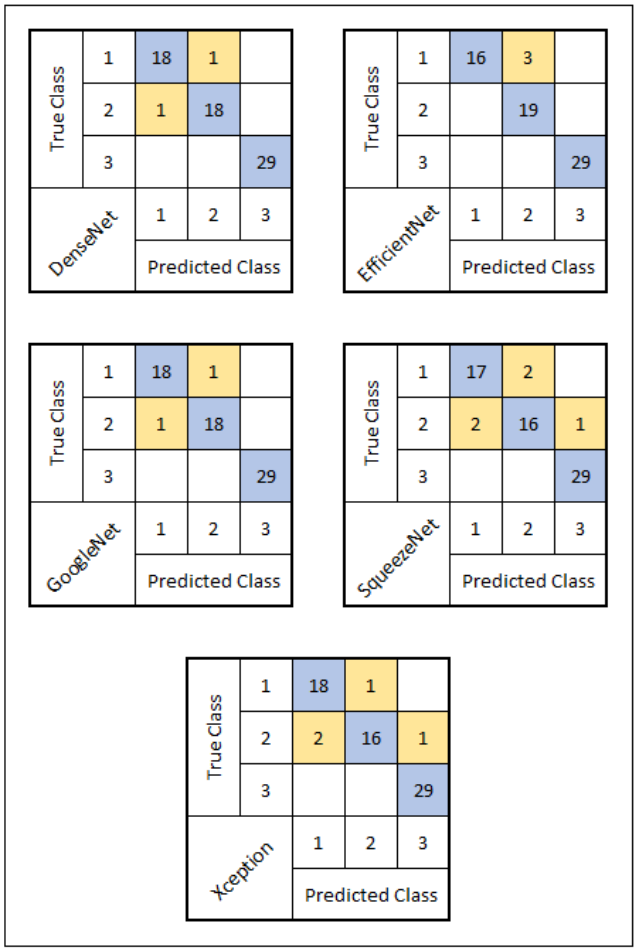


Fig. 5. The confusion matrices of the 3-class output variables with ROI marked images.

From Fig. 4, the prediction error rate between the 3rd and 4th classes is relatively high. This error is most likely due to the similar leaf sizes of the plants in these classes. To improve the accuracy rates of the developed model performances, the dataset was updated by merging the 3rd and 4th classes. Upon examining Table II and Fig. 5, it is observed that all models show high performance measurement metrics (>0.9). However, when looking at the runtime, it is noted that the runtime of the SqueezeNet models is quite short. According to these results, the GoogleNet model has a high superiority in terms of both performance and runtime.

TABLE III. THE PERFORMANCE OF THE 4-CLASS OUTPUT VARIABLES WITH ORIGINAL IMAGES WITHOUT ROI MARKINGS FOR EACH MODEL

	Precision	Recall	F1 Score	Acc	Time
Densenet	0,9487	0,9406	0,9440	**0,9552**	11:22
Efficient	0,8842	0,8748	0,8770	0,8955	04:29
Googlenet	0,7276	0,7272	0,7158	0,7313	02:19
Xception	0,9342	0,9260	0,9287	0,9254	13:22
Squeezenet	0,9605	0,9611	0,9605	**0,9552**	**00:53**

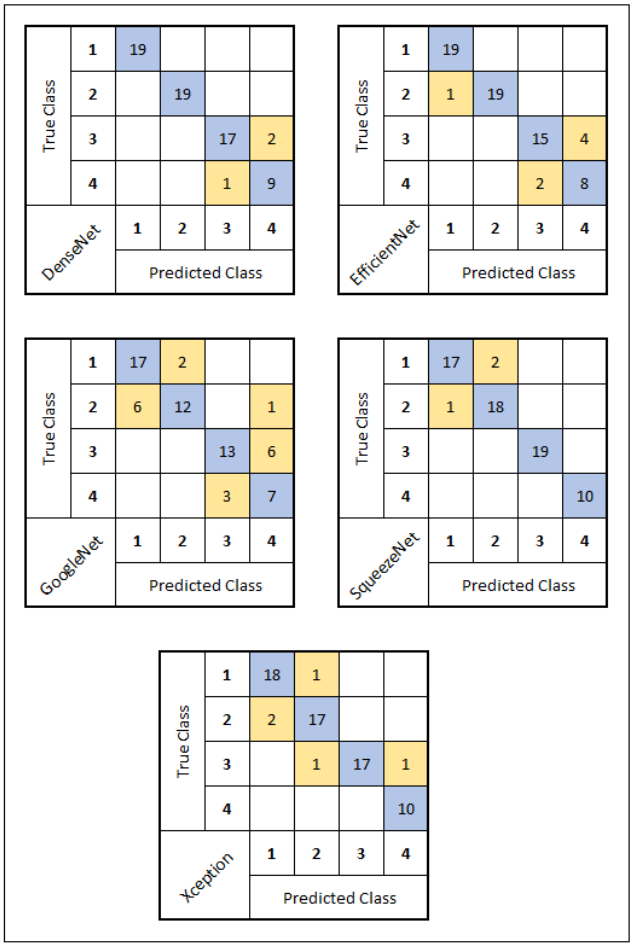


Fig. 6. The confusion matrices of the 4-class output variables with original images without ROI markings.

A 4-class database was created using photos without ROI markings. The objective here is to observe the effect of ROI marking on model performance. Upon examining Table III and Fig. 6(a), it is observed that the performance measurement metrics of the DenseNet, EfficientNet, Xception, and SqueezeNet models are higher than those of the GoogleNet model (>0.89). However, looking at the runtime, it is noted that the runtime of the SqueezeNet model is quite short. These results indicate that the SqueezeNet model has a high superiority in terms of both performance and runtime.

TABLE IV. THE PERFORMANCE OF THE 3-CLASS OUTPUT VARIABLES WITH ORIGINAL IMAGES WITHOUT ROI MARKINGS FOR EACH MODEL

	Precision	Recall	F1 Score	Acc	Time
DenseNet	0,9649	0,9683	0,9648	0,9701	23:34
Efficient	1,0000	1,0000	1,0000	**1**	07:26
GoogleNet	0,9474	0,6781	0,9473	0,9552	01:45
XCeption	0,9474	0,9545	0,9470	0,9552	16:14
Squeezenet	0,9359	0,9431	0,9355	0,9403	**01:08**

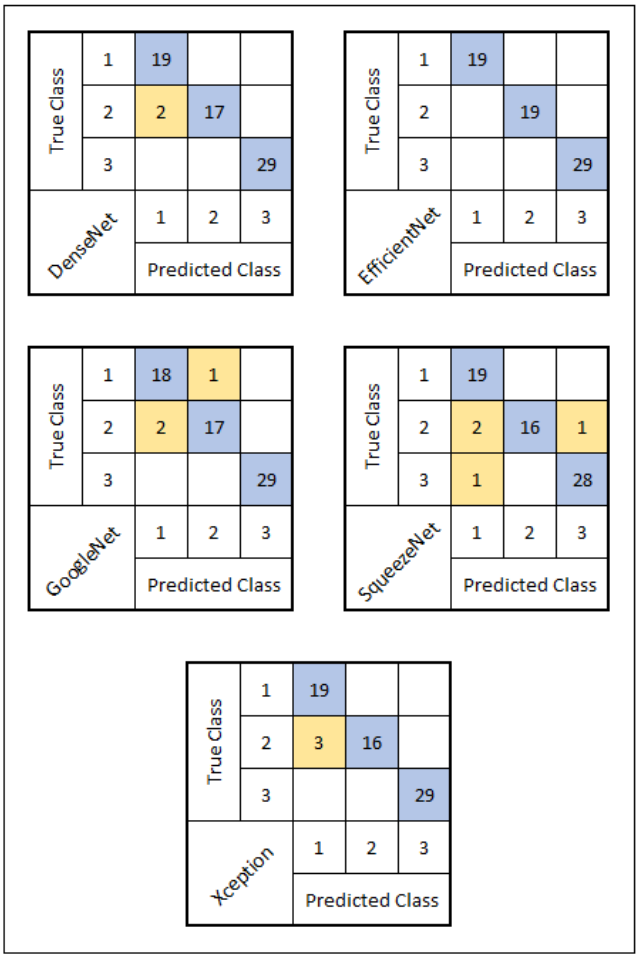


Fig. 7. The confusion matrices of the 3-class output variables with original images without ROI markings.

As observed in the dataset with ROI markings, the prediction error rate between the 3rd and 4th classes in photos without ROI markings is relatively high, as shown in Fig. 6. To improve the accuracy rates of the model performances, the dataset was updated by merging the 3rd and 4th classes, similar to the previous approach. Upon examining Table IV and Fig. 7, it is observed that the performance measurement metrics of all models are high (>0.94). However, looking at the runtime, it is noted that the runtime of the SqueezeNet model is quite short. According to these results, the EfficientNet model has high superiority in terms of performance, while the GoogleNet and SqueezeNet models have high superiority in terms of runtime.

IV. CONCLUSION

The aim of this study was to predict the growth stages of plants using a dataset of weed photographs through deep learning methods. The growth stages of the plants were divided into four different stages (classes). In line with this objective, ROI images of the leaves were extracted from the photos in the dataset using the Lifex program. This resulted in a total of 448 photos. Five different deep learning models were applied to this dataset.

According to the confusion matrix results, the accuracy rate for the leaves in the 4th class was lower because they resembled those in the 3rd class. As a result, the data for the 3rd and 4th classes were merged to form a new dataset, and the same deep learning methods were applied to this revised dataset.

Finally, a dataset was created using the raw, unmarked versions of the same images, and the previously mentioned deep learning models were applied again in both 3-class and 4-class formats.

In general, when the performance metrics of the models applied to the 4-class dataset marked with ROI was evaluated, no significant differences were observed (4 class ROI F1 Score: 0.8759). However, when the 3rd and 4th classes in the dataset were combined, an increase in the performance metrics of all models was observed (3 class F1 Score: 0.9649). Similarly, in the dataset created without ROI marking on the photos, the measurement parameters were found to be superior compared to the other conditions (4 class F1 Score: 96.05- 3 class F1 Score: 1.00).

This study is limited to the 4 weed species which are Lamb's Quarters (Chenopodium album), Jerusalem Oak Goosefoot (Dysphania botrys), Prickly Lettuce (Lactuca serriola), and Sow Thistle (sonchus oleraceus). In future studies, it is planned to extract the leaves from images using segmentation methods with the same dataset moreover other studies may be conducted on other weeds from different regions.

REFERENCES

[1] Zimdahl, R. L., Weed-crop competition: a review, second edition. USA. Blackwell, 2007.

[2] Knezevic, S. Z., & Datta, A., The critical period for weed control: revisiting data analysis. *Weed Science*, *63*(SP1), 188-202.

[3] Uludag, A., Uremis, I., Tursun, N., & Bukun, B. A, review on critical period for weed control in Turkey. Pp 37 *in The Proceedings of 6th International Weed Science Congress [17-22 June 2012, Hangzhou, China]*.

[4] Uremis, I., Uludag, A., Ulger, A., & Cakir, B., Determination of critical period for weed control in the second crop corn under Mediterranean conditions. *African Journal of Biotechnology*, 2009, *8*(18).

[5] Monaco, T. J., Weller, S. C., & Ashton, F. M., Weed science: principles and practices. John Wiley & Sons, 2002.

[6] Güzel, B., & Okatan, E., Agriculture and AI. Dynamics Changed by AI, 2022, 199-224 (in Turkish).

[7] LeCun, Y., Bengio, Y., & Hinton, G., Deep learning. nature, 521(7553), 2015, 436-444.

[8] Vasileiou, M., Kyrgiakos, L. S., Kleisiari, C., Kleftodimos, G., Vlontzos, G., Belhouchette, H., & Pardalos, P. M., Transforming weed management in sustainable agriculture with artificial intelligence: A systematic literature review towards weed identification and deep learning. Crop Protection, 2024, 176, 106522.

[9] Espejo-Garcia, B., Mylonas, N., Athanasakos, L., Fountas, S., & Vasilakoglou, I., Towards weeds identification assistance through transfer learning. Computers and Electronics in Agriculture, 2020, 171, 105306.

[10] Sunil, G. C., Zhang, Y., Howatt, K., Schumacher, L. G., & Sun, X., Multi-species Weed and Crop Classification Comparison Using Five Different Deep Learning Network Architectures. Journal of the ASABE, 2024.

[11] Pandey, S., Yadav, P. K., Sahu, R., & Pandey, P., Improving Crop Management with Convolutional Neural Networks for Binary and Multiclass Weed Recognition. In 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things, 2024, (IDCIoT) (pp. 539-543). IEEE.

[12] Garibaldi-Márquez, F., Flores, G., & Valentín-Coronado, L. M., Corn/Weed Plants Detection Under Authentic Fields based on Patching Segmentation and Classification Networks. Computación y Sistemas, 28(1), 2024, 271-282.

[13] Trong, V. H., Gwang-hyun, Y., Vu, D. T., & Jin-young, K., Late fusion of multimodal deep neural networks for weeds classification. Computers and Electronics in Agriculture, 2025, 175, 105506.

[14] Peteinatos, G. G., Reichel, P., Karouta, J., Andújar, D., & Gerhards, R., Weed identification in maize, sunflower, and potatoes with the aid of convolutional neural networks. Remote Sensing, 2020, 12(24), 4185..

[15] Chen, D., Lu, Y., Li, Z., & Young, S., Performance evaluation of deep transfer learning on multi-class identification of common weed species in cotton production systems. Computers and Electronics in Agriculture, 2022, 198, 107091.

[16] Gao, J., French, A. P., Pound, M. P., He, Y., Pridmore, T. P., & Pieters, J. G., Deep convolutional neural networks for image-based Convolvulus sepium detection in sugar beet fields. Plant methods, 2020, 16, 1-12.

[17] Ahmad, A., Saraswat, D., Aggarwal, V., Etienne, A., & Hancock, B., Performance of deep learning models for classifying and detecting common weeds in corn and soybean production systems. Computers and Electronics in Agriculture, 2021, 184, 106081

[18] Zhang, R., Wang, C., Hu, X., Liu, Y., & Chen, S., Weed location and recognition based on UAV imaging and deep learning. International Journal of Precision Agricultural Aviation, 2020, 3(1).

[19] Sharpe, S. M., Schumann, A. W., & Boyd, N. S., Goosegrass detection in strawberry and tomato using a convolutional neural network. Scientific Reports, 2020, 10(1), 9548.

[20] Tronrio, K., Puerto, A., Pedraza, C., Jamaica, D., & Rodríguez, L., A deep learning approach for weed detection in lettuce crops using multispectral images. AgriEngineering, 2020, 2(3), 471-488.

[21] Hasan, A. M., Sohel, F., Diepeveen, D., Laga, H., & Jones, M. G., A survey of deep learning techniques for weed detection from images. *Computers and electronics in agriculture*, 2021, *184*, 106067.

[22] Rai, N., Zhang, Y., Ram, B. G., Schumacher, L., Yellavajjala, R. K., Bajwa, S., & Sun, X., Applications of deep learning in precision weed management: A review. *Computers and Electronics in Agriculture*, 2023, *206*, 107698.

[23] Nioche C, Orlhac F, Buvat I., Texture — User guide, local image features 419 extraction, 2024. Available from: www.lifexsoft.org

[24] Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Campbell, J. P., Introduction to machine learning, neural networks, and deep learning. Translational vision science & technology, 2020, 9(2), 14-14.

[25] Goodfellow, I., Bengio, Y. & Courville, A. Deep Learning, MIT Press, 2016.

[26] Huang G, Liu Z, Van Der Maaten L, Weinberger K Q., Densely connected convolutional networks, In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, (4700−4708).

[27] Zhang A, Lipton Z C, Li M, Smola A J., Dive into deep learning, Cambridge University Press.

[28] Vasilev I., Advanced Deep Learning with Python: Design and implement advanced next-generation AI solutions using TensorFlow and PyTorch, Packt Publishing Ltd, 2019.

[29] Chollet F., Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 1251-1258.

[30] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Rabinovich A., Going deeper with convolutions, In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, 1−9.

[31] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K., Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. cite. arXiv preprint arxiv:1602.07360, 2016.

[32] Tan, M., & Le, Q., 2019, Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, 2019, (pp. 6105-6114). PMLR.

[33] Ravishankar N. Chityala, Kenneth R. Hoffmann, Daniel R. Bednarek, & Stephen, 524 Rudin "Region of interest (ROI) computed tomography", Proc. SPIE 5368, Medical 525 Imaging. Physics of Medical Imaging, 2004, https://doi.org/10.1117/12.534568

[34] Hossain, M. S., Shahriar, G. M., Syeed, M. M., Uddin, M. F., Hasan, M., Shivam, 464 S., & Advani, S., Region of interest (ROI) selection using vision transformer 465 for automatic analysis using whole slide images. Scientific Reports, 2023, 13(1), 11314.

[35] Bianco S, Cadene R, Celona L, Napoletano P., 2018, Benchmark analysis of representative deep neural network architectures. IEEE access, 2018, 6, 64270-64277.

[36] Hasanpour S H, Rouhani M, Fayyaz M, Sabokrou M., Lets keep it simple, using simple architectures to outperform deeper and more complex architectures. arXiv preprint arXiv:1608.06037, 2016.

[37] Mumuni A, Mumuni F., Data augmentation: A comprehensive survey of modern approaches, Array, 2022, 100258.

[38] Rebuffi S A, Gowal S, Calian D A, Stimberg F, Wiles O, Mann T A., Data augmentation can improve robustness, Advances in Neural Information Processing Systems, 2021, 34, 29935−29948.

[39] Kumar, S. (n.d.). Data augmentation increases accuracy of your model, but how? Medium. Retrieved December 23, 2024, from https://medium.com/secure-and-private-ai-writing-challenge/data-augmentation-increases-accuracy-of-your-model-but-how-aa1913468722

[40] Talukder, M. A., Layek, M. A., Kazi, M., Uddin, M. A., & Aryal, S., Empowering covid-19 detection: Optimizing performance through fine-tuned efficientnet deep learning architecture. Computers in Biology and Medicine, 2024, 168, 107789.

Target Detection of Leakage Bubbles in Stainless Steel Welded Pipe Gas Airtightness Experiments Based on YOLOv8-BGA

Huaishu Hou, Zikang Chen*, Chaofei Jiao
School of Mechanical Engineering, Shanghai Institute of Technology, Shanghai, 201418, China

***Abstract*—Gas-tightness experiment is an effective means to detect leakage of stainless steel welded pipe, and the vision-based bubble recognition algorithm can effectively improve the efficiency of gas-tightness detection. This study proposed a new detection network of YOLOv8-BGA using the YOLOv8 model as a baseline, which can achieve effective identification of leakage bubbles and bubble images are collected under different lighting conditions in a practical industrial inspection environment to create a bubble dataset. Firstly, a C2f_BoT module was designed to replace the C2f module in the backbone network, which improved the feature extraction capability of the model; secondly, the convolutional layer of the neck network was replaced by using the GSConv module, which achieved the model lightweighting; thirdly, the C2f-BM attention mechanism was added before the detection layer, which effectively improved the model performance; finally, the WIoU was used to improve the loss function, which improved the detrimental effect of small bubbles of low-quality samples in the dataset on the gradient, and significantly improved the convergence speed of the network. The experimental results showed that the average leakage bubble detection accuracy of the YOLOv8-BGA model algorithm reached 97.7%, which improved by 5.3% compared with the baseline, and meets the needs of practical industrial inspection.**

Keyword—Image processing; stainless steel welded pipe; non-destructive testing; YOLOv8; attention mechanism; loss function

I. INTRODUCTION

Stainless steel welded pipe has a wide range of uses, and is widely used in chemical industry, automobile manufacturing, shipbuilding and other industrial production areas [1], but it is prone to leakage problems, which has a greater impact on the subsequent use of the product. Gas tightness test is an effective means to test whether the stainless steel welded pipe leakage, when the welded pipe leakage exists, air bubbles are generated in the water, and the identification of air bubbles is an important indicator for judging the gas tightness. The traditional airtightness detection for workers to observe, the method is subject to subjective factors, easy to miss, low detection efficiency [2-4].Vision-based airtightness testing can exclude the influence of human subjective factors, assess the results through the indicators, make the test results more accurate, and improve the efficiency of product testing. Vision-based airtightness detection has traditional image processing method [5] and deep learning target detection based method [6].

Traditional image processing algorithms rely on classical image processing techniques based on edge contour and circumferential curvature fitting to manually extract the bubble edge features. Qaddoori [7] used the use of Hough's Circle Transform algorithm to identify the tiny bubbles in the graph based on Canny operator and segmented the centre and edges of the bubbles by two thresholds to calculate the average diameter and number of bubbles in the graph; Wen [8] designed a new image processing algorithm based on the concept of differential segmentation while considering the geometry and deflection angle of the bubbles; Akdemir [9] proposed a detection method using wavelet transform denoising and entropy threshold segmentation. The above proposed methods need to modify the parameters to respond to different detection environments, and have weak generalisation ability for identifying bubbles, and are prone to miss detection and false detection. Deep learning based target detection method has higher detection rate, stronger generalisation ability and better robustness for bubble recognition.

Deep learning target detection algorithms are classified into two types: a two-stage algorithm with high accuracy but slow speed, such as Faster R-CNN [10], Mask R-CNN [11]; and a single-stage algorithm with a simple structure and high computational efficiency, such as Krysko, N.V.et al [12] through the integrated application of non-destructive testing techniques, computer vision and convolutional neural networks, the surface of pipeline pitting and defects were classified and quantitatively analysed; Zhao et al [13] proposed a 3D quartz crucible bubble detection method, which significantly improved the detection accuracy of tiny bubbles by optimising the YOLOv5 network structure, introducing dilated convolution and ECA-Net mechanism, and combining Kalman filtering with Hungarian matching algorithm. Due to the different generalisation ability of the model in different detection environments. The real-time identification of bubbles is affected by the ripples generated by the bubble movement, the lighting environment, impurities in the water, sedimentation and other factors, which makes the collected bubble images have a lot of noise and missing bubble boundaries, which in turn affects the accuracy of target detection.

In this study, a model algorithm YOLOv8-BGA based on YOLOv8 is developed to achieve effective identification of leakage bubbles. Firstly, a C2f_BoT module is designed to replace the C2f module in the backbone network, which improves the feature extraction capability of the model; secondly, the convolutional layer of the neck network is replaced by using the GSConv module, which realises the model lightweight; again, the C2f-BM attention mechanism is

*Corresponding Author.

added before the detection layer, which effectively improves the model performance; and lastly, the loss function is improved by using the WIoU to improve the dataset The unfavourable effect of small bubbles of low-quality samples on the gradient significantly improves the convergence speed of the network.

II. Related Work

The first single-stage target detection YOLO algorithm was first proposed by Redmon [14] in 2015.The algorithm achieves fast identification and precise positioning of the target to be detected through a regression method. Currently, the YOLO series of algorithms has been developed to YOLOv10, and a variety of improved algorithms have been derived in academia. For example, Li [15] and others improved the internal structure of YOLOv5s and proposed a 'YOLOv5s-ShuffleNetV2-DWconv-Add' model, which provides an efficient acquisition method for fruit picking robots; Sun [16] and others proposed a 'Pconv-Wide lightweight' model for fruit picking robots. Pconv-Wide lightweight convolutional simplified YOLOv7 model, which increases the detection accuracy of UAVs on small targets while reducing the number of model parameters; Zhao [17] et al. proposed a Res-Clo network for denoising pre-processing of SAR images, and designed a DML-YOLOv8w network, which improves the performance of the model in multi-scale detection; YOLOv8 uses an Anchor-Free detection head [18], which improves the generalisation ability while reducing the model complexity, and improves the detection speed and accuracy compared to the previous YOLO algorithm.

YOLOv8 has four network model structures, YOLOv8n, YOLOv8s, YOLOv8m, and YOLOv8l, and the choice of the structure depends on the actual detection scenario and resource conditions. The YOLOv8n model is mainly divided into four parts: (1) Data enhancement strategies are applied to the image in the input part, such as Mosaic data enhancement and adaptive anchor frame computation to enhance the data diversity; (2) The backbone network is processed by multiple convolutional downsampling at key nodes, thus effectively reducing the size of the feature map and preserving the spatial information of the image. Then C2f modules of different sizes are used to capture multi-scale features; (3) The neck network accepts feature maps of various scales obtained through 8, 16, and 32 times downsampling from the backbone network, and fuses shallow high-resolution features and deep low-resolution features through PANet (path aggregation network), followed by fusing the fused feature maps with rich spatial and semantic information. Then the fused feature maps with rich spatial and semantic information are output to the Head part for further processing; (4) The Head network receives the multi-scale feature maps from Neck, performs up-sampling and splicing operations on them to match the resolution of the input image, and adjusts the number of channels through the feature conversion layer.

YOLOv8 has three Detect layers of different scales, 80×80, 40×40, and 20×20, which are used to detect large, medium, and small sized targets in the image, respectively. The three prediction layers output the category probability and bounding box location of each cell, and the stability of the prediction box is evaluated by a confidence layer.YOLOv8 adopts a Decoupled-Head structure to separate the classification and detection tasks to improve the detection efficiency. Finally, IOU (Intersection over union) [19] is used as the loss function for the bounding box regression, and then NMS (Non-Maximum Suppression) [20] removes the redundant prediction frames so as to avoid the same target from being detected repeatedly and retain the best results. This paper is based on the YOLOv8n algorithm to improve. The structure of the YOLOv8 model is shown in Fig. 1.

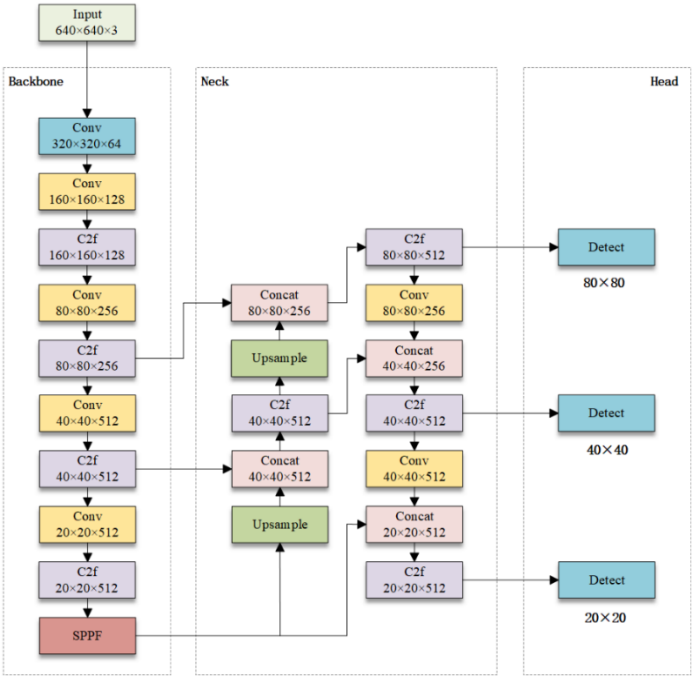


Fig. 1. YOLOv8 model structure.

III. Algorithm Design

A. Network Architecture Design

In this paper, four improvement modules for the YOLOv8n model are proposed. Firstly, the C2f_BoT module was constructed to replace the original C2f module of 40×40 and 20×20 sizes in the Backbone section, which improves the ability of the model to detect small and medium-sized bubbles; Secondly, the original Conv(Convolution) block is replaced by the GSConv in the neck network, which reduces the complexity of the model and improves the inference speed to achieve the effect of model lightweighting; Next, the C2f-BM attention mechanism is embedded in front of each detection head, which improves the detection accuracy of the model for small bubble targets; Finally, the WIoU loss function is introduced to speed up the network convergence. These improvements greatly improved the detection accuracy and detection speed of the original model in the stainless steel welded pipe gas-tightness inspection task. Fig. 2 shows the structure of the YOLOv8-BGA model obtained after the improvement.

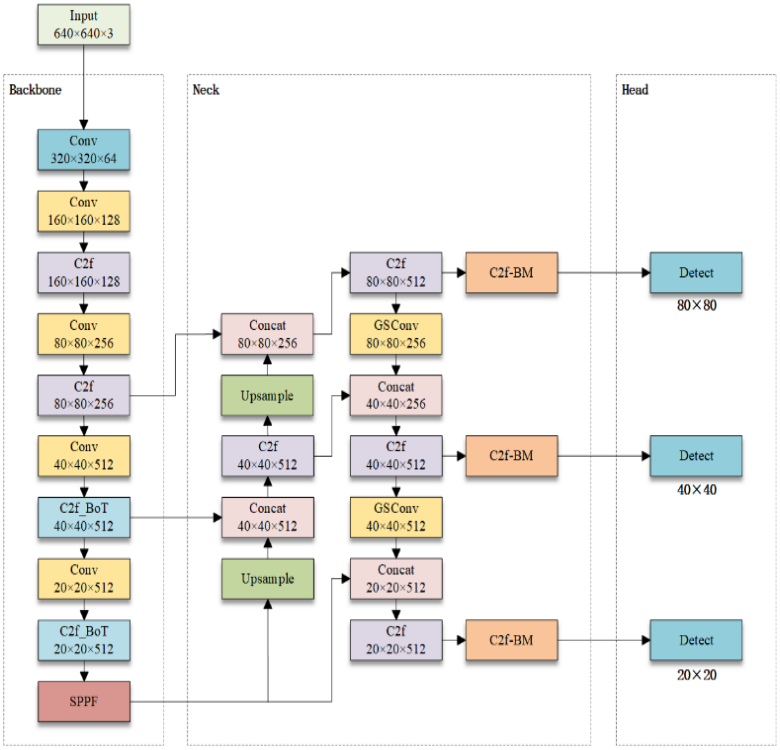
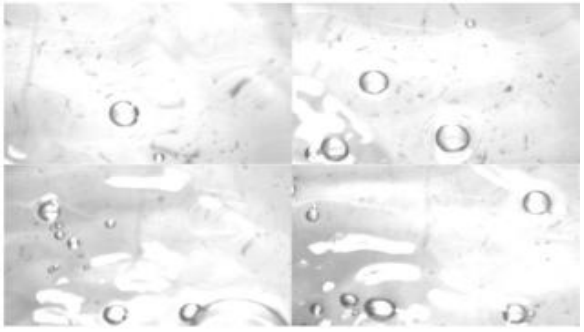


Fig. 2. YOLOv8-BGA model structure.

B. Mosaic Image Data Enhancement

Data enhancement of the self-constructed bubble dataset is performed using the Mosaic method at the YOLOv8n input. The Mosaic data enhancement algorithm improves the detection ability of the model in a small field of view by fusing multiple images into a single image according to a certain random scale. As shown in Fig. 3, four bubble images are randomly selected, certain parameters are set, and they are stitched into one image by random permutation, random size scaling, and random cropping.



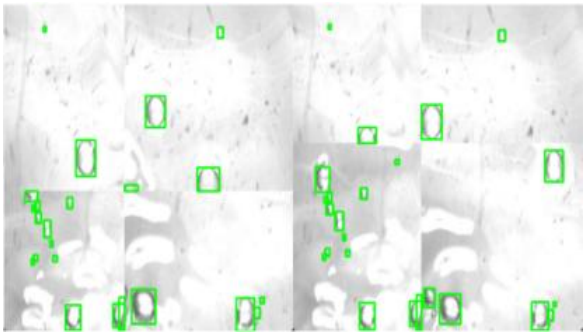


Fig. 3. Data enhancement.

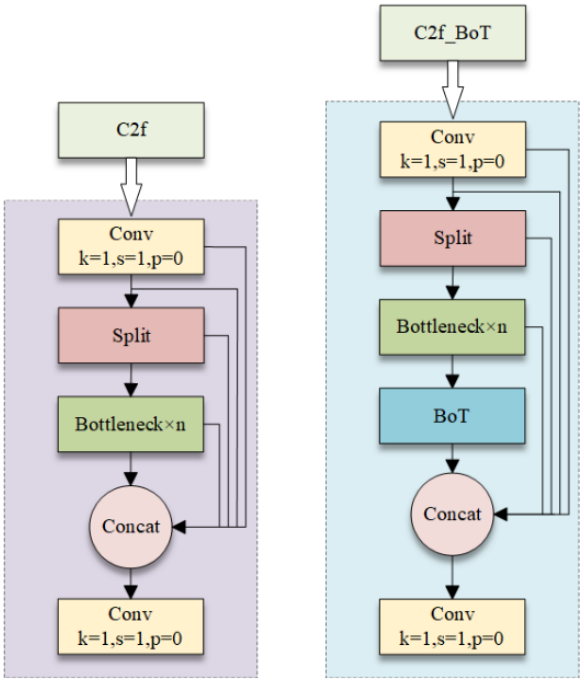
Using Mosaic data enhancement technique to strengthen the dataset can increase the data diversity, and the combined images obtained are more than the number of original images, which can get more small targets, and improve the generalisation ability of the trained model, which is a great improvement for detecting some tiny bubbles produced by subtle defects. And the network is able to batch process four image data in a single batch, which optimises the effect of the batch normalisation layer, thus reducing the computational burden on the GPU,

making it possible to achieve efficient training results on a single GPU.

C. C2f Module Improvements

The YOLOv8 network structure makes several uses of the C2f module [21], which is an improved residual block that enables feature fusion by connecting feature maps of different depths through splicing and upsampling operations. The spliced feature map fuses multi-scale features, which is helpful for the model to detect targets of different sizes. Compared with the traditional C3 module in YOLOv5, the C2f module increases the depth while ensuring smooth gradients, and the detection accuracy of the model is subsequently improved. However, for medium and small size targets, there are still cases of missed detection, in order to meet the accurate identification of tiny bubbles generated at the subtle defects of defective stainless steel welded pipes, and to increase the detection accuracy without increasing the additional computational cost, this paper proposes a C2f_BoT module for replacing the C2f module in the backbone network, and the structure of the C2f and the C2f_BoT network is shown in Fig. 4.

The BoT(Bottleneck Transformer) modules added after the last layer of Bottleneck in the C2f module, because the BoT module helps to improve the detection accuracy of the model for small and medium-sized targets, in order to maximise the utilisation of computational resources, this paper only replaces the C2f module with the 40×40 and 20×20 scales in the backbone network. The C2f_BoT module can effectively improve the accuracy. Moreover, the generalisation ability of the model can be further improved due to the MHSA attention mechanism in the BoT module. BoT network has a simple and powerful structure and is widely used in visual tasks such as image classification, target detection, semantic segmentation, etc. [22], and its structure is shown in Fig. 5.It replaces the 3×3 convolutional layers in the ResNet structure with the MHSA (Multi-Head Self-Attention), which automatically captures the dependencies between sequences when processing the input sequences and thus better understands the contextual information to improve the model performance. This operation significantly improves the baseline in the target detection task, resulting in lower latency. The C2f_BoT module with the added BoT structure improves the computational efficiency and detection accuracy of the model compared to the original C2f module, which is more suitable for the practical application of the model in industrial inspection.



(a)C2f network structure (b)C2f_BoT network structure

Fig. 4. C2f and C2f_BoT network structure.

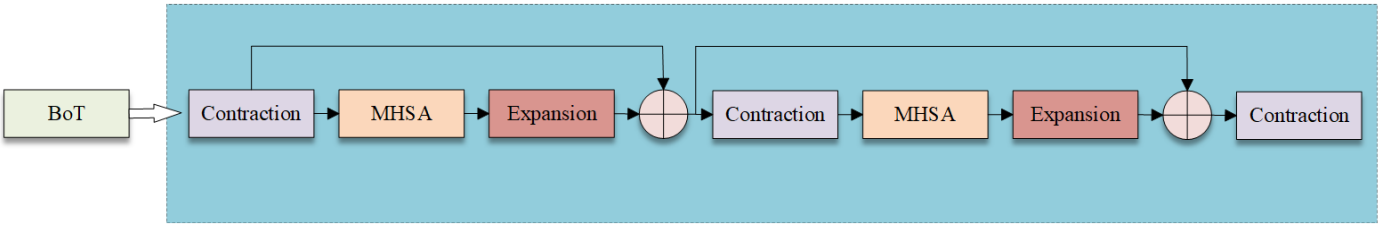


Fig. 5. Bottleneck transform structure.

D. Lightweight Processing

In order to reduce model complexity and reduce computational requirements, this paper uses the GSConv module to replace the original convolutional layer.GSConv is a hybrid of SConv (Standard convolution) and DSConv (Depth-wise separable convolution) [23] combined by the Shuffle convolution. DSConv incorporates deep convolution, which focuses on spatial feature extraction, and point-by-point convolution, which focuses on channel features. This structure performs deep convolution for each channel independently when processing features, and performs channel fusion at the output stage through a 1×1 convolutional layer to reduce the number of computations and parameters. Since DSConv separates each channel at the input image, some of the bubble image features are lost, which is degraded for airtightness detection accuracy. And GSConv can reduce the number of parameters and computation of the model while ensuring the training speed and detection accuracy. Its structure is shown in Fig. 6.

In this paper, the GSConv structure is introduced in the Neck part. In the Neck part, the channel dimension of the feature map extracted by the network is large, in order to maximally retain the circulation of feature information in the spatial and channel dimensions, and to avoid the loss of detail information, the original convolutional layer is replaced with GSConv in this part, which can reduce the model complexity under the premise of guaranteeing a certain model accuracy. If GSConv is used in all parts of the network, it will increase the depth of the network and thus prolong the inference time.

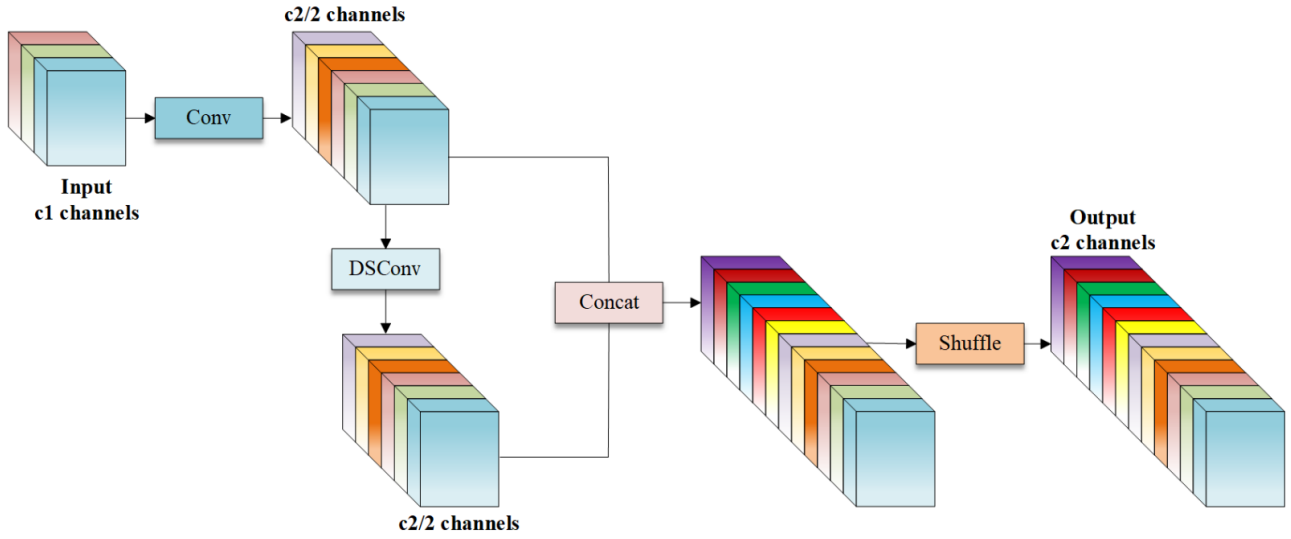


Fig. 6. GSConv structure.

E. C2f-BM

The attention mechanism is a resource allocation scheme that mimics human vision, which effectively allocates computational resources, prioritises critical tasks and alleviates information overload.YOLOv8 often faces the problem of feature loss in stainless steel welded pipe airtightness experiments for detecting small target bubbles, because small bubbles occupy fewer pixels in the feature map, and as the network undergoes many times of downsampling, the feature details of the small bubbles may be overlooked which makes the model performance degraded and generates problems such as leakage detection, which has a significant impact on the leakage detection generated by fine defects in steel pipes. Especially in the complex scene in the water, due to the water surface ripples caused by leakage, the impurity precipitation in the water and the pixel difference between the light and dark areas, the noise information in the background may occupy most of the feature space, so that the model is interfered with and cannot be accurately focused on the target area, resulting in the inability to accurately locate the target, and thus leakage detection.

To solve the above problems, the C2f-BM attention mechanism is designed in this section and embedded in front of the YOLOv8 detection head, the structure of which is shown in Fig. 7.

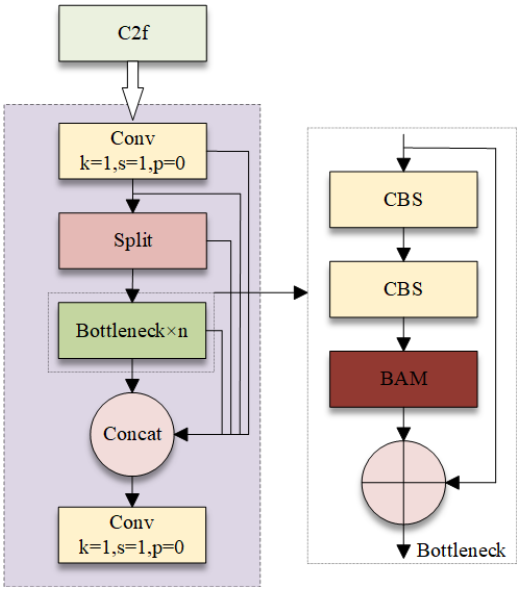


Fig. 7. G2f-BM structure.

The BAM (Bottleneck Attention Module) [24] is a simple and efficient attention mechanism, the structure of which is shown in Fig. 8, which mimics how the human visual system focuses on the critical parts of an image by separating the information from both channel and spatial pathways.

Channel attention focuses on enhancing the identification of feature channels relevant to the target detection task, while spatial attention focuses on the spatial location in the image, helping the model to focus more on the region where the small target is located, thus reducing the loss of feature information of the small target. The BAM combines the results of these two to generate a comprehensive Attention Map. This mechanism effectively improves the model's performance in the detection of small bubble target. This mechanism effectively improves the performance of the model in small bubble target detection, especially in complex detection scenarios in water. Through the spatial attention mechanism, the BAM mechanism can effectively suppress the irrelevant information in the background, reduce the noise interference, and make the model more focused on the feature space of the small target to enhance the robustness of the model. The C2f module itself optimises the feature extraction process, and by embedding the BAM mechanism, the model can focus on important features in the early feature extraction stage to improve the feature quality and model performance in the subsequent stages. Moreover, since C2f improves the gradient flow, the introduction of BAM can further optimise the gradient distribution, which makes the model more stable to be trained when dealing with complex visual tasks. Due to the lightweight nature of the BAM mechanism, it does not significantly increase the extra computational burden during embedding, allowing the whole model to improve performance while maintaining efficient computation.

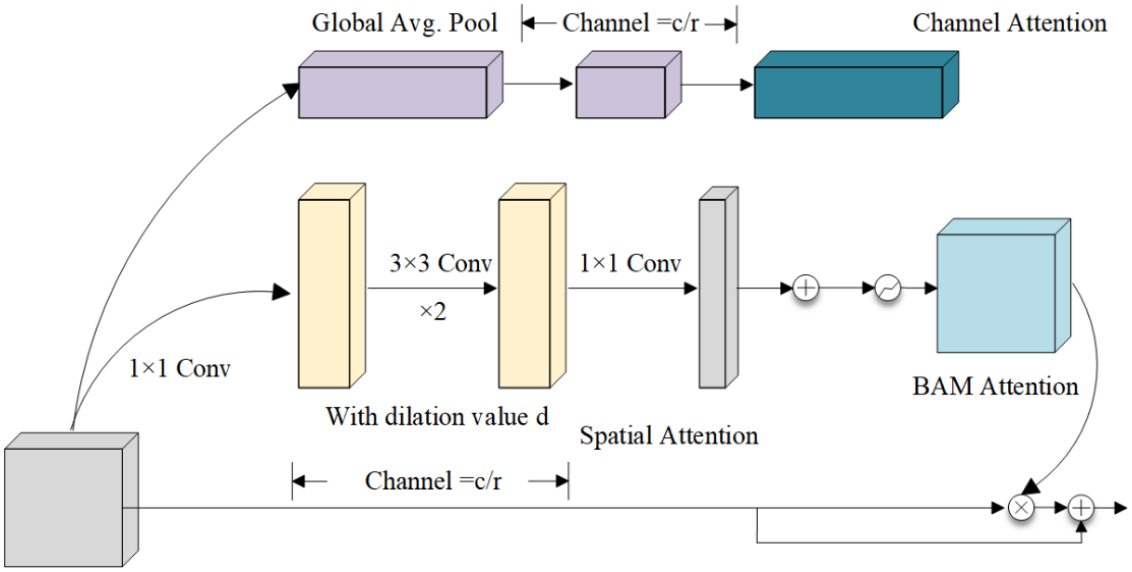


Fig. 8. BAM structure.

BAM calculates Attention Map through two branches: channel and space, and this paper introduces its calculation method from the following three points:

1) .Channel Attention (CA): The expectation maximisation algorithm aims to find maximum likelihood solutions for the hidden variables. In this step, the posterior probability distribution of the hidden variables under the current model parameters is calculated. In fact, it is to calculate the weights and responsibilities of each base for each pixel, the formula is shown in Eq. (1).

$$F_c = AvgPool(F) \tag{1}$$

where the global average pooling of the input feature map $F \in R^{C\times H\times W}$ is performed to obtain the average value of each channel to obtain the vector $F_c \in R^{C\times 1\times 1}$. C is the number of channels, H is the height, and W is the width.

Next, a multilayer perceptron (MLP) containing hidden layers is used to estimate the inter-channel attention from the channel vector F_c ,the formula is shown in Eq. (2).

$$MLP(F_c) = BN(W_1(\mathrm{Re}bLU(W_0F_c+b_0)+b_1)) \tag{2}$$

where, $W_0 \in R^{\frac{C}{r}\times C}$, $b_0 \in R^{\frac{C}{r}}$, $W_1 \in R^{\frac{C}{r}\times C}$, $b_1 \in R^{C}$, the size of the hidden layer is $\frac{C}{r}$, and r is the reduction ratio.

After the MLP, a batch normalisation layer is added to scale the output, the formula is shown in Eq. (3).

$$CA = BN(MLP(F_c)) \tag{3}$$

2) Spatial Attention (SA): First, the feature map F is projected to a smaller size using a 1×1 convolution to integrate and compress the feature map across channel dimensions, the formula is shown in Eq. (4).

$$F_R = Conv_{1\times1}(F) \tag{4}$$

Next, the contextual information is effectively utilised with two 3×3 null convolutions, the formula is shown in Eq. (5).

$$F_D = DilatedConv_{3\times3}(F_R) \tag{5}$$

Finally, the feature map is again reduced to a spatial attention map using 1×1 convolution, the formula is shown in Eq. (6).

$$SA = Conv_{1\times1}(F_D) \tag{6}$$

3) Combination of CA and SA:

$$M(F) = \sigma(M_C(F) + M_S(F)) \tag{7}$$

where $M_C(F)$ is the channel attention mapping, $M_S(F)$ is the spatial attention mapping, and σ is the *Sigmoid* activation function. Finally, the 3D attention map $M(F)$ is multiplied element-by-element with the input feature map, and then added to the original input feature map to obtain the refined feature map $F^{'}$, the formula is shown in Eq. (8).

$$F^{'} = F + F \otimes M(F) \tag{8}$$

where $\otimes$ denotes element-by-element multiplication.

This combination of BAM mechanisms is simple yet effective in adaptively assigning higher weights to small targets by balancing information from both the channel and spatial branches, while facilitating gradient flow.

F. Loss Function Improvement

The original YOLO family of algorithms uses IoU (Intersection over union) to calculate the bounding box regression loss, which refers to the ratio of the intersection and concatenation set of the true and predicted frames, as shown in Eq. (4).

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{9}$$

Where A denotes the true frame and B denotes the predicted frame, $A \cap B$ is its intersection, $A \cup B$ is its union set.

When the similarity ratio between the predicted frame and the real frame is higher, it means that the detection is better. However, there are shortcomings, in the detection when the real frame and the predicted frame do not intersect, that is, when $A \cap B = 0$, it is impossible to determine the size of the distance between the predicted frame and the real frame.

Therefore, in order to solve the instability brought by the above situation to the detection, YOLOv8 adopts CIoU (Complete-IoU) [25] instead of IoU. CIoU adds distance and aspect ratio to IoU, and its calculation formula is shown in Eq. (10)- Eq. (12).

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(A,B)}{c^2} + \alpha V \tag{10}$$

$$V = \frac{4}{\pi^2}(\arctan\frac{\omega^A}{\omega^A} - \arctan\frac{\omega^B}{\omega^B}) \tag{11}$$

$$\alpha = \frac{V}{1 - IoU + V} \tag{12}$$

where, $\rho^2(A,B)$ is the Euclidean distance between the centroids of the true and predicted frames, V is a parameter indicating the consistency of the aspect ratio, and α is a parameter used to balance the ratio.

Using CIoU, the minimum value of the distance between the real and predicted frames, the aspect ratio and the distance between the centre points of the bounding boxes can be calculated, which improves the stability of the target box regression. However, CIoU only considers the centre distance of the two bounding boxes, and the matching of the boundaries cannot be accurately assessed, leading to an impact on the detection effect when the target shape changes. The WIoU loss function, on the other hand, introduces a weight function on the basis of IoU, which makes it flexible to adjust the weights between different samples even in more complicated situations. For samples with poorer labelling quality in the dataset, WIoU performs better compared to other boundary loss functions.

In this paper, WIoU (Wise-IoU) is replaced as the loss function, which is proposed on the basis of Focal-EIoU [25].Most of the research on loss function in recent years assumes that the samples in the dataset are of high quality, and is committed to improving the fitting ability of the bounding box loss, while when there are low-quality samples in the dataset, it will jeopardise the model detection performance if the regression of the bounding box on the low-quality samples is improved, while Focal-EIoU is proposed to solve the problem, and its formula is shown in Eq. (13) - Eq. (14).

$$L_{EIoU} = 1 - IoU + \frac{\rho^2(A,B)}{c^2} + \frac{\rho^2(\omega^A,\omega^B)}{c_\omega{}^2} + \frac{\rho^2(h^A,h^B)}{c_h{}^2} \tag{13}$$

$$L_{Focal-EIoU} = IoU^{\gamma} L_{EIoU} \tag{14}$$

where γ is the hyperparameter used to control the curvature of the curve and its focusing mechanism is static. To fully exploit the potential of the non-monotonic focusing mechanism, WIoU [26] uses a dynamic non-monotonic mechanism to assess the quality of the anchor frame, which gives it a better performance when facing the targets with different geometric factors, and the detection performance of the samples from low-quality datasets is improved, and improves the model's generalisation ability. The WIoU loss function expression is shown in Eq. (15) - Eq. (17).

$$L_{WIoU} = \frac{\beta}{\delta\alpha^{\beta-\delta}} R_{WIoU} L_{IoU} \tag{15}$$

$$R_{WIoU} = exp(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(\omega_g{}^2 + h_g{}^2)^*})p \tag{16}$$

$$L_{IoU} = 1 - IoU \tag{17}$$

where β describes the outlier of the anchor frame mass, α and δ are hyperparameters. ω_g and h_g are the dimensions of the minimum enclosing frame, (x, y) and (x_{gt}, y_{gt}) are the coordinates of the centre points of the anchor and target frames. Using the WIoU loss function in the improved model, the performance of target detection can be improved by introducing the weighting factor, and for targets with different sizes and shapes, the WIoU can give more reasonable weights to different samples, so that the improved model can better learn the characteristics of bubbles with different sizes and shapes during the training process, and improve the robustness of the stainless steel welded pipe airtightness detection task.

IV. LEIS EXPERIMENTS AND ANALYSIS OF RESULTS

A. Experimental Methodology Flow

The experimental methodology of this paper is specifically divided into the steps of data preparation, model construction, model training, model validation, and result analysis. The flow chart of the experimental method is shown in Fig. 9.

The specific work is as follows:

1) Data preparation. It is necessary to collect the datasets (D1, D2, D3, D4) of stainless steel defective welded pipe leakage detection, perform data enhancement on the D1 and D2 datasets, and divide the datasets into a training set and a validation set.

2) Model construction. Taking YOLOv8n as the base model, the model improvement work in section III of this paper is carried out on it in order to obtain the improved YOLOv8-BGA model.

3) Model training. Train the model on D1 and D2 datasets respectively.

4) Model validation. Use the trained model weights to validate the model on the validation sets of D1, D2, D3, and D4 datasets, and calculate Precision, Recall, and mAP metrics.

5) Result analysis. Compare the detection performance of different models, analyze the generalization ability of models in different environments, and conduct ablation experiments to verify the effectiveness of each improvement module.

B. Experimental Setup

The experiments in this paper were conducted under Windows 11 operating system, using a CPU model i7-13620H with 16G of RAM, a graphics card NVIDIA RTX4060, accelerating the GPU using CUDA11.8 and CUDNN8.8.1, and running under the Pytorch2.0.0 deep learning framework. In this experiment, the official website YOLOv8n model weights are used as the basic network model, batchsize is set to 32 and epoch is 200.an industrial face array camera with model number MV-CS004-10GM is used, and the camera is shown in Fig. 10.

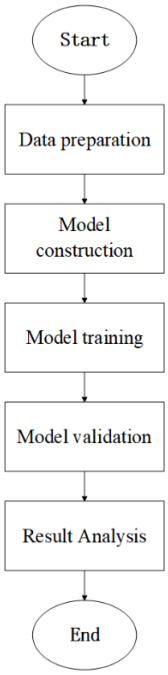


Fig. 9. Experimental methodology.

Fig. 10. Industrial camera.

An airtightness tester model LS11Z-100 was used for inflation and pressurization, as shown in Fig. 11.

Fig. 11. Airtightness tester diagram.

The light source is a strip light source and the schematic diagram of the image acquisition device is shown in Fig. 12.

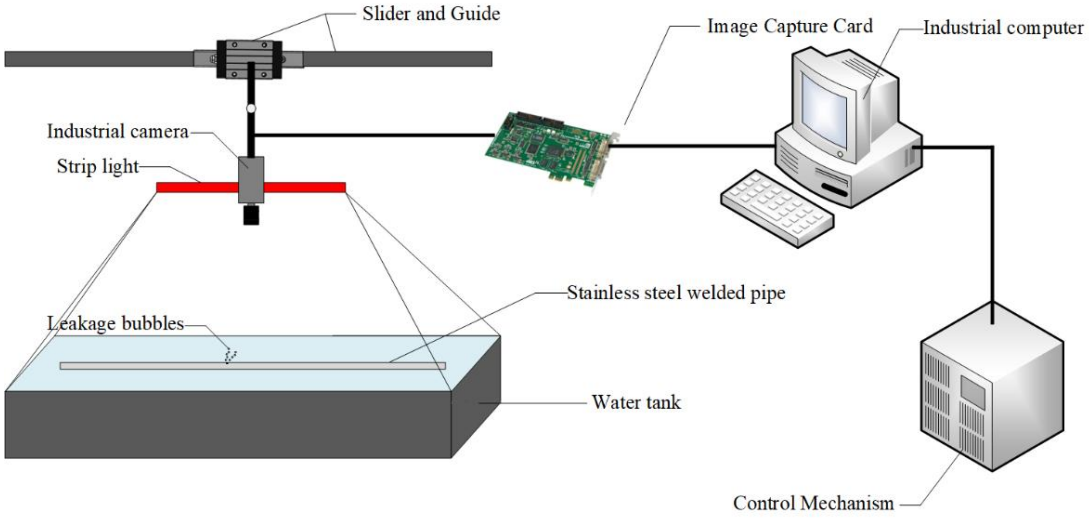
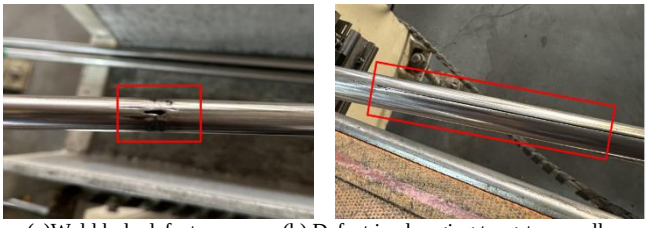


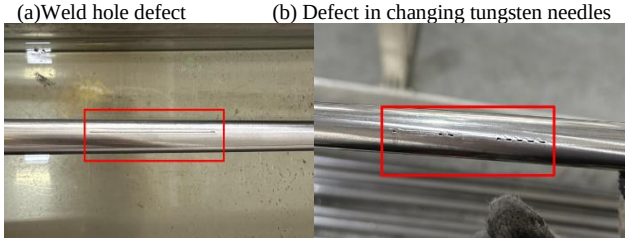
Fig. 12. Image acquisition device.

C. Experimental Data Set

The dataset in this paper was acquired in an image acquisition platform built in a real industrial inspection environment. The stainless steel defective welded pipe used for the experiment is obtained from the processing workshop, and several common defective welded pipes are shown in Fig. 13.



(a)Weld hole defect (b) Defect in changing tungsten needles



(c) Soldering defect (d) Skipped weld defects

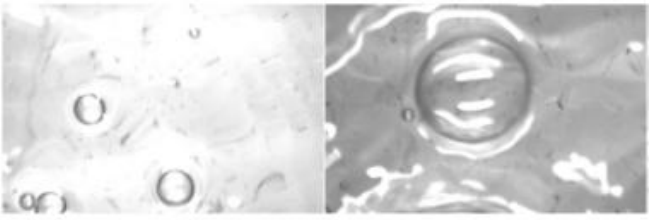
Fig. 13. Stainless steel welded pipe common defects diagram.

In order to evaluate the generalisation ability of the improved model, two datasets, D1 and D2, were acquired in two different lighting environments (bright and dimmer), and secondly, two noisy images were acquired in these two lighting environments as D3 and D4 datasets. All the dataset images were labelled and their basic characteristics are detailed in Table I.

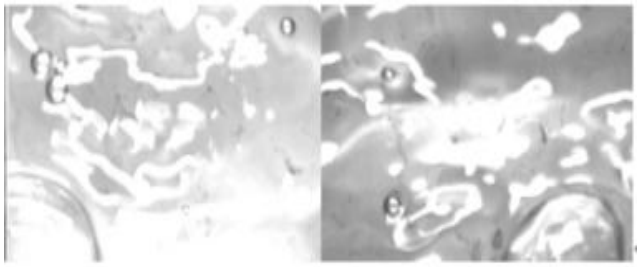
TABLE I BASIC INFORMATION ON THE D1-D4 DATA SETS

Dataset	Feature	train set	val set
D1	Sufficient light	4640	1160
D2	dusky	5080	1270
D3	Sufficient light, noise	0	945
D4	Dusky, noise	0	975

Due to the defects of the original data, Mosaic data enhancement is performed on D1 and D2 datasets, and the images of D1 dataset are expanded to 5800, D2 dataset is expanded to 6350, and D3 and D4 datasets are not processed. The D1 and D2 datasets are divided into training set and validation set with the ratio of 8:2. Some images of D1-D4 datasets are shown in Fig. 14.



D1 D2



D3 D4

Fig. 14. Partial images of the dataset.

D. Assessment of Indicators

In the field of deep learning target detection, the main metrics for model performance evaluation are P (Precision), R (Recall) and mAP (mean AP), and the model complexity can be evaluated by GFLOPS (Giga Floating Point Operations) and Parameters [27]. FPS were used to evaluate the performance of the model. In this paper, the model evaluation is also based on these criteria.p, R and mAP are calculated as shown in Eq. (18)-Eq. (21):

$$P = \frac{TP}{TP + FP} \tag{18}$$

$$R = \frac{TP}{TP + FN} \tag{19}$$

$$AP = P(r)dr \tag{20}$$

$$mAP = \frac{1}{N} AP \tag{21}$$

In Eq. (13) - Eq. (16), TP is the number of positive samples correctly detected; FP is the number of positive samples incorrectly detected; FN is the number of negative samples incorrectly detected; AP is the average precision; and N is the number of all predictions.

E. Error Analysis and Discussion

In the process of detecting leakage bubbles in stainless steel welded pipes using the YOLOv8-BGA model, several types of errors were observed. Small bubbles, particularly those generated by subtle defects, are more likely to be missed due to their low pixel occupancy in the feature map. The repeated downsampling in the network can lead to the loss of these fine details. Additionally, bubbles that appear in regions with high background noise or similar textures are prone to incorrect detection. For example, in the presence of water impurities or ripples, the model may confuse these with actual bubbles. These factors collectively contribute to the challenges faced during the detection process.

The impact of noise and ripples on detection accuracy cannot be overlooked. The movement of bubbles can generate ripples on the water surface, which introduce additional noise into the image. These ripples can interfere with the model's ability to accurately locate and identify bubbles, especially in complex underwater scenes. Variations in lighting conditions can cause significant reflection and refraction effects on the water surface, leading to false positives or false negatives as the model may misinterpret the light patterns as bubbles. Overlapping bubbles and those with unclear or irregular boundaries also pose challenges, as the model may struggle to distinguish individual bubbles or accurately delineate bubble contours. These issues highlight the need for further improvements to enhance the model's robustness and accuracy.

To address these challenges, several strategies were implemented. Mosaic data augmentation was employed to increase the diversity of the training dataset, helping the model learn to detect small targets more effectively. The C2f-BM attention mechanism was introduced to allow the model to focus more on critical regions, reducing the impact of noise and improving the detection accuracy for small bubbles. Additionally, the WIoU loss function was utilized to enhance the model's fitting ability, especially for low-quality samples with small bubbles, thereby improving the overall detection performance. Through these improvements, the YOLOv8-BGA model demonstrated enhanced robustness and accuracy in detecting leakage bubbles under various conditions. Future work will continue to explore advanced techniques and additional data collection under varied conditions to further improve the model's performance.

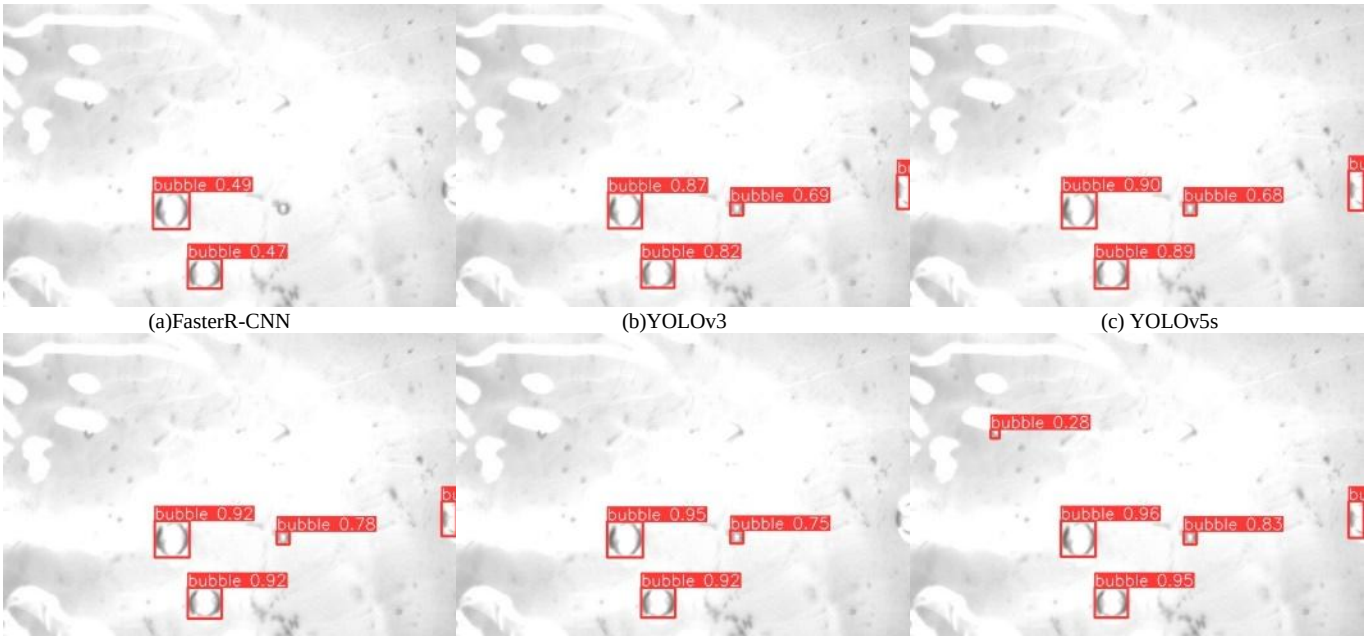
F. Experimental Results and Analysis

1) Comparison experiment: In order to verify the performance of the improved model in this paper, a comparison experiment is first taken. Several common target detection algorithms are selected and trained on D1 dataset under the same experimental environment and experimental configuration, and validated. The experimental results are detailed in Table II.

TABLE II COMPARATIVE EXPERIMENTAL RESULTS OF SOME MAINSTREAM ALGORITHMS

Model	P/%	R/%	mAP/%t	Parameters/M	FLOPS/G	FPS (F/S)
Faster R-CNN	60.1	63.7	64.6	14.3	19.8	25
YOLOv3	84.6	85.4	87.3	54	125.6	45
YOLOv5s	90.1	89.8	91.6	7.0	15.8	65
YOLOv8n	90.6	90.8	92.4	3.0	8.1	85
YOLOv8s	89.9	90.4	91.5	11.1	28.4	50
YOLOv8-BGA	95.4	96.3	97.7	2.7	7.5	89

Table II shows that the improved algorithm in this paper has excellent training effect on the D1 dataset, and has certain improvement compared with other algorithms. Compared with Faster R-CNN, it has the biggest improvement, with an average accuracy improvement of 34.1%; compared with YOLOv3, the average accuracy improvement is 11.4%; compared with YOLOv5s and YOLOv8s, it has an improvement of 7.1% and 7.2%, respectively; this paper improves the improved model by using YOLOv8n as the baseline, and the improved model has an average accuracy improvement of 5.3% compared with YOLOv8n.The FPS value of YOLOv8-BGA is 89, which is significantly higher than that of other algorithms, indicating that it not only improves the detection accuracy, but also optimizes the inference speed of the model, which can make the detection task achieve a good balance between real-time and accuracy. In addition, while the detection accuracy is improved, the number of parameters and inference time of the model are lower than the other five algorithms. It can be seen that YOLOv8-BGA has higher detection accuracy in the dataset of this paper, and at the same time, the model complexity is lower, which can better meet the actual industrial detection needs of stainless steel defective welded pipe bubbles. The comparison of the detection results of YOLOv8-BGA and the other algorithms is shown in Fig. 15.



(a)FasterR-CNN (b)YOLOv3 (c) YOLOv5s

(d) YOLOv8n (e) YOLOv8s (f) YOLOv8-BGA

Fig. 15. Comparison of detection results between YOLOv8-BGA and other algorithms.

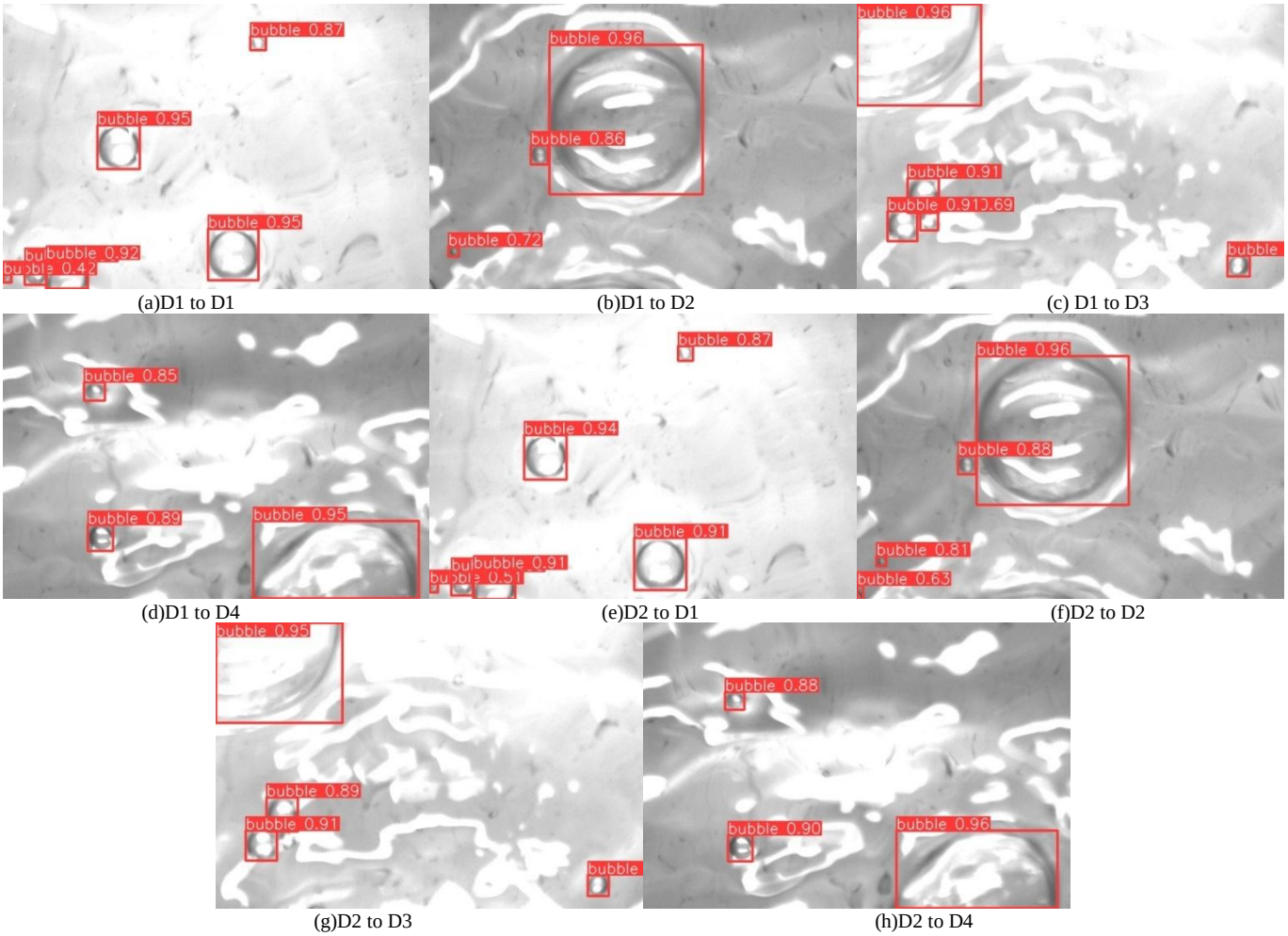
2) Verification of generalisation capabilities: In the actual detection process, the image quality is affected by a variety of factors such as illumination, water reflection and ripples, so it is crucial to evaluate the generalisation ability of the model. In this section of experiments, the improved algorithm is trained on D1 and D2 datasets under the same configuration, and the corresponding validation sets are validated with the respective training results. Precision and recall are chosen to evaluate the model performance, and the experimental results are detailed in Table 3.In this experiment, the influence of lighting conditions on the leakage bubble detection results of stainless steel welded pipe based on YOLOv8-BGA model is emphatically analyzed. In this section, two lighting environments is set up, bright and dark, and the results show that the lighting conditions have a significant impact on the model detection performance. The results show that the lighting conditions have a certain influence on the model detection performance.

TABLE III COMPARISON OF TEST RESULTS

mission	P/%	R/%
D1 Validated D1	95.4	96.3
D2 Validated D2	93.9	94.3

In low-illumination conditions, the overall brightness of the image is reduced, which directly leads to a decrease in image contrast. According to the fundamental principles of image processing, contrast is a key factor in distinguishing targets from the background. When contrast is reduced, the differences between small bubbles and the background become less distinct, especially for small bubbles generated by minor defects, which occupy fewer pixels in the feature map. In low-contrast images, the features of these small bubbles become even more difficult to extract. For example, in our experimental dataset, when the illumination is dim, the edge details of small bubbles are blurred, making it challenging for the model to accurately identify their contours. This results in larger deviations between the predicted and ground-truth bounding boxes when calculating the loss function, thereby reducing the detection accuracy of the model. Specifically, when validating on the dataset D2 under low-illumination conditions, the model's Precision is lower compared to that on the dataset D1 under bright illumination conditions. For instance, the Precision is 95.4% when D1 is validated on D1, while it is 93.9% when D2 is validated on D2.In bright illumination conditions, although the overall brightness of the image is sufficient, strong reflections become a significant factor affecting detection results. When light strikes the surface of the water and bubbles, reflection and refraction occur. According to the principles of optics, the angle of incidence is equal to the angle of reflection. When light is incident at a large angle, the intensity of the reflected light increases. Under bright illumination, due to the undulations of the water surface and the movement of bubbles, the direction of the reflected light constantly changes, causing the bubble boundaries to appear blurred in the image. This blurred boundary interferes with the model's accurate judgment of bubbles because the model relies on clear boundary features for localization and identification. For example, in the experiment, we observed that under bright illumination with undulating water surfaces, the model is prone to misjudging reflective areas as bubbles or inaccurately locating the boundaries of bubbles, thereby affecting Recall and Precision.

Then, in order to verify the generalisation ability of the improved algorithm under different environmental conditions, the D2, D3, and D4 datasets were validated using the training results of the D1 training set, and the D1, D3, and D4 were validated using the training results of the D2 training set, respectively, and the results are shown in Fig. 16.



(a)D1 to D1 (b)D1 to D2 (c) D1 to D3

(d)D1 to D4 (e)D2 to D1 (f)D2 to D2

(g)D2 to D3 (h)D2 to D4

Fig. 16. Comparison of some test results.

In assessing the performance of the model in various environments, precision and recall are also used as evaluation metrics, as detailed in Table IV.

TABLE IV RESULTS OF DIFFERENT ENVIRONMENTS

Mission	P/%	R/%
D1 Validated D2	88.7	85.6
D1 Validated D3	92.4	93.6
D1 Validated D4	85.9	86.1
D2 Validated D1	86.9	83.2
D2 Validated D3	83.4	82.9
D2 Validated D4	90.8	91.5

Comparing Table IV, it can be found that the validation is carried out under datasets with different environments, and although the precision and recall of the improved algorithm have decreased compared with those under the corresponding datasets, the model as a whole still maintains a good performance. The experimental results illustrate that the improved YOLOv8-BGA algorithm has good generalisation ability under different detection environments, and can effectively detect air bubbles generated by stainless steel defective welded pipes in real time under a variety of detection environments. Using the model trained under the D1 dataset to validate the D3 dataset is 9% more accurate compared to the model trained under the D2 dataset to validate the D3 dataset; similarly, using D1 to validate D4 is 4.9% less accurate compared to D2 to validate D4, which indicates that the performance of the improved model trained under the D1 dataset is higher than the performance of the model trained under the D2 dataset. Because the D1 dataset and the D3 dataset, the D2 dataset and the D4 dataset have the same kind of lighting conditions, which indicates that the lighting conditions have a greater impact on the training results of this model, and the image quality of the D1 dataset is higher than that of the D2, which may affect the performance of the model generalisation ability.

3) Comparative experiments with multiple datasets: To further validate the generalization and robustness of the improved model across diverse fields and industries, this section devises experiments to assess the performance of YOLOv8n and YOLOv8-BGA on publicly available datasets, namely the Bubble Image Database and UF-120.The Bubble Image Database, disseminated by researchers from the University of Queensland, encompasses 5,184 original images and 25,920 augmented images. It comprises five bubble categories: “Fine bubbles fully covering the viewing window”,

"fine bubbles partially covering the viewing window", "Co-existence of coarse and fine bubbles", "Only coarse bubbles", and "No bubbles". The UF-120 dataset consists of 120 high-quality underwater bubble images, representing a wide array of underwater scenarios and faithfully mirroring the complexity inherent in underwater environments. The experimental outcomes are presented in Table V and Table VI.

TABLE V EXPERIMENTAL RESULTS OF BUBBLE IMAGE DATABASE

Model	mAP/%	Parameters/M	FLOPS/G	FPS(F/S)
YOLOv8n	82.1	3.1	8.2	70
YOLOv8-BGA	82.7	2.8	7.6	72

TABLE VI EXPERIMENTAL RESULTS OF UF-120

Model	mAP/%	Parameters/M	FLOPS/G	FPS(F/S)
YOLOv8n	89.1	3.0	8.2	77
YOLOv8-BGA	89.8	2.7	7.6	75

Evident from the findings presented in Table V, the YOLOv8-BGA model exhibits an increment of 0.6 in the mAP. Concurrently, the model experiences a reduction of 9.7% in the Parameters and 7.3% in the FLOPS, with no discernible alteration in the detection speed. Inspection of the results in Table 6 reveals that the YOLOv8-BGA model registers a mAP increase of 0.7, accompanied by decreases of 10% in Parameters and 7.3% in FLOPS. The detection speed remains commendably stable, showing no significant fluctuations. Experimental outcomes on publicly accessible datasets, including the Bubble Image Database and UF-120, incontrovertibly demonstrate that, within the domain of bubble detection, the YOLOv8-BGA model surpasses the YOLOv8n model with respect to accuracy, model complexity, and computational efficiency. Overall, the YOLOv8-BGA model demonstrates good utility value and robustness in practical industrial applications.

4) Ablation experiment: In order to test the effectiveness and model performance of adding BoT module, replacing GSConv module, adding C2f-BM attention mechanism and improving loss function in the improved algorithm on the gas tightness detection of stainless steel defective welded pipe, ablation experiments are carried out on YOLOv8-BGA algorithm under D1 dataset, and the results of the ablation experiments are shown in Table VII in detail.

Analysing the experimental results in Table VII, the YOLOv8n model is used as the baseline, and the improved modules are added individually in sequence for comparison.YOLOv8-BoT replaces the 40×40 and 20×20 C2f modules in the Backbone section with the C2f_BoT module, which improves the average accuracy by 0.9% without any change in the number of Parameters and FLOPS, which is due to the fact that the BoT module is able to automatically identify and capture the dependencies in the input sequence through its multi-head self-attention mechanism, which helps the model to extract features more efficiently and improves the detection accuracy, the introduction of the BoT module did not significantly increase the model complexity, so the FPS was not significantly reduced; YOLOv8-GSConv is to replace the original convolutional layer in the Neck part with the GSConv module, and the number of parameters is reduced by about 0.2M, and the FLOPS is reduced by 0.4G.The GSConv module reduces the amount of computation through deeply separable convolution, which improves the lightness of the model and the speed of inference, so the FPS is improved, the model achieves a lightweight effect while maintaining the detection accuracy; YOLOv8-BM is to add the C2f-BM attention mechanism in front of each detection layer, the number of parameters is reduced by about 0.1M, the FLOPS is reduced by 0.2G, and the average accuracy is improved by 4.4%, which indicates that the expectation maximisation algorithm reduces the computational volume of the model while the detection accuracy is improved dramatically, the attention mechanism performs well on this dataset, improving performance along with inference speed and FPS values; YOLOv8-WIoU introduces the WIoU loss function, and the average accuracy of the model is improved by 1.9%, and the WIoU loss function improvesthe degree of model fitting, which in turn improves the accuracy of model recognition. The introduction of the WIoU loss function, which mainly optimizes the training process, has a small impact on the inference speed, so the FPS value is unchanged; YOLOv8-BGA is the final model after adding all of the above modules, and the number of references is lowered by about 0.3M, FLOPS is reduced by 0.6G, and the average accuracy is improved by 5.3%.After combining all the improvements, the model is optimized in all aspects, and the inference speed is further optimized while maintaining high accuracy, and the final FPS value of YOLOv8-BGA reaches 89.

TABLE VII RESULTS OF ABLATION EXPERIMENTS

Model	BoT	GSConv	C2f-BM	WIoU	mAP/%	Parameters/M	FLOPS/G	FPS(F/S)
YOLOv8n					92.4	3.0	8.1	85
YOLOv8-BoT	√				93.3	3.0	8.1	84
YOLOv8-GSConv		√			92.1	2.8	7.7	87
YOLOv8-BM			√		96.8	2.9	7.9	86
YOLOv8-WIoU				√	94.3	3.0	8.1	85
YOLOv8-BGA	√	√	√	√	97.7	2.7	7.5	89

The experimental results verify that the YOLOv8-BGA model obtained after the improvement of this paper has been improved compared with the original model in terms of reasoning speed, detection accuracy, model complexity, etc., and is able to better complete the task of stainless steel defective welded pipe airtightness detection compared with other algorithms proposed above.

V. CONCLUSION

1) The improved model is better than the original YOLOv8n model in terms of inference speed, detection accuracy, and model complexity, and can achieve effective identification of leakage bubbles.

2) With the introduction of the C2f-BoT module and the WIoU loss function, the improved model possesses excellent detection capability for tiny bubbles and reduces leakage detection.

3) The improved model has a strong generalisation ability, and can also have a good detection ability in poorer airtightness detection environments.

4) Due to the limitations of the experimental conditions, the accurate linkage between real-time detection, positioning and alarm at the leakage of stainless steel welded pipe needs to be further investigated in order to improve the popularisation of the improved model in the field of industrial airtightness detection.The model algorithm is a useful exploration and reference for other different products airtightness detection.

REFERENCES

[1] Li, L., Yu, T., Xia, J., Gao, Y., Han, B., Gao, Z.: Failure analysis of L415/316L composite pipe welded joint.Engineering Failure Analysis, 158, 107981.(2024)

[2] Jiao, C., Zhang, R., Hou, H., Zhao, Z., Tang, C.and Yun, H.: Research on ultrasonic non-destructive evaluation algorithm for ultimate tensile strength of stainless-steel welded pipe welds based on width learning.Nondestructive Testing and Evaluation, pp.1-15.(2024)

[3] Heim, D.and Miszczuk, A.: Modelling building infiltration using the airflow network model approach calibrated by air-tightness test results and leak detection.Building Services Engineering Research and Technology, 41(6), pp.681-693.(2020)

[4] F.Gao, J.Lin, Y.Ge, S.Lu and Y.Zhang.: A Mechanism and Method of Leak Detection for Pressure Vessel: Whether, when, and how.IEEE Transactions on Instrumentation and Measurement 69(9): 6004-6015.(2020)

[5] Luo, Q., Fang, X., Liu, L., Yang, C.and Sun, Y.: Automated visual defect detection for flat steel surface: A survey.IEEE Transactions on Instrumentation and Measurement, 69(3), pp.626-644.(2020)

[6] Ren, Z., Fang, F., Yan, N.and Wu, Y.: State of the art in defect detection based on machine vision.International Journal of Precision Engineering and Manufacturing Green Technology, 9(2), pp.661-691.(2021)

[7] Qaddoori, A.S., Saud, J.H., Hamad, F.A.: A classifier design for micro bubble generators based on deep learning technique.Materials Today: Proceedings, 2023, 80: 2684-2696.(2023)

[8] Wen, J., Sun, Q., Sun, Z., Gu, H.: An improved image processing technique for determination of volume and surface area of rising bubble.International Journal of Multiphase Flow, 104, 294-306.(2018)

[9] Akdemir, B., Öztürk, S.: Glass surface defects detection with wavelet transforms.International Journal of Materials, Mechanics and Manufacturing ,3(3): 170-173.(2015)

[10] Xu, J., Ren, H., Cai, S., Zhang, X.: An improved faster R-CNN algorithm for assisted detection of lung nodules.Computers In Biology and Medicine, 153: 106470.(2023)

[11] Wang, W., Xu, X., Yang, H.: Intelligent Detection of Tunnel Leakage Based on Improved Mask R-CNN.Symmetry, 2024, 16(6): 709.(2024)

[12] Krysko, N.V., Shchipakov, N.A., Kozlov, D.M., Kusyy, A.G., & Skrynnikov, S.V.: Classification and Sizing of Surface Defects in Pipelines Based on the Results of Combined Diagnostics by Ultrasonic, Eddy Current, and Visual Inspection Methods of Nondestructive Testing.Russian Journal of Nondestructive Testing, 59(12): 1315-1323.(2023)

[13] Zhao Q, Zheng C, Ma W.: An Improved Crucible Spatial Bubble Detection Based on YOLOv5 Fusion Target Tracking.Sensors, 22(17): 6356.(2022)

[14] Redmon, J., Divvala, S., Girshick, R.and Farhadi, A.: You only look once: Unified, real-time object detection.Proceedings of the IEEE conference on computer vision and pattern recognition.(2016)

[15] Li, M., Zhang, J., Liu, H., Yuan, Y., Li, J., Zhao, L.: A lightweight method for apple-on-tree detection based on improved YOLOv5.Signal, Image and Video Processing, 1-15.(2024)

[16] Sun, F., He, N., Wang, X., Liu, H., Zou, Y.: YOLOv7-P: a lighter and more effective UAV aerial photography object detection algorithm.Signal, Image and Video Processing, 1-9.(2024)

[17] Zhao, S., Tao, R., Jia, F.: DML-YOLOv8-SAR image object detection algorithm.Signal, Image and Video Processing, 1-13.(2024)

[18] Zhang, Y., Zhang, H., Huang, Q., Han, Y., Zhao, M.: DsP-YOLO: An anchor-free network with DsPAN for small object detection of multiscale defects.Expert Systems with Applications, 241: 122669.(2024)

[19] Yan, J., Wang, H., Yan, M., Diao, W., Sun, X., Li, H.: IoU-adaptive deformable R-CNN: Make full use of IoU for multi-class object detection in remote sensing imagery.Remote Sensing, 11(3): 286.(2019)

[20] Kang, S.H., Palakonda, V., Kim, I.M., Kang, J.M.and Yun, S.: Enhanced Non-Maximum Suppression for the Detection of Steel Surface Defects.Mathematics, 11(18): 3898.(2023)

[21] Zou, Y.and Fan, Y.: An Infrared Image Defect Detection Method for Steel Based on Regularized YOLO.Sensors, 24(5): 1674.(2024)

[22] Nakai, K., Chen, Y.W., Han, X.H.: Enhanced deep bottleneck transformer model for skin lesion classification.Biomedical Signal Processing and Control, 78: 103997.(2022)

[23] Huang, D., Tu, Y., Zhang, Z., Ye, Z.: A Lightweight Vehicle Detection Method Fusing GSConv and coordinate attention mechanism.Sensors, 24(8): 2394.(2024)

[24] Chen, Z., Tian, S., Yu, L., Zhang, L., Zhang, X.: An Object Detection Network Based on YOLOv4 and Improved Spatial Attention Mechanism.Journal of Intelligent & Fuzzy Systems, 42(3), 2359-2368 (2022)

[25] Aswal, D., Shukla, P., Nandi, G.C.: Designing effective power law-based loss function for faster and better bounding box regression.Machine Vision and Applications, 32(4): 87.(2021)

[26] Yang, X., Liu, C., Han, J.: Reparameterized underwater object detection network improved by cone-rod cell module and WIOU loss.Complex & Intelligent Systems, 1-16.(2024)

[27] Li, S., Zhang, X., Shan, R.: Enhanced YOLOv5 for Efficient Marine Debris Detection.Engineering Letters, 32(8).(2024)

Broccoli Grading Based on Improved Convolutional Neural Network Using Ensemble Deep Learning

Zaki Imaduddin[1], Yohanes Aris Purwanto[2]*, Sony Hartono Wijaya[3], Shelvie Nidya Neyman[4]
Department of Computer Science, IPB University, Bogor, Indonesia[1, 3, 4]
Department of Mechanical and Biosystem Engineering, IPB University, Bogor, Indonesia[2]
Departement of Informatics, Sekolah Tinggi Teknogi Terpadu Nurul Fikri, Depok, Indonesia[1]

***Abstract*—The demand for broccoli in Indonesia has been increasing significantly, with an annual growth of approximately 15% to 20%. However, the supply availability remains insufficient, and its quality is often inconsistent. Therefore, a grading process is needed to classify broccoli into grades A, B, and C based on color, size, and shape. Currently, the grading process is carried out solely by market intermediaries, while farmers and the general public have a limited understanding of this process. This research developed an automated grading method using a Convolutional Neural Network (CNN) based on two broccoli images' top and side views. Three main parameters, namely color, size, and shape, were identified from these images and used as grading determinants. An ensemble deep learning technique was applied by training each parameter separately using several CNN models, namely ResNet50, EfficientNetB2, VGG16, and Improved CNN. These were then combined in the testing phase using a voting technique. The test was conducted 64 times with various model combinations to achieve the best accuracy. A significant contribution of the Improved CNN lies in the shape feature, which achieved a maximum performance of 95%. This study also compared evaluation metrics such as precision, recall, F-Score, and accuracy across different model combinations.**

***Keywords*—Grading; convolution neural network; ensemble deep learning; voting**

I. INTRODUCTION

Broccoli (Brassica oleracea L. var. italica) is a widely cultivated cruciferous vegetable valued for its high nutritional content and economic significance. It is a rich source of essential vitamins such as C, K, and A [1] and bioactive compounds such as glucosinolates, which are studied for their potential health benefits. Globally, broccoli is a key commercial vegetable, with strong export demand from regions such as the United States, China, and Europe. In Indonesia, the demand for broccoli has increased by up to 20% annually, driven by growing consumption in restaurants, hotels, modern retail markets, and exports [2]. However, domestic supply is constrained by a lack of standardized quality grading, creating inconsistencies that disadvantage farmers and traders alike. Broccoli grading is crucial in determining market value and quality, with morphological attributes such as size, shape, color, and compactness as key indicators. The morphology of the broccoli head is particularly significant, as it reflects the crop's overall quality and resilience to environmental stress [3]. Traditionally, grading has been performed manually, leading to inconsistencies due to subjective evaluation. Various approaches have been employed to assess broccoli quality, including dry geometric and weight measurements, mass spectrometry analysis, and non-contact sensor technologies [4]. Among these, image-based methods using RGB digital cameras and deep learning algorithms have emerged as promising solutions because of their non-destructive nature, cost-effectiveness, and high accuracy. Recent studies have demonstrated the potential of Convolutional Neural Networks (CNNs) in broccoli grading, utilizing algorithms such as Mask R-CNN and ResNet for tasks such as detecting harvest readiness, estimating weight, and analyzing color attributes.

Research conducted by Blok et al. employed the Mask R-CNN algorithm [5] and successfully detected 229 out of 232 harvest-ready broccoli heads across three cultivars. The study concluded that the algorithm demonstrated better generalizability across multiple broccoli cultivars. Previous studies have also contributed by combining the Viewpoint Feature Histogram (VFH) with a Support Vector Machine (SVM), enabling precise broccoli detection and facilitating automated systems for detecting and measuring broccoli heads. This approach proved effective in achieving the goal of determining the optimal harvest timing [6].

A recent study by Zhou [4] developed a dataset of 100 broccoli head images captured using a custom-designed imaging system under controlled conditions. The study employed an improved ResNet CNN algorithm to extract broccoli pixels from the background and estimate their weights. Additionally, the particle Swarm Optimization Algorithm (PSOA) and Otsu method were applied to evaluate the broccoli quality, achieving an accuracy rate of 0.896 [4]. However, Zhou et al. study was limited to color as the sole criterion, whereas grading broccoli typically requires multiple parameters to assess its overall quality.

Another study compared the performance of various models, including ResNet, DenseNet, MobileNetV2, NASNet, and EfficientNetB2, to determine the best model for grading apples and bananas [7]. EfficientNet yielded the highest accuracies of 99.2% for the training data and 98.6% for the testing data. However, this approach has not been applied to broccoli datasets.

This study introduces several contributions, and those are: 1) the development of a broccoli grading model based on three features: color, size, and shape, where the data are divided into three independent features, each assessed without reliance on the others; 2) optimization of convolutional and classification

models for the broccoli grading process; and 3) comparison of grading results using multiple CNN models.

The grading criteria in this study differ from those of previous research because of the distinct physical characteristics of domestically produced broccoli compared to imported varieties. The criteria include shape (degree of roundness), color, size, and flower compactness (density). Image acquisition was conducted from two perspectives: top and side views. The research modified several deep learning models, including ResNet50, EfficientNetB2, VGG16, and Improved CNN.

In addition, the study utilized a parallel Ensemble Learning technique during the training phase. This approach allows models to be developed independently, with no interdependencies, ensuring that errors from one model are less likely to align with errors from others. Consequently, the weaknesses of one model can be mitigated by those of the other. Ensemble learning has been successfully applied across various fields and often outperforms single-model approaches [8].

The predictions from all deep learning models were aggregated and reused during the testing phase, where a voting mechanism was applied to make classification decisions. This image-based grading algorithm aims to enhance broccoli's post-harvest quality and standardization while advancing the agricultural industry in Indonesia, particularly benefiting broccoli farmers.

II. Literature Review

The utilization of Convolutional Neural Network (CNN) architectures had seen rapid growth since 2012, when Krizhevsky's breakthrough in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) demonstrated the CNN's efficiency for image classification across various domains [9][10]. Specifically in agriculture, CNN-based approaches have been employed for fruit classification and detection, highlighting CNN's pivotal role of CNNs in image processing for this sector. CNN operations typically involve six essential layers [10][12]:

- Input Layer: This layer accepts raw images as input and forwards them to the subsequent layers for feature extraction.
- Convolution Layer: Each output connects to a small region in the input using a weight matrix (filter or kernel). Multiple filters can be applied within each convolution layer to generate several 2D outputs stacked into an output volume.
- ReLU (Rectified Linear Unit): Negative values in the output of the convolution layers are replaced with zero, accelerating the training process.
- Pooling Layer: This process downsamples feature maps to achieve translation invariance. Feature map dimensions are reduced using average and max pooling techniques.
- Fully Connected Layer: This final layer integrates all filtered image data, converting it into labels and categories.
- Softmax Layer: Positioned before the output layer, this layer generates decimal probabilities for each class, ranging from 0 to 1, enabling CNNs to extract, process, and classify image data features efficiently.

TABLE I. CNN-Based Research on Broccoli Plants

Author	CNN Model	Result
[5]	Mask Region-based Convolutional Neural Network	Successfully detected 175 out of 176 test datasets
[6]	KNN & SVM	95.2%
[13]	3D information based on convolution neural network	Below 90%
[14]	improved resnet	Below 90%
[15]	Organised Edges Segmentation (OES) and Organised Region Growing Segmentation (ORG	Low generalization level
[16]	Gaussian Mixture Model	97.9%

Table I compares various CNN models and other methods in detection or classification tasks based on referenced studies. Mask R-CNN performed well, detecting 175 out of 176 test datasets, while the Gaussian Mixture Model achieved the highest accuracy at 97.9%. KNN & SVM also performed well with 95.2%, whereas 3D CNN and improved ResNet had accuracies below 90%. OES & ORG showed a low generalization level, indicating limitations in handling new data. the researchers adopted the parallel Ensemble Learning technique, which integrates multiple models or classifiers to improve the prediction accuracy compared to using a single model [8]. Each model or classifier within the ensemble is generally trained on slightly different datasets (distinct training data or varying feature selection methods).

Ensemble Learning aggregates predictions from multiple base models to produce more stable and accurate results. Two primary strategies introduce diversity among the baseline classifiers: homogeneous ensembles and heterogeneous. Homogeneous ensembles consist of baseline classifiers of the same type, in which each classifier is trained on different datasets. The feature selection method remained consistent across the datasets. By contrast, Heterogeneous ensembles comprise baseline classifiers of various types, allowing each classifier to adopt distinct methodologies in processing the training data [11]. Heterogeneous ensembles typically demonstrate superior generalization capabilities than homogeneous ensembles [17]. Several commonly used ensemble techniques include the following:

- Averaging computes the average prediction from multiple models.
- Bagging (Bootstrap Aggregating) involves training multiple models on random subsets of training data and subsequently combining their outputs. Bagging aims to stabilize models by reducing the variance [18].
- Random Forest is a bagging method that employs numerous decision trees to generate predictions.
- Stacking combines predictions from multiple base models and utilizes another model (a meta-model) to aggregate these predictions. Boosting enhances the

model performance by assigning greater importance to previously misclassified data points. A prominent boosting algorithm is AdaBoost [19], which improves the performance of the decision trees.

Among ensemble methods, bagging and boosting are the most frequently applied in classification tasks. These approaches are widely recognized for their robust theoretical foundations and exceptional empirical results [11]. Existing studies highlight that bagging and boosting are particularly effective when applied to decision tree models [20].

III. Research Method

This research involves several methodological stages to achieve broccoli grading classification: Data Acquisition, Data Pre-processing, Model Training, Model Evaluation, and Grading (Fig. 1).

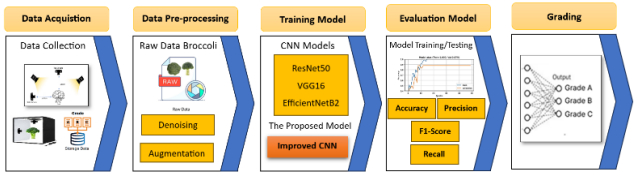


Fig. 1. Methodology of classification and grading broccoli.

A. Data Acquisition

The first stage of the research process was the dataset collection. The broccoli images were gathered and classified into grades with the assistance of experts. The dataset consisted of 450 broccoli images captured from the top and side views, which were categorized into grades A, B, and C.

B. Data Pre-Processing

After the data were collected, the images were processed through pre-processing stages.

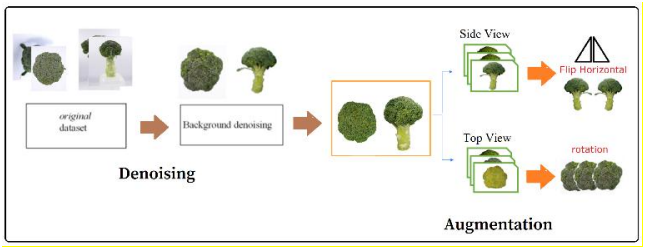


Fig. 2. The architecture of denoising and data augmentation process.

In Fig. 2, the first step is denoising, which removes unnecessary backgrounds to improve image quality. The second step is augmentation, which involves transforming the images using rotation techniques for top-view images and flipping them for side-view images. The denoising and data augmentation process aimed to increase the variation in the dataset.

C. Training Model

At this stage, the data training process uses ensemble deep learning. The ensemble learning technique enables the classification of three distinct feature subsets, color, size, and shape, by training each feature independently without relying on the others. This independent training allows the model to capture and store information from each feature more effectively, reducing the risk of interference between features. As a result, during the testing process, the model can focus on accurately reading and determining the grade of each test data point. This approach enhances the model's generalization ability across different feature types and improves overall classification performance.

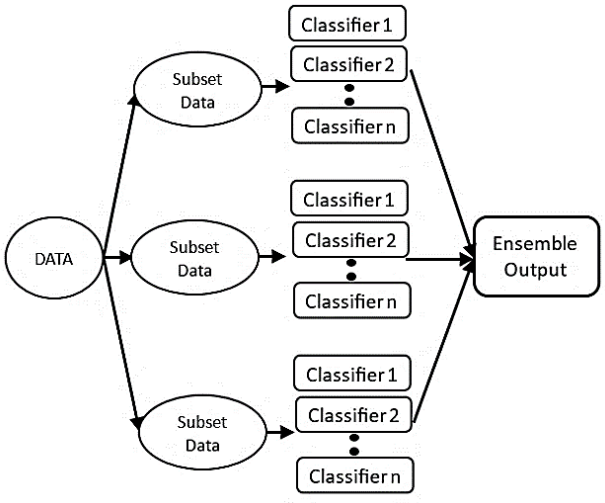


Fig. 3. Development of training architecture using ensemble learning.

As shown in Fig 3, each subset is trained using several convolutional neural network (CNN) models, which are known to be effective for image recognition. The models used included ResNet50, VGG16, EfficientNetB2, and the proposed model, which is an Improved CNN.

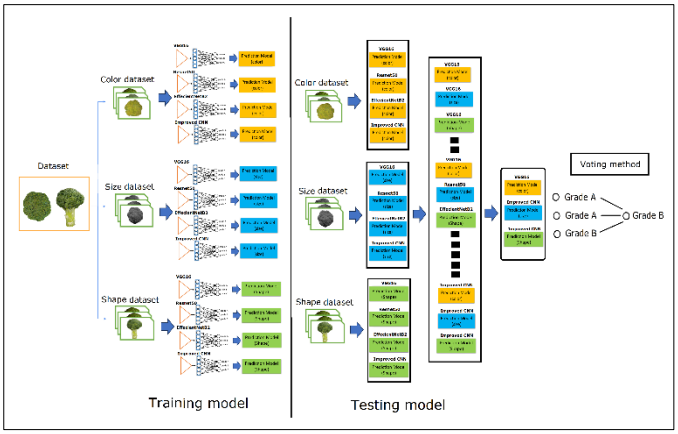


Fig. 4. The architecture training and testing data using ensemble deep learning.

In Fig. 4, each subset based on color, size, and shape is trained using several Convolutional Neural Network (CNN) models: ResNet50, VGG16, EfficientNetB2, the proposed model, and the Improved CNN. The outputs of these models were used in the testing phase.

D. Evaluation Model

The trained models were evaluated using performance metrics such as accuracy, precision, recall, and F1-Score. Below are some equations for this method.

$$\text{Acc} = \frac{(\text{TP}+\text{TN})}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (1)$$

The Accuracy (Acc) formula was used to measure the model's performance by calculating the number of True Positive and True Negative elements as the numerator and the total number of elements in the Confusion Matrix as the denominator. The True Positive and True Negative elements represent the correct predictions made by the model and are located on the matrix's main diagonal. Meanwhile, the denominator includes all the elements incorrectly classified by the model outside the main diagonal. Therefore, accuracy indicates how well the model can make correct classifications for positive and negative cases compared with the entire dataset [9].

$$Precision = \frac{TP}{(TP + FP)} \quad (2)$$

Precision measures the model's accuracy in identifying positive cases, indicating how well the model avoids misclassifying negative cases as positive cases.

$$Recall = \frac{TP}{(TP + FN)} \quad (3)$$

Recall or Sensitivity indicates how well the model can remember or recognize all the existing positive cases. This metric is important because it ensures the model does not miss significant positive cases. The higher the recall value, the better the model captures all positive cases.

$$F1 - score = 2.\frac{(precision \times recall)}{precision + recall} \quad (4)$$

The F1 score is a metric that combines Precision and Recall to assess the performance of a classification model. The F1-score reached its optimal value when the model had high precision and recall, indicating that it effectively identified and recognized all positive cases in the data. Conversely, the F1-score will be low if the Precision or Recall is low, signaling inadequate model performance. Therefore, the F1 score provides an overview of how well the model can balance Precision and Recall, with one being the best and zero the worst.

E. Grading

In the final stage, the model's prediction results are translated into specific categories (e.g., Grade A, B, C) based on a voting technique, where the decision is based on the lowest grade. This process was used to classify broccoli quality according to predefined standards.

IV. Result and Discussion

A. Data Acquisition

Based on Fig. 5, all broccoli data that had been labeled and collected were placed in a photo box studio and photographed individually using a 12MP SLR camera. The labeling process was conducted in collaboration with three experts experienced in grading transit locations before distribution to supermarkets. The broccoli samples were then categorized into grades A, B, and C. Subsequently, images were captured from the top and side angles, maintaining a consistent distance of 20 cm between the object and the camera. Photobox Studio was equipped with adjustable lighting settings to ensure optimal image quality and minimize noise interference from the surrounding environment. This process was designed to produce high-quality images suitable for further analysis.

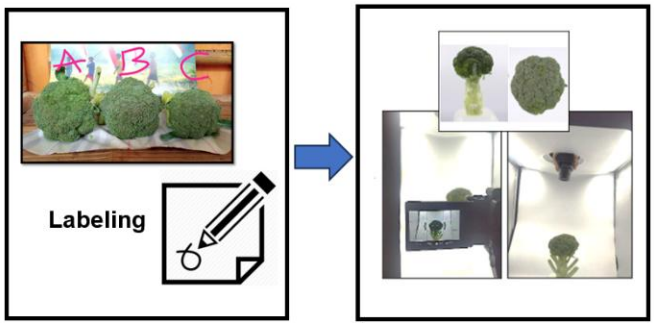


Fig. 5. Collecting data with photobox studio.

B. Data Pre-Processing

Table II shows the distribution of the dataset used for broccoli grading based on two perspectives, namely the top view and side view, as well as the impact of the data augmentation process. The table is divided into grades A, B, and C. In the top view, each grade (A, B, and C) contained 100 samples from the original dataset and 100 augmented samples, indicating that data augmentation doubles the dataset size for the top view. Meanwhile, each grade contained 50 samples from the original dataset and 50 augmented samples in the side view, demonstrating a similar augmentation process but fewer samples than in the top view. Overall, the total number of samples for the top view, including original and augmented data, was 600 (300 original + 300 augmented). In comparison, the total for the side view was 100 samples (50 original + 50 augmented), resulting in a combined dataset of 700 samples. This table emphasizes that the dataset is balanced across the three grade categories and illustrates how data augmentation is applied to both perspectives to enhance the model training performance.

TABLE II. Dataset Augmentation

Grade	Top view (original)	Augmentation	Side View (original)	Augmentation
Grade A	100	100	50	50
Grade B	100	100	50	50
Grade C	100	100	50	50
Tota Dataset	600		100	
	700			

After the data were augmented, data splitting was performed, in which 80% of the data were allocated for training and 20% for testing.

C. Training Model

At this stage, the data training process utilizes an ensemble deep learning approach with heterogeneous ensemble specifications, a technique that has shown better generalization than single models and homogeneous ensembles [21]. The broccoli dataset was divided into three subsets and processed using the four classifiers.

TABLE III. SUMMARY OF CNN AND PROPOSED METHOD

Hyperparameter	CNN Models			
	ResNet50	***EfficientNet B2***	***VGG16***	***Improved CNN***
Kernel 1	ResNet50 Architecture	EfficientNet B2 Architecture	VGG16 Architecture	32 (3x3)
Kernel 2				64 (3x3)
Kernel 3				128 (3x3)
Activation Function	ReLU	ReLU	ReLU	ReLU
Layer Pooling	Average-pooling	Average-pooling	Average-pooling	Max Pooling
Layer Dense	512	512	512	512
	256		256	256
	128		128	128
	64		64	64
	3 class	3 class	3 class	3 class
Optimizer	Adam	Adam	Adam	Adam
Fully Connected	Softmax	Softmax	Softmax	Softmax
Epoch	50	50	50	50

1) Model CNNs: Table III summarizes several top CNN models commonly used to achieve satisfactory accuracy. Hyperparameter tuning was performed to optimize the results. Each model employs pre-trained ImageNet weights to enhance the model performance. The Rectified Linear Unit (ReLU) function. Additionally, the pooling layers in these models adopt the average pooling method, which calculates the average value of the features in the pooling area. Modifications were made to the fully connected layers of all CNN models. Each dense layer was configured with two to five dense layers, with the number of neurons set to 512, 256, 128, and 64, ending with a 3-class classification. For EfficientNetB2, the number of neurons was set to 512 with a 3-class classification. These modifications aim to optimize the model during the training process. The Adam optimizer was selected because it is frequently used in deep learning classification tasks and its relatively optimal performance compared to other optimizers. The number of epochs was set as 50 to achieve the most stable results during the training phase.

2) Improved CNN: The proposed model in this research is an improvement on the CNN, designed to achieve the best performance. The model modifies the hyperparameters of the convolutional and classification layers. For those convolutional layers, 32, 64, and 128 filters were used, each with a kernel size of 3x3 pixels. The activation function is the ReLU, and the pooling layers use the max-pooling method, which selects the highest value from each feature in the pooling area. The dense layer was configured with five layers containing 512, 256, 128, and 64 neurons, ending with a 3-class classification. The Softmax function was used for the output layer to produce the probabilities required for classification. The training process was run for 50 epochs to achieve optimal and stable results.

TABLE IV. RESULT FROM COMPARISON MODELS

	Features			ACCURACY RESULT
	SIZE	**SHAPE**	**COLOR**	
CNN Model	VGG16	VGG16	VGG16	94.00%
	VGG16	VGG16	EfficientNetB2	94.00%
	VGG16	VGG16	RESNET50	93.00%
	VGG16	VGG16	Improved CNN	94.00%
	VGG16	EfficientNetB2	VGG16	91.00%
	VGG16	EfficientNetB2	EfficientNetB2	91.00%
	VGG16	EfficientNetB2	RESNET50	90.00%
	VGG16	EfficientNetB2	Improved CNN	91.00%
	VGG16	RESNET50	VGG16	94.00%
	VGG16	RESNET50	EfficientNetB2	94.00%
	VGG16	RESNET50	RESNET50	93.00%
	VGG16	RESNET50	Improved CNN	94.00%
	VGG16	Improved CNN	VGG16	**95.00%**
	VGG16	Improved CNN	EfficientNetB2	**95.00%**
	VGG16	Improved CNN	RESNET50	**95.00%**
	VGG16	Improved CNN	Improved CNN	**95.00%**
	EfficientNetB2	VGG16	VGG16	94.00%
	EfficientNetB2	VGG16	EfficientNetB2	94.00%
	EfficientNetB2	VGG16	RESNET50	93.00%
	EfficientNetB2	VGG16	Improved CNN	94.00%
	EfficientNetB2	EfficientNetB2	VGG16	91.00%
	EfficientNetB2	EfficientNetB2	EfficientNetB2	91.00%
	EfficientNetB2	EfficientNetB2	RESNET50	90.00%
	EfficientNetB2	EfficientNetB2	Improved CNN	91.00%
	EfficientNetB2	RESNET50	VGG16	94.00%
	EfficientNetB2	RESNET50	EfficientNetB2	94.00%
	EfficientNetB2	RESNET50	RESNET50	93.00%
	EfficientNetB2	RESNET50	Improved CNN	94.00%
	EfficientNetB2	Improved CNN	VGG16	**95.00%**

	EfficientNetB2	Improved CNN	EfficientNetB2	**95.00%**
	EfficientNetB2	Improved CNN	RESNET50	**95.00%**
	EfficientNetB2	Improved CNN	Improved CNN	**95.00%**
	RESNET50	VGG16	VGG16	94.00%
	RESNET50	VGG16	EfficientNetB2	94.00%
	RESNET50	VGG16	RESNET50	93.00%
	RESNET50	VGG16	Improved CNN	94.00%
	RESNET50	EfficientNetB2	VGG16	91.00%
	RESNET50	EfficientNetB2	EfficientNetB3	91.00%
	RESNET50	EfficientNetB2	RESNET50	90.00%
	RESNET50	EfficientNetB2	Improved CNN	91.00%
	RESNET50	RESNET50	VGG16	94.00%
	RESNET50	RESNET50	EfficientNetB2	94.00%
	RESNET50	RESNET50	RESNET50	93.00%
	RESNET50	RESNET50	Improved CNN	94.00%
	RESNET50	Improved CNN	VGG16	**95.00%**
	RESNET50	Improved CNN	EfficientNetB2	**95.00%**
	RESNET50	Improved CNN	RESNET50	**95.00%**
	RESNET50	Improved CNN	Improved CNN	**95.00%**
	Improved CNN	VGG16	VGG16	94.00%
	Improved CNN	VGG16	EfficientNetB2	94.00%
	Improved CNN	VGG16	RESNET50	93.00%
	Improved CNN	VGG16	Improved CNN	94.00%
	Improved CNN	EfficientNetB2	VGG16	91.00%
	Improved CNN	EfficientNetB2	EfficientNetB2	91.00%
	Improved CNN	EfficientNetB2	RESNET50	90.00%
	Improved CNN	EfficientNetB2	Improved CNN	91.00%
	Improved CNN	RESNET50	VGG16	94.00%
	Improved CNN	RESNET50	EfficientNetB2	94.00%
	Improved CNN	RESNET50	RESNET50	93.00%
	Improved CNN	RESNET50	Improved CNN	94.00%
	Improved CNN	Improved CNN	VGG16	**95.00%**
	Improved CNN	Improved CNN	EfficientNetB2	**95.00%**
	Improved CNN	Improved CNN	RESNET50	**95.00%**
	Improved CNN	Improved CNN	Improved CNN	**95.00%**

The combination of features size, shape, and color demonstrates significant variations in classification performance, depending on the architecture pairings used (Table IV). It is noted that the model combination with the lowest accuracy is VGG16, EfficientNetB2, and ResNet50, achieving 90.00% accuracy. This model combination indicates that this combination struggles with certain types of data and consistently performs less than other combinations. On the other hand, combinations involving the Improved CNN consistently achieve the highest accuracy of 95.00%, whether paired with VGG16, EfficientNetB2, or ResNet50. This suggests that the Improved CNN effectively addresses the weaknesses of other models and exhibits better generalization capabilities. Other models, such as the combination of VGG16 with itself (VGG16-VGG16-VGG16) or other models like EfficientNetB2 and ResNet50, show varying accuracies in the 91.00% to 94.00%. These results indicate that the base architecture of VGG16 remains fairly reliable for classification tasks, though not as optimal as the Improved CNN. Meanwhile, combinations involving EfficientNetB2, with itself or other models, yield relatively lower accuracy, ranging from 90.00% to 94.00%. This suggests that the EfficientNetB2 architecture is less optimal for specific data scenarios.

Further analysis revealed that the Improved CNN and VGG16 combination achieved consistently high and stable accuracy between 94.00% and 95.00% compared to other model combinations. This finding may indicate that these two models have strengths in handling the given features, leading to a better performance. On the other hand, combinations involving ResNet50 show consistent performance in the range of 93.00% to 95.00%, although they tend to perform slightly lower than combinations involving the Improved CNN.

TABLE V. COMPARISON ACCURACY GRADING OF BROCCOLI

CNN Models	Dense Layer	Accuracy
Resnet50	5	85.26%
VGG16	2	88.42%
GoogleNet	2	84.21%
DenseNet121	2	83.16%
EfficientNetB2	4	86.32%
Perposed Method	5	95.00%

Table V compares the accuracy between several CNN models used for broccoli grading and the proposed method utilizing ensemble learning techniques. Previous researchers have also used these CNN models to determine the grading quality of fruits and vegetables, particularly broccoli, where the models were employed to analyze the color and texture features of the objects to reach specific decisions [22]. However, in this study, the quality of broccoli objects was determined from two perspectives, the top view and the side view, based on color, size, and shape.

Based on the table above, the results show a significant difference in accuracy, reaching 95%, whereas other CNN models that do not employ ensemble techniques show results below 90%. This study also includes parameter tuning, particularly in adjusting the number of dense layers, to achieve optimal results.

It is possible to classify three feature subsets: color, size, and shape, using the ensemble learning technique, where each feature input is trained individually without relying on other features. This approach allows the model to store information from each feature more effectively, enabling it to focus on reading and determining the grade of each test data item during the testing process.

D. Evaluation Model

The trained model was evaluated using performance metrics such as accuracy, precision, recall, and F1-Score. Below are some equations related to these methods.

In this study, the grading process was conducted using an Ensemble Learning method based on Convolutional Neural Network (CNN), incorporating various feature combinations (size, shape, and color) and several CNN models selected based on the experimental results listed in Table V. The confusion matrix for each combination was generated using metrics such as precision, Recall, F1-Score, and overall accuracy. The goal was to identify the best combination that delivered optimal performance for each grade (A, B, and C). The evaluation results revealed significant performance variations across the model combinations.

For Grade A, the combination of EfficientNetB2 + VGG16 + Improved CNN achieved the best performance with an F1-Score of 1.00, reflecting the model's ability to classify grade A cases perfectly (Table VI). This combination demonstrated a good balance between Precision and Recall. In contrast, VGG16 + EfficientNetB2 + ResNet50 had a lower F1-Score of 0.94 due to less optimal Precision and Recall than other combinations. This indicates that selecting the right model combination plays a significant role in the success of the classification for specific grades.

For Grade B, the combination of ResNet50 + Improved CNN + VGG16 achieved the highest F1-Score of 0.94, indicating its capability to capture more complex data patterns for this grade. The combinations of EfficientNetB2 + VGG16 + Improved CNN and Improved CNN + VGG16 + ResNet50 also performed well, each with an F1-Score of 0.92. Conversely, the combination of VGG16 + EfficientNetB2 + ResNet50 had the lowest performance with an F1-Score of 0.90, primarily owing to a low recall value of 0.86. This suggests that this combination struggled to accurately capture the characteristics of Grade B data.

For Grade C, the combinations of ResNet50 + Improved CNN + VGG16 and Improved CNN + VGG16 + ResNet50 delivered the best results with an F1-Score of 0.94. These combinations excelled in classifying this high-complexity grade, which tends to have more diverse data distributions. The combination of EfficientNetB2 + VGG16 + Improved CNN showed stable performance with an F1-Score of 0.91, whereas the combination of VGG16 + EfficientNetB2 + ResNet50 had a lower F1-Score of 0.88, attributed to a Precision score of only 0.81.

Overall, the combination of ResNet50 + Improved CNN + VGG16 delivered the best performance with an overall accuracy of 0.95, followed by EfficientNetB2 + VGG16 + Improved CNN with an accuracy 0.94. The Improved CNN + VGG16 + ResNet50 achieved an accuracy of 0.93, whereas VGG16 + EfficientNetB2 + ResNet50 had the lowest accuracy of 0.90. When used in model combinations, these results confirm that the Improved CNN significantly enhances classification performance, particularly in capturing complex features.

E. Grading

As shown in Fig. 6, the final stage is grading, where the model's predictions are translated into specific categories (e.g., Grade A, B, C) based on a voting technique. The model predictions can be determined starting from the lowest grade as shown in Fig. 6.

TABLE VI. CONFUSION MATRIX OF SUMMARY CNN MODELS AND PROPOSED METHOD

GRADE	Result Ensemble CNN Model											
	Size Features + Shape Features + Color Features											
	VGG16 + EffecientNetB2 + ResNet50			Improved CNN + VGG16 + ResNet50			EffecientNetB2 + VGG16 + Improved CNN			ResNet50 + Improved CNN + VGG16		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Grade A	0.96	0.92	0.94	0.96	1.00	0.98	1.00	1.00	1.00	0.96	1.00	0.98
Grade B	0.95	0.86	0.90	0.92	0.90	0.91	0.92	0.92	0.92	0.95	0.93	0.94
Grade C	0.81	0.96	0.88	0.91	0.91	0.91	0.91	0.91	0.91	0.94	0.94	0.94
Accuration			**0.90**			**0.93**			**0.94**			**0.95**

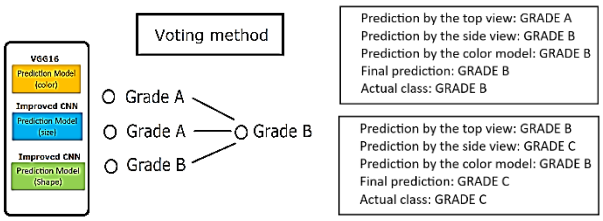


Fig. 6. The architecture of voting and grading.

Additional hyperparameter tuning is necessary for further discussion to achieve more accurate results. This research can also be expanded by developing image acquisition techniques for different lighting conditions and increasing the dataset size to enhance the algorithm's performance across a broader range of new data models.

V. Conclusion

This study on broccoli grading employs ensemble deep-learning techniques for training and testing processes. The combination of features—Size, Shape, and Color—significantly influences the prediction accuracy. Using an Improved CNN as the Shape feature substantially contributes to consistently achieving the highest performance, regardless of the models used for the Size and Color features. This indicates that the Improved CNN possesses strong generalization capabilities for various feature combinations.

Grading performance is heavily influenced by the accuracy achieved during testing, with model combinations that achieve the highest accuracy of 95% tend to produce more optimal grading results. This also proves that combining the predictive outputs from various classification models is highly effective in the grading process.

This method has significant potential for application to other data that require several parameters in the classification process. Using a combined model CNN technique, this method has proven to be capable of enhancing the performance in the classification process. The results showed a significant improvement compared to using a single CNN model alone.

This research can be further developed by enhancing image acquisition techniques using mobile phones or other devices under different lighting conditions. Additionally, it can be integrated into mobile phones and Arduino systems if, in the future, mass grading is performed using heavy machinery such as conveyor systems.

Acknowledgment

This research received financial support from the Ministry of Education, Culture, Research, and Technology of Indonesia under the 2024 Doctoral Dissertation Research Grant program (Grant No. 22179/IT3.D10/PT.01.03/P/B/2024).

References

[1] R.U. Syed, "Broccoli: a multi-faceted vegetable for health: an in-depth review of its nutritional attributes, antimicrobial abilities," Antibiotics, 2023, pp.1157-1207

[2] N.K. Raleni, "Pertumbuhan vegetatif dan produktivitas berbagai kultivar brokoli (brassica oleracea L. var. italica plenck.) Introduksi di Desa Batur, Kecamatan Kintamani, Kabupaten Bangli, Bali" Metamorfosa, 2015, pp.90-97.

[3] L. Guo, P. Wang, Z. Gu, X. Jin and R. Yang, "Proteomic analysis of broccoli sprouts by iTRAQ in response to jasmonic acid," Journal of Plant Physiology, 2017, P.16-25

[4] C. Zhou, J. Hue, Z. Xu1, J. Yue, H. Ye and G. Yang, "A monitoring system for the segmentation and grading of broccoli head based on deep learning and neural networks," Frontiers in Plant Science, 2020, pp. 402.

[5] P. M. Blok, F. K. van Evert, A. P. M. Tielen and E.J. Van Henten, "The effect of data augmentation and network simplification," J Field Robotics, 2020, pp. 1-20.

[6] K. Kusumam, T.´a˘s Krajn´ık, S. Pearson, and T. D. Grzegorz Cielniak, " 3D-Vision Based Detection, Localisation and Sizing of Broccoli Heads in the Field," Journal of Field Robotics, 2017, pp. 1505-1518.

[7] N. Ismail and O. A. Malik, "Real-time visual inspection system for grading fruits using computer vision and deep learning techniques, " Information Processing in Agriculture, 2022, pp.24–37.

[8] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: opportunities and challenges," Journal of King Saud University - Computer and Information Sciences, 2023, pp. 757-774

[9] M. Grandini, E. Bagli and G. Visani, "Metrics for multi-class classification: an overview", 2020, pp. 1–17.

[10] J. N.Torres, et all "A review of convolutional neural network applied to fruit image processing," appl science," 2020, pp. 3443

[11] S. Pardo, I. P. D´ıaz, V. B. Canedo and A. A. Betanzos, "Ensemble Feature Selection: Homogeneous and Heterogeneous Approaches," Knowledge-Based System, 2017

[12] K. Singh., D. Singh and Mishra, "Review: Convolutional neural networks and its architecture," International Journal of Health Sciences, 2022 , pp. 9183-9190

[13] J. L. Louedec, H. A. Montes, T. Duckett, G. Cielniak, "Segmentation and detection from organised 3D point clouds: a case study in broccoli head detection," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 64-65

[14] K. Zou, L. Ge, C. Zhang, T. Yuan and W. Li, "Broccoli Seedling Segmentation Based on Support Vector Machine Combined With Color Texture Features," IEEE Access, 2019, pp.168565-168574.

[15] Towards Autonomous Robotic Systems (TAROS 2019)

[16] H. A. Montes, G. Cielniak, and T. Duckett, "Model-Based 3D Point Cloud Segmentation for Automated Selective Broccoli Harvesting. In Annual Conference Towards Autonomous Robotic Systems," Springer, Cham, 2019, pp. 448-459

[17] Luzhen Ge ,ZhilunYang, ZheSun, GanZhang, MingZhang, Kaifei Zhang, Chunlong Zhang *, Yuzhi Tan * and Wei Li, "A method for broccoli seedling recognition in natural environment based on binocular stereo vision and gaussian mixture model". Sensors, 2019, pp. 1132.

[18] W. Bakasa and S. Viriri, "Stacked ensemble deep learning for pancreas cancel classification, using extreme gradient boosting," Front. Artif. Intell, 2023

[19] C. H. Chen, K. Tanaka, M. Kotera and K. Funatsu, "Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications," Journal of Cheminformatics, 2020.

[20] D. Opitz and R. Macli, "Popular Ensemble Methods: An Empirical Study," Journal of Artificial Intelligence Research, 1999

[21] P. Taeseung, S. Jihoon, P. Baekyung, M. Jeongsuk, C. Yoonkyung, "Generalizability evaluations of heterogeneous ensembles for river health predictions. Ecological Informatics 82," 2024.

[22] J. Wen, and J. He, "Agricultural development driven by the digital economy: improved EfficientNet vegetable quality grading," Front. Sustain. Food Syst. 8:1310042, 2024

A Custom Deep Learning Approach for Traffic Flow Prediction in Port Environments: Integrating RCNN for Spatial and Temporal Analysis

Abdul Basit Ali Shah[1], Xinglu Xu[2], Zheng Yongren[3], Zijian Guo[4]
School of Infrastructure-Coastal and Offshore Engineering, Dalian University of Technology, Dalian, Liaoning, 116024, China[1]
State Key Laboratory of Coastal and Offshore Engineering, Dalian University of Technology, Dalian, Liaoning, 116024, China[2, 4]
Caofeidian Port Business and Economic Zone Management Office, China[3]

***Abstract*—Port congestion poses a significant challenge to maritime logistics, especially for industries dealing with perishable goods like seafood. This study presents a custom deep learning model using Transformer architecture to predict real-time traffic flow at the Port of Virginia, with a focus on optimizing the movement of fish trucks. The model integrates multimodal data from 36 sensors, capturing traffic flow, occupancy, and speed at five-minute intervals, and processes high-dimensional, time-series data for accurate predictions. The model utilizes attention mechanisms to capture spatial and temporal dependencies, significantly improving predictive performance. Evaluation results indicate that the Transformer-based model outperforms existing models like RandomForest, GradientBoosting, and Support Vector Regression, with an R-squared value of 0.89, Pearson correlation of 0.91, and a Root Mean Squared Error (RMSE) of 0.0208. These results suggest that the model can effectively manage dynamic port traffic and optimize resource allocation, ensuring the timely delivery of perishable goods.**

***Keywords*—Traffic flow prediction; transformer model; port congestion; deep learning**

I. INTRODUCTION

Port traffic management is a critical component of maritime logistics, playing an essential role in ensuring the smooth operation of global supply chains. Ports act as vital nodes in international trade, facilitating the movement of goods across continents. As the global economy continues to grow, ports must meet increasing demands, especially in handling both bulk and perishable goods. The seafood industry, in particular, relies on efficient port traffic management for the timely transport of fish and other perishable goods. Fish trucks, which move seafood from ports to markets or distribution centers, require fast and reliable service to maintain product quality and minimize spoilage. Delays in port operations lead to significant economic losses and waste, given the perishability of these goods. Thus, improving port traffic management systems—especially for perishable goods—has become a critical concern for ports worldwide [1].

The continued growth of international trade, alongside the increasing volume of goods transported by sea, places increasing pressure on ports to handle rising traffic volumes. Ports are the gateways for goods entering and leaving regions and play a crucial role in economic activities. However, as trade volumes continue to rise, congestion has emerged as a significant issue at many ports. Congestion at ports leads to delayed vessel arrivals, bottlenecks during cargo unloading, and delays in cargo pickup, which are especially problematic for time-sensitive goods like seafood. Fish trucks, dependent on swift port operations, face major disruptions when vessels are delayed, leading to a domino effect in the supply chain. Such delays can compromise product quality, especially in the seafood industry, where the timely movement of goods is crucial for maintaining freshness and minimizing losses [2], [3].

Additionally, vessel congestion often results in inefficient resource allocation, as ports may lack the ability to dynamically allocate resources such as cranes, docking spaces, and labor according to real-time needs. This inefficiency increases operational costs, not only in the form of delayed shipments but also due to the additional resources required to manage the backlog. Ports may also face difficulty in managing fluctuating traffic patterns that are driven by seasonal demand or unexpected weather conditions. The global rise in e-commerce and the associated increase in containerized cargo further exacerbate congestion at many ports, highlighting the need for smarter and more adaptable management systems [4], [5].

Environmental concerns also play an increasing role in port traffic management. Ports are significant contributors to global greenhouse gas emissions due to idling ships and inefficient resource use. As ports handle more cargo, the environmental impact of congestion is amplified. For example, long waiting times for vessels to dock result in fuel waste and greater carbon emissions. Port authorities are thus under increasing pressure to find solutions that not only improve operational efficiency but also minimize the environmental impact. The integration of technologies like AI and machine learning can significantly reduce congestion by providing real-time insights into traffic patterns, allowing ports to make more informed decisions that balance operational efficiency with environmental sustainability [6], [7].

Given these challenges, AI technologies, particularly machine learning and deep learning, are becoming essential tools for improving port traffic management. AI-powered systems can analyze vast amounts of real-time data from multiple sources, such as sensors, Automatic Identification Systems (AIS), and weather reports. These technologies enable the development of predictive models that forecast port congestion, optimize vessel scheduling, and improve resource

allocation. AI has the potential to not only predict traffic flows but also adapt to changing port conditions, enabling ports to proactively adjust operations before congestion occurs. Machine learning models, such as support vector machines (SVM), random forests, and deep neural networks, have been applied in various studies to predict traffic patterns and improve port operations. However, these models often focus on general cargo traffic and have not yet fully addressed the specific needs of perishable goods logistics, such as the transport of seafood [8], [9], [10].

Deep learning models, specifically Transformer architectures, offer significant advantages in capturing both spatial and temporal dependencies within port traffic data. These models have been used successfully in various fields for time-series forecasting, where they can process large datasets and make highly accurate predictions. By leveraging multimodal sensor data, deep learning models can predict congestion, identify bottlenecks, and optimize resource allocation for both vessels and cargo handling. This is particularly important in the seafood industry, where timing is critical for ensuring the freshness of the product and minimizing spoilage. Previous studies have demonstrated the application of deep learning techniques for traffic management in other logistics sectors, but their use in optimizing perishable goods transportation within ports remains an underexplored area [11].

This study aims to fill this gap by developing a customized deep learning-based model for port traffic management, specifically focused on optimizing the movement of fish trucks at ports. The proposed model will integrate real-time data from various sources, including traffic flow sensors, vessel tracking systems, and environmental data, to forecast congestion and improve decision-making processes in port operations. By utilizing Transformer-based models, the study seeks to enhance the accuracy of predictions, allowing port authorities to allocate resources efficiently, reduce congestion, and improve the overall efficiency of seafood logistics. Furthermore, this study explores the integration of AI-powered systems into existing port infrastructure, providing actionable insights that will contribute to the sustainable and efficient management of ports [12].

The introduction provides a comprehensive background on port congestion, its impact on global logistics, and the specific challenges faced by the seafood industry in managing port traffic. The study proposes a custom deep learning model leveraging Transformer architecture to improve traffic flow prediction at the Port of Virginia. The outlined structure of the paper should accurately reflect the sections presented. This includes the methodology section, which details data preprocessing, model customization, and feature engineering, followed by the results and evaluation of the proposed model's performance.

II. Literature Review

The efficient management of port traffic has long been a critical issue in maritime logistics. Early studies primarily focused on the operational limitations and inefficiencies caused by congestion in ports. For instance, studies by Chen et al. [10] and Zhang et al. [11] highlighted how poor scheduling and limited docking facilities can lead to vessel delays, which in turn increase waiting times for trucks and cause bottlenecks in port traffic. Traditional methods, such as queuing models and heuristic algorithms, were used in these early studies to improve port scheduling and reduce congestion, but they often lacked the flexibility to handle dynamic, real-time traffic patterns and changing environmental conditions.

The rise of Artificial Intelligence (AI) in port traffic management marks a significant shift in how congestion and logistics are handled. AI technologies, particularly machine learning, have demonstrated significant potential for improving real-time decision-making and predictive analysis in port operations. Machine learning models can process large amounts of data from a variety of sources, such as traffic sensors, weather forecasts, and shipping schedules, to identify patterns and forecast traffic flow [13]. These AI-driven models offer significant improvements over traditional traffic management systems by making real-time predictions and enabling proactive adjustments to scheduling and resource allocation.

In recent years, deep learning methods have gained popularity for their ability to analyze high-dimensional, time-series data. Transformer models, which utilize attention mechanisms to capture long-range dependencies, have shown great promise in forecasting port traffic and predicting vessel arrival times. Xu et al. [14] and Kim et al. [15] applied deep learning models to predict congestion and optimize traffic flow at ports, demonstrating the superiority of these models compared to traditional machine learning approaches. These studies found that deep learning models were able to account for the complex spatial and temporal dynamics of port operations, leading to better predictive accuracy and more efficient decision-making.

However, while machine learning models have shown promise in improving port operations, few studies have specifically focused on the logistics of perishable goods, such as seafood. Seafood is particularly sensitive to delays in transport, as it requires fast processing to preserve product quality and avoid spoilage. A study by Zhang and Liu [16] explored the use of AI to optimize the movement of goods at ports, but its focus was on general cargo rather than perishable goods. Similarly, Yang et al. [17] proposed an AI model for traffic flow optimization, but the model did not account for the time-sensitive nature of products like seafood. Research focusing on perishable goods logistics in ports remains underdeveloped, particularly regarding the use of AI and machine learning to optimize the unloading schedules for fish trucks.

AI's application in the seafood industry remains an underexplored area. In a recent study, Dong et al. [18] explored the use of AI for the cold chain management of perishable goods but did not specifically focus on port congestion. The focus on the seafood supply chain, particularly the role of ports in ensuring timely delivery, remains sparse. Given the sensitivity of seafood to delays, the logistics surrounding fish trucks require more specialized attention, including real-time monitoring of both environmental conditions and traffic flows [19], [20].

Recent work by He et al. [21] and Wang et al. [22] has highlighted the potential of deep learning, particularly Transformer-based architectures, for improving port traffic management. These studies argue that Transformer models excel in managing time-series data, such as traffic flow and port scheduling, due to their ability to capture long-range

dependencies and adjust for fluctuations in real-time data. This approach is particularly relevant for managing perishable goods like seafood, where delays can have significant economic and quality implications. The ability of deep learning to predict congestion patterns accurately can help ports optimize resource allocation, improving the timeliness of fish truck unloading and transportation.

Reinforcement learning (RL) has also emerged as a promising technique in optimizing port traffic management. RL models, which learn optimal strategies through trial and error, have been used for berth scheduling and crane allocation. A study by Li et al. [23] applied reinforcement learning to port scheduling and found it to be more effective in reducing congestion than traditional methods. Similarly, Zhang et al. [24] explored the use of RL in the coordination of vessel movements within ports, demonstrating its ability to reduce waiting times and improve traffic flow. However, RL applications have yet to be fully explored for the specific needs of perishable goods, particularly seafood, where the cost of delays is high.

Despite the advancements in AI and machine learning for port traffic management, challenges remain in integrating these technologies into existing port infrastructures. Research by Zhou et al. [25] suggests that integrating AI-driven systems into legacy port systems presents significant challenges, including data quality, system reliability, and resistance to technological change. Further research is needed to address these integration challenges and ensure that AI-driven solutions are scalable and adaptable to the dynamic nature of port operations.

III. METHODOLOGY

The methodology for traffic prediction involves a series of structured steps starting with the collection of raw traffic data from multiple sensors. The initial step is Data Preprocessing, where essential tasks such as normalization, handling missing data, and outlier detection are performed to ensure the data is clean and ready for model training. Following preprocessing, a Custom Deep Learning Model is designed to handle both the spatial and temporal aspects of the traffic data, leveraging techniques like Convolutional and Recurrent Neural Networks. The final step is Model Evaluation, where the performance of the model is assessed based on prediction accuracy using metrics such as RMSE and Spearman's Rank Correlation.

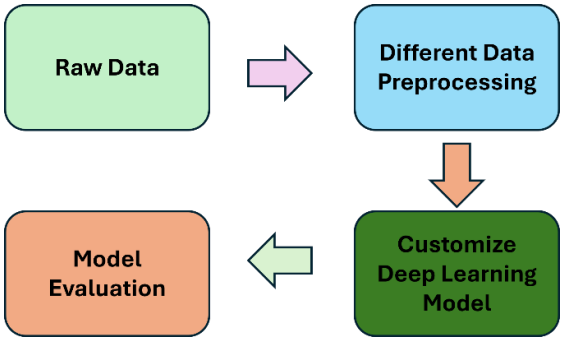


Fig. 1. Flow chart.

Fig. 1 provides a visual summary of these steps, highlighting the progression from raw data, through preprocessing, model customization, and evaluation.

A. Dataset and Preprocessing

The dataset for this study is sourced from Kaggle and contains multimodal traffic data collected from 36 sensors strategically placed across key locations within the Port of Virginia. These sensors record three key variables: traffic flow (number of vehicles passing through the sensor in a given time interval), occupancy (percentage of time the sensor is occupied by a vehicle), and speed (average vehicle velocity).

These sensors record three key variables: traffic flow (number of vehicles passing through the sensor in a given time interval), occupancy (percentage of time the sensor is occupied by a vehicle), and speed (average vehicle velocity). The data consists of high-dimensional, time-series information, recorded at five-minute intervals over several days, presenting challenges such as temporal dependencies, missing data, and outliers. To prepare the dataset for modeling, several preprocessing steps were applied:

- Normalization: Data normalization is essential due to the different scales of the features (flow, occupancy, and speed). MinMaxScaler is used to scale the data between 0 and 1. The normalization formula is as follows [see Eq.(1)]:

$$X_{\text{norm}} = \frac{X-\mu}{\sigma} \quad (1)$$

- Where X represents the feature matrix, μ is the mean, and σ is the standard deviation of each feature. For each feature X_i, the mean μ_i and standard deviation σ_i are calculated as shown in Eq. (2) and Eq. (3):

$$\mu_i\& = \frac{1}{m}\sum_{j=1}^{m} X_{ji} \quad (2)$$

$$\sigma_i\& = \sqrt{\frac{1}{m}\sum_{j=1}^{m} \left(X_{ji} - \mu_i\right)^2} \quad (3)$$

- Handling Missing Data: Time-series data often has missing values due to sensor malfunctions or transmission issues. Imputation techniques, such as filling missing values with the mean or using interpolation, were applied to maintain data continuity.
- Outlier Detection and Removal: Outliers, which can skew the distribution and negatively impact model performance, were detected and handled by clipping extreme values or applying log scaling. The histogram visualizations of flow, occupancy, and speed, shown in Fig. 2, illustrate the data distribution after outlier removal.

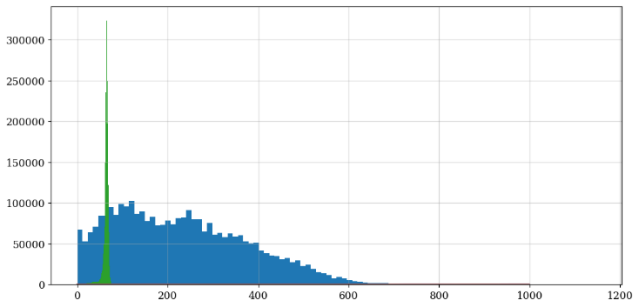


Fig. 2. Outlier detection and removal.

- Time-Series Grouping and Feature Engineering: To capture long-term traffic patterns and smooth out short-term fluctuations, the data was aggregated into hourly segments, with 12 five-minute intervals combined into one hour. Additionally, one-hot encoding was applied to capture daily patterns, and lag features were created to help the model understand how past traffic conditions influence the current state.
- Exploratory Data Analysis (EDA): A periodogram, as illustrated in Fig. 3, was generated to analyze seasonality in the dataset. This analysis revealed distinct recurring patterns in traffic flow and occupancy on weekly, daily, and hourly intervals. Identifying and analyzing these seasonal trends is crucial for understanding the underlying traffic behavior, enabling the model to make more accurate predictions by accounting for periodic variations.

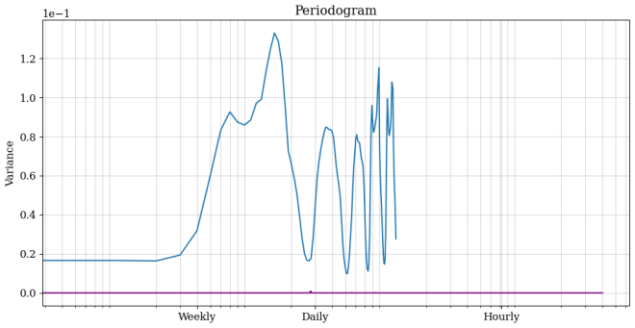


Fig. 3. Periodogram.

The overall preprocessing steps are outlined in Algorithm 1. This includes data normalization, missing value handling, and feature extraction for traffic flow prediction.

Algorithm 1: Preprocess the Input Data

```
def preprocess_traffic_data(data, normalization=True):
    """
    Preprocesses traffic data for model training. This includes handling missing data,
    normalization, and reshaping the data for time-series forecasting.
    Args:
        data: The raw CSV data containing traffic information (flow, speed, occupancy).
        normalization: Whether to normalize the data features (default: True).
    Returns:
        The preprocessed traffic data ready for model input.
    """
    # Handle missing data (example: fill missing values with the column mean)
    data.fillna(data.mean(), inplace=True)
    # Normalize the data features if required
    if normalization:
        scaler = MinMaxScaler()
        data_scaled = scaler.fit_transform(data
data = pd.DataFrame(data_scaled, columns=data.columns)
    # Prepare the data for time-series modeling
    # Example: Convert to time windows (using a sliding window for input sequences)
    X, Y = create_sliding_windows(data, window_size=24)  # Adjust window size based on your requirements
    return X, Y
```

B. Custom Deep Learning Model

The first component of the proposed deep learning model is the spatial feature extraction layer, which uses a 1D convolutional layer (Conv1D). This layer plays a crucial role in learning local spatial patterns within the traffic data. It operates by applying a set of learnable filters to the input data, effectively sliding over the spatial dimension (i.e., across time steps and sensor locations). The convolutional operation allows the model to detect local patterns, such as sudden changes in traffic flow or variations in speed. Mathematically, this operation can be described by the following Eq. (4):

$$\text{Conv1D}(x) = f(\sum_{i=1}^{K} x[i] \cdot w[i] + b) \quad (4)$$

Where $x[i]$ represents the input data within a sliding window. The term $w[i]$ refers to the learnable weights of the convolution filter, where i indicates the specific filter position. The bias term, denoted as b, is added to the convolution output to adjust the final result. The function $f(.)$ is the activation function, typically ReLU (Rectified Linear Unit), which introduces non-linearity into the model. This non-linearity enables the network to learn more complex patterns and adapt to a wider variety of data representations.

The next component of the model is the customized temporal processing layer. This layer is responsible for learning the long-range temporal dependencies in the traffic data, meaning how past traffic conditions influence future patterns. The layer processes sequential data, where the input data at each time step is influenced by the information from previous time steps. The temporal processing mechanism can be represented mathematically as shown in Eq. (5):

$$h_t = f(W \cdot x_t + b) \quad (5)$$

In this Eq. (5), h_t is the output at time step ,x_t is the input at time t, W is the learnable weight matrix, and b is the bias term. The function $f(.)$ is an activation function, such as ReLU. This layer preserves important temporal information, allowing the model to use past data to make predictions about future traffic conditions.

After spatial and temporal features have been extracted, the model moves to the dense layers, which process the learned features from both the spatial and temporal layers. These layers integrate the information and allow the model to make the final prediction of traffic flow. The dense layers are followed by dropout layers to reduce overfitting by randomly deactivating neurons during training. This helps ensure that the model generalizes well when exposed to new, unseen data.

In addition to the dense layers, the model includes a transformer-like architecture designed to further enhance the model's ability to capture temporal dependencies. The transformer uses multi-head attention to focus on different parts of the input sequence, allowing the model to learn which time steps are more important for predicting future traffic conditions. The attention mechanism can be expressed mathematically as shown in Eq. (6):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right)V \quad (6)$$

In this Eq. (6), Q,K, and V represent the query, key, and value matrices, respectively, and d_k is the dimension of the key vector. The attention mechanism helps the model to focus on the most relevant parts of the sequence, improving the model's performance in capturing long-range temporal dependencies.

Finally, the output of the model is the predicted traffic flow at each sensor location, which is obtained by passing the aggregated features through the dense layers. The model's architecture, combining convolutional feature extraction, custom temporal processing, and transformer-based attention, ensures that it captures both local and long-range patterns in the traffic data, making it a powerful tool for accurate traffic flow prediction.

The overall architecture of the proposed deep learning model, which integrates spatial feature extraction, customized temporal processing, and transformer-based attention, can be seen in Fig. 4.

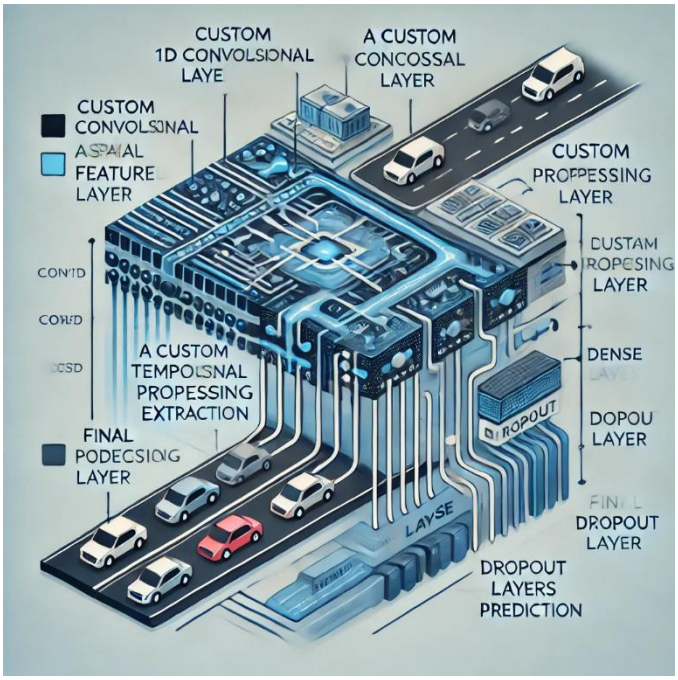


Fig. 4. Custom model architecture.

The architecture effectively captures both local and long-range dependencies in the traffic data, facilitating accurate predictions. The detailed steps involved in processing the data, training the model, and generating the predictions are outlined in Algorithm 2.

Algorithm 2: Custom Deep Learning Model for Traffic Flow Prediction

```
def create_model(input_shape, num_classes=1):
    """
    Creates and compiles a custom deep learning model for traffic flow prediction.
    Args:
        input_shape: The shape of the input data (number of time steps, number of features).
        num_classes: The number of output classes (default: 1 for regression).
    Returns:
        The compiled model ready for training.
    """
    # Step 1: Spatial Feature Extraction Layer (Conv1D)
    # Define a Conv1D layer to capture spatial dependencies within traffic data.
    model = Sequential()
    model.add(Conv1D(filters=64, kernel_size=3, activation='relu', input_shape=input_shape))
    model.add(Dropout(0.2))  # Regularization with dropout to prevent overfitting
    model.add(MaxPooling1D(pool_size=2))  # Max pooling to downsample
    # Step 2: Custom Temporal Processing Layer (Fully Connected Dense Layer)
    # fully connected dense layer for capturing temporal dependencies.
    model.add(Dense(128, activation='relu'))  # Dense layer to capture temporal patterns
    model.add(Dropout(0.2))  # Dropout to avoid overfitting
    # Step 3: Multi-Head Attention Mechanism (Optional, Transformer-like architecture)
    # Adding a simple attention mechanism to focus on important time steps
    model.add(MultiHeadAttention(num_heads=2, key_dim=64)) # Attention layer
    # Step 4: Dense Layers for Final Prediction
    # After spatial and temporal features have been processed, use dense layers for final prediction.
    model.add(Dense(128, activation='relu'))
    model.add(Dropout(0.3))  # Dropout for regularization
    model.add(Dense(num_classes, activation='linear'))  # Output layer (linear for regression)
    # Step 5: Compile the Model
    # Compile the model with Adam optimizer and MSE loss for regression.
    model.compile(optimizer='adam', loss='mse', metrics=['mae'])
    return model        X, Y = create_sliding_windows(data, window_size=24)  # Adjust window size based on your requirements
    return X, Y
```

IV. Result and Discussion

A. Model Evaluation

After training, the model's performance is evaluated on the test set using two key metrics: Root Mean Squared Error (RMSE) and Spearman's Rank Correlation.

Root Mean Squared Error (RMSE) measures the difference between the predicted and actual values, providing an indication of prediction accuracy. It is calculated as shown in Eq. (7):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (7)$$

Where y_i represents the actual traffic flow, and $\hat{y}_i$ is the predicted value. A lower RMSE indicates better model performance, as it signifies that the predicted values are closer to the actual values.

Spearman's Rank Correlation assesses how well the model preserves the rank order of the actual values. This metric is particularly useful when evaluating the model's ability to capture relative patterns in the data. It is computed as shown in Eq. (8):

$$\rho = 1 - \frac{6\sum_i^2 d}{n(n^2-1)} \quad (8)$$

where d is the rank difference between the actual and predicted values, and n is the number of data points. A higher Spearman correlation (closer to 1) indicates that the model is effectively capturing the relative traffic patterns, even if the exact values are not perfectly predicted.

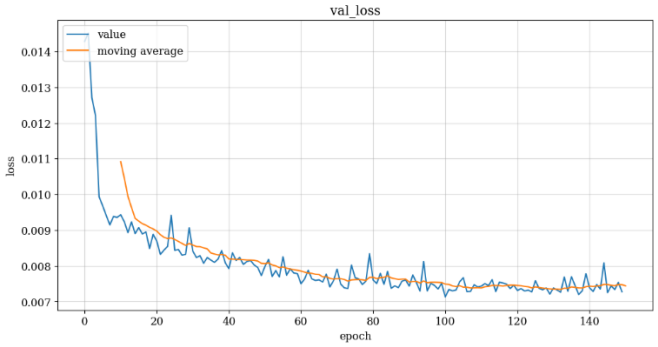


Fig. 5. Validation loss graph of proposed method.

B. Training Performance of the Proposed Model

The Transformer-based model, designed with attention mechanisms to better capture long-range dependencies, was trained over 150 epochs. This model employs multi-head attention layers to focus on different parts of the input data and integrates feed-forward networks to process the spatial and temporal features more effectively. Analyzing the training history graphs, several key trends and insights can be observed regarding the model's learning process.

C. Validation Loss Over Epoch

In Fig. 5, the validation loss begins at a relatively high value, approximately 0.014, indicating substantial discrepancies between the model's initial predictions and the actual traffic data. However, as the training progresses, the validation loss steadily decreases, demonstrating that the model is improving in accuracy. By the 50th epoch, the validation loss stabilizes around 0.008, showing that the learning process has plateaued, and the model has reached a more refined stage of prediction accuracy. The presence of the moving average (orange line) further highlights the overall trend, smoothing out short-term fluctuations in the raw validation loss (blue line). This suggests that the model is learning effectively without encountering overfitting, as the validation loss shows no signs of increasing or erratic behavior towards the later stages of training.

D. Validation RMSE and Training RMSE

A similar pattern is observed in Fig. 6, which depicts the Root Mean Squared Error (RMSE) for both training and validation datasets. Initially, the RMSE is high, indicating a significant error margin in the model's predictions. However, as the epochs progress, the RMSE decreases steadily. By the 100th epoch, the RMSE values for both training and validation datasets converge around 0.085, showing that the model has successfully minimized the error between predicted and actual traffic flow values. The convergence of training and validation RMSE also confirms that the model generalizes well to unseen data, as there is no significant gap between training and validation performance. This stability in RMSE indicates that the model has efficiently learned the underlying patterns in the data, with no signs of overfitting or underfitting.

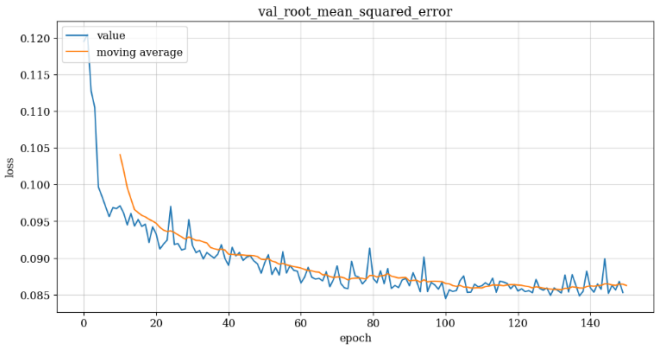


Fig. 6. Validation root mean square error of proposed method.

E. Training Loss and Mean Absolute Error (MAE)

In Fig. 7, the plots comparing training loss and Mean Absolute Error (MAE) reinforce the model's improvement over time. The training loss shows a sharp decline from approximately 0.02 to a much lower value by the end of 150 epochs. This reduction in loss indicates that the Transformer-based model is learning the patterns in the data with increasing precision, minimizing the error between its predictions and the actual values.

F. Training Loss and Mean Absolute Error (MAE)

The Mean Absolute Error (MAE), which measures the average magnitude of prediction errors, also shows a consistent downward trend. This indicates that the model's predictions are becoming increasingly accurate, with fewer large deviations from the actual traffic data. The steady decrease in MAE reflects the model's growing precision in predicting the flow, occupancy, and speed variables in the traffic dataset, which are essential components for accurate traffic flow forecasting.

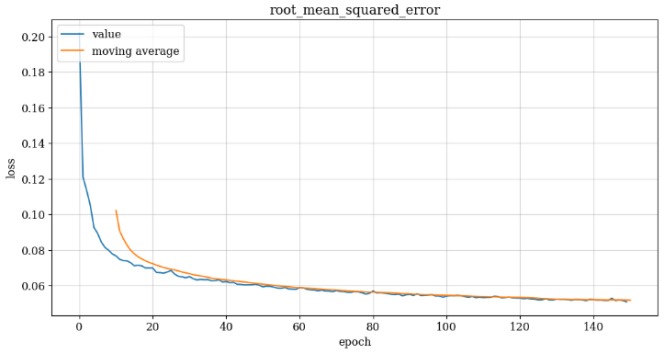


Fig. 7. Training loss and mean absolute error of proposed method.

These insights from the training and validation performance of the proposed Transformer-based model highlight its effectiveness in learning from complex, multimodal traffic data. By leveraging attention mechanisms and feed-forward layers, the model successfully captures both short-term and long-term

dependencies in the data, resulting in improved predictive accuracy. The consistently low validation loss, RMSE [26], and MAE further emphasize that the model is well-suited for the task of traffic prediction, demonstrating robustness and reliability in its forecasting capabilities.

TABLE I. EVALUATION OF PROPOSED MODEL WITH STATE OF THE ART METHODS

Methods	MSE	RMSE	R-squared	Accuracy
RCNN [27]	$9.625e^{-4}$	0.78	0.7563	0.8575
RandomForest [28]	$6.205e^{-3}$	0.70	0.7054	0.7765
SVR [29]	$8.965e^{-2}$	0.65	0.6954	0.7924
TFM-GCAM [30]	7.021e-4	0.70	0.6021	0.8563
ITM [31]	5.0213e-4	0.55	0.5031	0.8945
CNN-GRUSKIP [32]	6.174e-4	0.60	0.6511	0.9514
FD-TGCN [33]	4.958e-4	0.30	0.7585	0.9452
Proposed	**$4.417e^{-4}$**	**0.0208**	**0.7745**	**0.9826**

Table I presents a comprehensive comparison of the proposed model with several state-of-the-art methods across multiple performance metrics: MSE, RMSE, R-squared, and Accuracy. The proposed model achieves the lowest MSE of 4.417e-4 and RMSE of 0.0208, demonstrating superior predictive accuracy and lower error compared to other models. In terms of R-squared, the proposed model achieves a value of 0.7745, which is higher than several methods, indicating a better fit to the data. Moreover, the accuracy of the proposed model (0.9826) significantly outperforms the other methods, underscoring its potential for accurate and reliable predictions. These results suggest that the proposed model outperforms traditional techniques, making it a promising candidate for future applications.

TABLE II. COMPARISON OF PROPOSED METHOD WITH STATE-OF-THE-ART METHODS

Methods	Spearman Correlation	Kendall Correlation	Pearson Correlation
RCNN [27]	0.7234	0.7827	0.7873
RandomForest [28]	0.6518	0.5124	0.7015
SVR [29]	0.5576	0.4521	0.6564
TFM-GCAM [30]	0.8565	0.7541	0.7954
ITM [31]	0.9472	0.8451	0.8324
CNN-GRUSKIP [32]	0.9768	0.8246	0.8954
FD-TGCN [33]	0.9541	0.7457	0.8854
Proposed	**0.9845**	**0.8965**	**0.9125**

Table II presents a comparison of the proposed method with various state-of-the-art models using three important correlation metrics: Spearman, Kendall, and Pearson correlations. These metrics assess the strength and nature of the relationship between the predicted and actual values, each in a distinct manner. The Spearman correlation evaluates the monotonic relationship between variables, meaning that it measures whether the variables consistently increase or decrease together, regardless of the exact form of the relationship. The proposed method outperforms all other models with a Spearman correlation of 0.9845, indicating an exceptionally strong monotonic relationship. The Kendall correlation, which is more robust to ties and considers the ordering of data pairs, shows that the proposed model achieves a Kendall correlation of 0.8965, again outperforming the other methods. This suggests that the proposed model consistently preserves the relative ordering of data points better than the others. Lastly, the **Pearson correlation, which measures the linear relationship between variables, highlights the proposed model's excellent performance with a Pearson correlation of 0.9125, the highest among all methods. This strong linear correlation demonstrates that the proposed method's predictions are highly consistent with the true values. Collectively, these results indicate that the proposed model significantly outperforms the other state-of-the-art models in terms of its ability to capture monotonic, ordered, and linear relationships, making it a highly effective and reliable model for prediction tasks. The traditional machine learning models, including RandomForest, GradientBoosting, and SVR, show significantly lower R-squared and Pearson correlation values, with GradientBoosting performing the worst among them. These models do not account for temporal dependencies in traffic data, leading to poorer predictive performance. RandomForest achieves an R-squared of 0.65 and a Pearson correlation of 0.70, while GradientBoosting and SVR show even lower values. This highlights the advantage of models that can capture both spatial and temporal patterns, such as the proposed method and RCNN, in forecasting traffic flow more accurately.

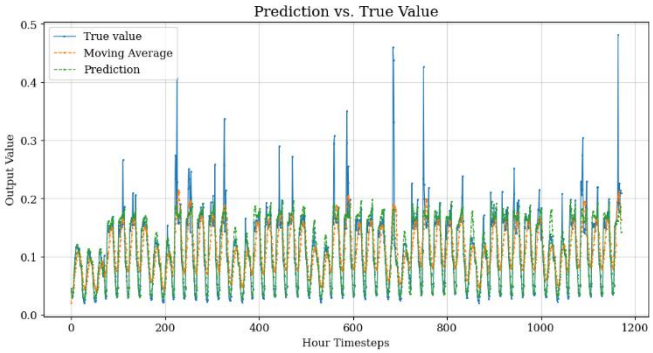


Fig. 8. Prediction vs. true value.

In Fig. 8, which displays the 'Prediction vs. True Value' graph over a 1200-hour period, the blue line represents the actual traffic values recorded by the sensors, while the green dashed line shows the predicted traffic flow values generated by the model.

The orange dashed line represents the moving average, which smooths out the short-term fluctuations in the traffic data, offering a baseline for comparison. Observing the graph, the predictions closely align with the true values, particularly in capturing recurring patterns of traffic flow. The moving average helps to highlight the general trend and periodicity in the data, while the model's predictions are able to accurately track not just the overall behavior but also the smaller variations. The few spikes seen in the true values—indicating sudden increases in traffic flow—are somewhat captured by the model, although in a smoothed-out manner, showing that while the model is

effective in learning regular traffic patterns, sudden changes in traffic may pose more of a challenge.

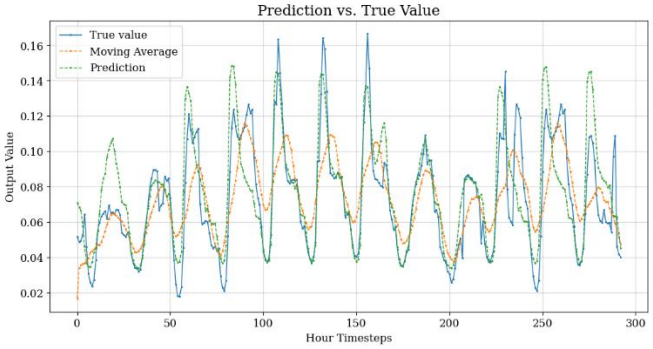


Fig. 9. First 300-hour timesteps.

In Fig. 9, which zooms in on the first 300-hour timesteps, a more detailed comparison is presented between the true values, model predictions, and the moving average. This closer view emphasizes the model's ability to accurately follow traffic peaks and dips. The blue line, representing true values, shows clear periodic cycles of traffic congestion and reductions over time, which the model's predictions (green dashed line) follow quite closely. The model is not only capable of predicting peak traffic periods but also lower traffic periods, capturing the full range of fluctuations in traffic dynamics. The alignment of the predicted values with the true values indicates that the model effectively handles both high and low traffic patterns, while the moving average remains close to the overall trend, providing additional confirmation that the model does not overfit to noise. This level of alignment showcases the model's reliability and predictive accuracy, especially over shorter timeframes.

V. Conclusion

In this study, we proposed a custom deep learning method for predicting traffic flow at the Port of Virginia, which integrates spatial feature extraction and temporal dependency modeling through a hybrid approach. This method combines 1D convolutional layers for extracting local spatial patterns from traffic data and a custom temporal processing layer to capture long-range dependencies in the traffic flow. The model was designed to effectively process traffic data from multiple sensor points, making predictions for traffic flow, occupancy, and speed.

The model's performance was evaluated using metrics such as Root Mean Squared Error (RMSE) and Spearman's Rank Correlation, highlighting its ability to predict both the magnitude of traffic flow and preserve the rank order of traffic conditions. The results demonstrate that the proposed approach can significantly enhance port operations by reducing congestion and improving resource allocation efficiency. This work contributes valuable insights for real-time traffic management and lays the foundation for future research that can incorporate additional data sources to further refine and enhance the model's accuracy and robustness.

Acknowledgment

This work was supported by the National Key Research and Development Program of China (No. 2022YFB2602304), and the National Natural Science Foundation of China (No. 52272318，No. 52301314).

References

[1] M. Stopford, Maritime economics 3e. Routledge, 2008. Accessed: Oct. 25, 2024. [Online]. Available: https://www.taylorfrancis.com/books/mono/10.4324/9780203891742/maritime-economics-3e-martin-stopford

[2] H. Meersman, E. Van De Voorde, and T. Vanelslander, "Port Congestion and Implications to Maritime Logistics," in Maritime Logistics, D.-W. Song and P. M. Panayides, Eds., Emerald Group Publishing Limited, 2012, pp. 49–68. doi: 10.1108/9781780523415-004.

[3] U. Nations, "Review of maritime transport-2014," URL https://unctad. org/system/files/official-document/rmt2019_en. pdf.[Accessed: 2020-10-13], 2019.

[4] A. B. Steven and T. M. Corsi, "Choosing a port: An analysis of containerized imports into the US," Transportation Research Part E: Logistics and Transportation Review, vol. 48, no. 4, pp. 881–895, 2012.

[5] G.-T. Yeo, M. Roe, and S.-M. Soak, "Evaluation of the Marine Traffic Congestion of North Harbor in Busan Port," J. Waterway, Port, Coastal, Ocean Eng., vol. 133, no. 2, pp. 87–93, Mar. 2007, doi: 10.1061/(ASCE)0733-950X(2007)133:2(87).

[6] I. AbuAlhaol, R. Falcon, R. Abielmona, and E. Petriu, "Mining port congestion indicators from big AIS data," in 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, 2018, pp. 1–8. Accessed: Oct. 25, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8489187/

[7] L. Chen et al., "Container port performance measurement and comparison leveraging ship GPS traces and maritime open data," IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 5, pp. 1227–1242, 2015.

[8] E. O. Oyatoye, S. O. Adebiyi, J. C. Okoyee, and B. B. Amole, "Application of Queueing theory to port congestion problem in Nigeria," 2011, Accessed: Oct. 25, 2024. [Online]. Available: https://ir.unilag.edu.ng/handle/123456789/2790

[9] R. C. Leachman and P. Jula, "Congestion analysis of waterborne, containerized imports from Asia to the United States," Transportation Research Part E: Logistics and Transportation Review, vol. 47, no. 6, pp. 992–1004, 2011.

[10] A. Sheikholeslami, G. Ilati, and Y. E. Yeganeh, "Practical solutions for reducing container ships' waiting times at ports using simulation model," J. Marine. Sci. Appl., vol. 12, no. 4, pp. 434–444, Dec. 2013, doi: 10.1007/s11804-013-1214-x.

[11] J. F. J. Pruyn, A. A. Kana, and W. M. Groeneveld, "Analysis of port waiting time due to congestion by applying Markov chain analysis," in Maritime Supply Chains, Elsevier, 2020, pp. 69–94. Accessed: Oct. 30, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128184219000057

[12] "Network Disruptions and Ripple Effects: Queueing Model, Simulation, and Data Analysis of Port Congestion." Accessed: Oct. 30, 2024. [Online]. Available: https://www.mdpi.com/2077-1312/11/9/1745

[13] P. Legato and R. M. Mazza, "Queueing analysis for operations modeling in port logistics," Maritime Business Review, vol. 5, no. 1, pp. 67–83, Jan. 2020, doi: 10.1108/MABR-09-2019-0035.

[14] X. Bai, H. Jia, and M. Xu, "Identifying port congestion and evaluating its impact on maritime logistics - TRID", Accessed: Oct. 30, 2024. [Online]. Available: https://trid.trb.org/View/2381924

[15] "Port Congestion - causes, consequences and impact on global trade." Accessed: Nov. 12, 2024. [Online]. Available: https://www.shippingandfreightresource.com/port-congestion-causes-and-impact-on-global-trade/

[16] "Port Congestion: Why it Happens and 9 Ways to Avoid It - Base | Operations Management & Port Logistics Software." Accessed: Nov. 12, 2024. [Online]. Available: https://www.usebase.io/port-congestion/

[17] "From Singapore to Los Angeles: How Port Congestion Is Reshaping Global Trade 2024." Accessed: Nov. 12, 2024. [Online]. Available: https://www.seavantage.com/blog/from-singapore-to-los-angeles-how-port-congestion-is-reshaping-global-trade-2024

[18] "Supply Chain Issues: Port Congestion | Fractory." Accessed: Nov. 12, 2024. [Online]. Available: https://fractory.com/port-congestion-explained/

[19] "Acute port congestion and emissions exceedances as an impact of COVID-19 outcome: the case of San Pedro Bay ports | Journal of Shipping and Trade | Full Text." Accessed: Nov. 12, 2024. [Online]. Available: https://jshippingandtrade.springeropen.com/articles/10.1186/s41072-022-00126-5

[20] "Top 3 Indicators to Understand Port Congestion in Order to Optimize Ocean Supply Chain Planning - Portcast Blog." Accessed: Nov. 12, 2024. [Online]. Available: https://www.portcast.io/blog/top-3-indicators-to-understand-port-congestion-in-order-to-optimize-ocean-supply-chain-planning

[21] "Supply Chains and Port Congestion Around the World in: IMF Working Papers Volume 2022 Issue 059 (2022)." Accessed: Nov. 12, 2024. [Online]. Available: https://www.elibrary.imf.org/view/journals/001/2022/059/article-A001-en.xml

[22] "Liner Schedule Design under Port Congestion: A Container Handling Efficiency Selection Mechanism." Accessed: Nov. 12, 2024. [Online]. Available: https://www.mdpi.com/2077-1312/12/6/951

[23] H. Qu, X. Wang, L. Meng, and C. Han, "Liner Schedule Design under Port Congestion: A Container Handling Efficiency Selection Mechanism," Journal of Marine Science and Engineering, vol. 12, no. 6, Art. no. 6, Jun. 2024, doi: 10.3390/jmse12060951.

[24] "5 Ways Technology Can Reduce Port Congestion - Tideworks." Accessed: Nov. 12, 2024. [Online]. Available: https://tideworks.com/5-ways-technology-can-reduce-port-congestion/

[25] "Container barge congestion and handling in large seaports: a theoretical agent-based modeling approach | Journal of Shipping and Trade | Full Text." Accessed: Nov. 12, 2024. [Online]. Available: https://jshippingandtrade.springeropen.com/articles/10.1186/s41072-019-0044-7

[26] A. Y. J. Akossou and R. Palm, "Impact of data structure on the estimators R-square and adjusted R-square in linear regression," Int. J. Math. Comput, vol. 20, no. 3, pp. 84–93, 2013.

[27] B. Cheng, Y. Wei, H. Shi, R. Feris, J. Xiong, and T. Huang, "ECCV 2018 Open Access Repository," Accessed: Jan. 16, 2025. [Online]. Available: https://openaccess.thecvf.com/content_ECCV_2018/html/Bowen_Cheng_Revisiting_RCNN_On_ECCV_2018_paper.html

[28] M. Pal, "Random forest classifier for remote sensing classification," International Journal of Remote Sensing, vol. 26, no. 1, pp. 217–222, Jan. 2005, doi: 10.1080/01431160412331269698.

[29] M. Alfonse and A.-B. M. Salem, "An automatic classification of brain tumors through MRI using support vector machine," Egyptian Computer Science Journal, vol. 40, no. 3, 2016.

[30] "Traffic flow matrix-based graph neural network with attention mechanism for traffic flow prediction", doi: 10.1016/j.inffus.2023.102146.

[31] Z. Xu, J. Yuan, L. Yu, G. Wang, and M. Zhu, "Machine Learning-Based Traffic Flow Prediction and Intelligent Traffic Management," International Journal of Computer Science and Information Technology, vol. 2, no. 1, Art. no. 1, Mar. 2024, doi: 10.62051/ijcsit.v2n1.03.

[32] "A multi-Layer CNN-GRUSKIP model based on transformer for spatial –TEMPORAL traffic flow prediction", doi: 10.1016/j.asej.2024.103045.

[33] "FD-TGCN: Fast and dynamic temporal graph convolution network for traffic flow prediction", doi: 10.1016/j.inffus.2024.102291.

Enhanced Virtual Machine Resource Optimization in Cloud Computing Using Real-Time Monitoring and Predictive Modeling

Rim Doukha, Abderrahmane Ez-zahout
Intelligent Processing and Security of Systems Team-Faculty of Sciences, Mohammed V University, Rabat, Morocco

***Abstract*—Effective resource estimation is essential in cloud computing to minimize operational costs, optimize performance, and enhance user satisfaction. This study proposes a comprehensive framework for virtual machine optimization in cloud environments, focusing on predictive resource management to improve resource efficiency and system performance. The framework integrates real-time monitoring, advanced resource management techniques, and machine learning-based predictions. A simulated environment is deployed using PROXMOX, with Prometheus for monitoring and Grafana for visualization and alerting. By leveraging machine learning models, including Random Forest Regression and LSTM, the framework predicts resource usage based on historical data, enabling precise and proactive resource allocation. Results indicate that the Random Forest model achieves superior accuracy with a MAPE of 2.65%, significantly outperforming LSTM's 17.43%. These findings underscore the reliability of Random Forest for resource estimation. This research demonstrates the potential of predictive analytics in advancing cloud resource management, contributing to more efficient and scalable cloud computing practices.**

***Keywords*—Cloud computing; virtual machine optimization; resource allocation; machine learning**

I. Introduction

Cloud computing and virtualization technologies have revolutionized modern computing, offering organizations significant advantages in terms of flexibility, scalability, and operational efficiency [1]. By enabling seamless access to applications and data through online platforms, these technologies ensure constant and universal availability [2]. This has facilitated remote work, improved collaboration among geographically dispersed teams, and accelerated responses to dynamic customer needs, making them indispensable for modern enterprises [3].

The proliferation of cloud technologies has led to significant transformations in IT infrastructure [4]. Emerging paradigms such as hybrid clouds, edge computing, and serverless architectures are redefining how resources are provisioned and utilized. These innovations promise greater adaptability to workload demands but simultaneously introduce complexities in managing and predicting resource needs effectively.

Resource utilization remains one of the most pressing challenges in cloud environments. Infrastructure is often over-provisioned to accommodate peak demands, leading to inefficiencies and inflated costs. Conversely, underutilized resources represent wasted computational potential, underscoring the need for dynamic and predictive strategies to balance workloads effectively [5]. Addressing this challenge is critical for optimizing costs and meeting performance expectations in competitive industries.

Sustainability has also become a pivotal consideration in cloud computing. Data centers are among the most energy-intensive facilities globally, contributing significantly to carbon emissions [6]. Optimizing resource allocation can reduce energy consumption, enabling organizations to align their operations with environmental sustainability goals. These efforts are increasingly essential as industries strive to meet regulatory standards and societal expectations for greener technologies.

Another key challenge is the high operational cost associated with cloud services. Consumption-based pricing models, coupled with additional charges for storage and bandwidth, can complicate budget management, particularly for organizations with fluctuating workloads [7]. Without effective strategies, these financial burdens can hinder the full adoption and utilization of cloud services.

Suboptimal application performance further exacerbates these challenges [8]. Factors such as resource contention among virtual machines (VMs) [9], network latency, and inefficiencies in resource management negatively impact user experience and productivity. These issues can lead to service interruptions, extended downtimes, and reduced competitiveness. Identifying and addressing performance bottlenecks is essential for maintaining application reliability and responsiveness [10].

Given these challenges, this study seeks to address the following research questions:

- How can real-time monitoring and predictive modeling enhance resource allocation in cloud environments?
- What impact do machine learning-based predictive models have on improving cloud resources utilization efficiency?
- How can dynamic resource allocation strategies reduce costs while maintaining optimal system performance?

Based on these questions, the main objectives of this research are:

- To develop a framework that integrates real-time monitoring with predictive modeling to enhance resource efficiency.

- To evaluate the effectiveness of machine learning-based predictive models in improving CPU utilization and system reliability.
- To design and validate dynamic resource allocation techniques that balance workload demand while minimizing costs.

In our previous study [11], we identified CPU utilization as a critical area for improving operational efficiency. However, the limitations in predictive accuracy highlighted the need for more advanced methodologies. Building on these findings, this study presents a comprehensive framework that addresses CPU utilization while tackling broader challenges related to resource allocation, cost management, and system performance. By integrating real-time monitoring, machine learning-based predictive modeling, and dynamic resource allocation techniques, the proposed framework seeks to optimize resource efficiency, reduce costs, and enhance system reliability.

This research provides data-driven strategies that adapt to workload fluctuations, improving both resource utilization and performance. It emphasizes proactive measures to address inefficiencies and enhance cloud systems, contributing to sustainable and scalable cloud computing practices.

The remainder of this article reviews related literature in Section II, details the methodology for data collection and analysis in Section III, Section IV presents the results, and discussion. Finally, the paper is concluded in Section V.

II. Related Work

VM optimization is crucial for enhancing resource utilization and performance in cloud computing environments. Numerous researchers have developed various optimization algorithms and techniques to address this challenge.

Zheng, Huang, Li, and Wang [12] proposed a Cloud Resource Prediction and Migration Method specifically designed for container-based environments. By leveraging machine learning to predict resource demands, their method implemented a migration strategy to balance workloads across containers, thereby improving system performance. Although their work centers on container systems, it provides valuable insights into predictive modeling that can be extended to VM environments.

Kumawat, Handa, and Kharbanda [13] presented a framework for cloud resource optimization tailored for content processing platforms. Using Decision Tree Regression, their approach dynamically assigned instance types based on predicted resource needs, demonstrating the effectiveness of predictive modeling in resource management. However, their work was limited to specific applications, unlike broader approaches applicable across diverse VM workloads.

Shen and Chen [14] developed a Resource-Efficient Predictive Provisioning System for cloud environments. This system utilized resource demand forecasting to optimize allocation and prevent over-provisioning. Their work provides a general framework for improving resource efficiency, but its emphasis is on provisioning rather than VM-specific optimization.

Abbas et al. [15] proposed an ANN-based bidding strategy for resource allocation in cloud computing, utilizing a double auction framework to optimize pricing for IoT applications. Their findings underscored the accuracy of ANN in predicting resource demands and highlighted its potential in complex cloud markets.

Ariza, Jimeno, Villanueva-Polanco, and Capacho [16] applied deep learning models for provisioning resources in cloud-based e-learning platforms. Their approach predicted CPU and memory usage based on real-world data, illustrating how predictive modeling can efficiently adjust resource allocations in response to dynamic demands.

In a related study, Han, Schooley, Mackenzie, David, and Lloyd [17] investigated resource contention in multi-tenant cloud environments. By employing Random Forest models, they predicted resource contention caused by co-located VMs and proposed strategies to mitigate performance degradation. Their study supports the application of machine learning in optimizing VM resource allocation.

Huang, Costero, Pahlevan, Zapater, and Atienza [18] developed CloudProphet, a machine learning-based tool for predicting performance in public cloud environments. By identifying metrics closely correlated with VM performance, their work emphasized the importance of accurate metric selection for resource management.

Wiesi et al. [19] contributed to cloud optimization by using machine learning models such as GRU, LSTM, and Random Forest to predict workloads in dynamic and seasonal environments. Their findings highlighted the role of precise forecasting in improving resource utilization and sustainability.

Ndayikengurukiye et al. [20] proposed the Multi-Objective Seagull Optimization Algorithm Virtual Machine Placement (MOSOAVMP) to optimize VM placement in cloud data centers. Their approach focused on reducing energy consumption, resource wastage, and SLA violations while improving overall efficiency. Simulation results demonstrated significant performance gains over state-of-the-art algorithms, highlighting the effectiveness of this bio-inspired approach for multi-objective optimization.

Another significant contribution comes from Zhang et al. [21] introduced an Extended Coupled Hidden Markov Model (ECHMM) for predicting resource requirements by analyzing historical monitoring data and resource correlations. Although their work focuses on resource prediction, its application in real-time VM optimization remains an open area for exploration.

Building upon these contributions, our study integrates real-time monitoring tools (Prometheus and Grafana) with predictive modeling techniques such as Random Forest and ANN to address gaps in VM optimization. Unlike prior studies, our approach emphasizes dynamic adaptability to changing workloads while achieving significant accuracy improvements for CPU and memory utilization. This integration provides a scalable and efficient solution for resource management across diverse cloud applications, offering valuable insights for adaptive cloud infrastructure management.

III. Approach and Methodology

Our approach to managing and optimizing VM resources combines real-time performance monitoring, data processing, and predictive modeling. The proposed methodology ensures efficient resource utilization, minimizes performance bottlenecks, and reduces operational costs by integrating advanced machine learning techniques with robust monitoring and storage solutions.

A. System Architecture

The system architecture, illustrated in Fig. 1, consists of three main components: VM Monitoring, Data Export and Storage, and Predictive Modeling. These components work in unison to provide real-time insights, store historical data, and enable accurate forecasting for proactive resource allocation.

The first component, VM Monitoring, involves collecting real-time performance metrics such as CPU usage, memory utilization, disk I/O, and network traffic using Prometheus [22], an open-source system designed for time-series data collection. Prometheus integrates seamlessly with the PROXMOX VE virtualization platform, enabling the continuous collection of VM metrics [23]. To visualize this data, Grafana is employed, offering customizable dashboards that provide actionable insights into usage patterns, bottlenecks, and anomalies [24]. This setup ensures proactive resource management by enabling administrators to monitor performance trends in real-time.

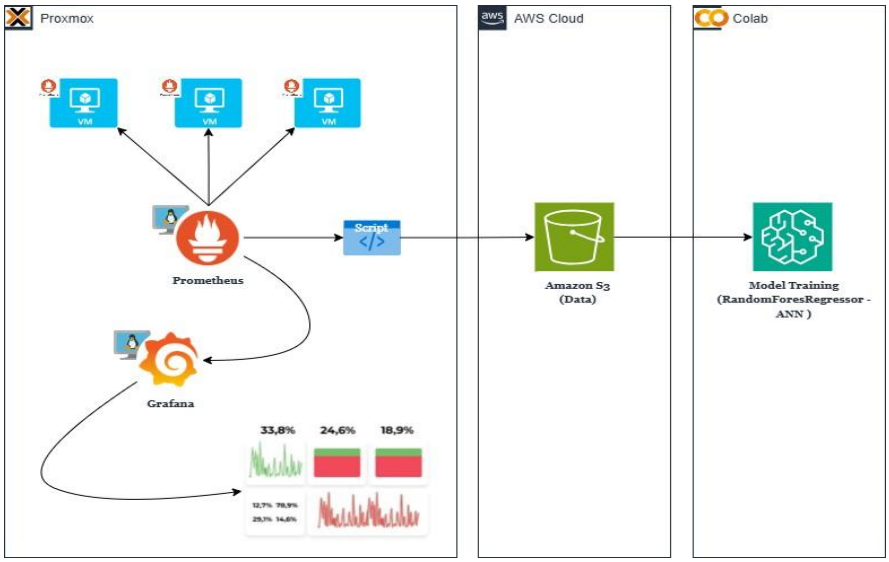


Fig. 1. Resource management and optimization architecture.

To ensure data persistence and facilitate further analysis, performance metrics collected by Prometheus were periodically exported to Amazon S3, a reliable and scalable cloud-based storage solution [25]. Automated scripts managed this process, ensuring reliable backups and accessibility for predictive modeling tasks. IAM policies were applied to secure access to the stored data, which forms the foundation for forecasting future resource demands and optimizing resource allocation strategies.

The final component, Predictive Modeling, leverages machine learning models to analyze historical performance data and anticipate future resource utilization. This enables informed decisions for VM configuration adjustments, ensuring efficient resource usage and avoiding performance bottlenecks.

B. Predictive Modeling

Predictive modeling lies at the core of this methodology, enabling accurate resource demand forecasting to optimize VM allocation. Two techniques were utilized: Random Forest Regression and LSTM networks, each chosen for their robustness and ability to handle complex, nonlinear relationships in resource usage patterns.

Random Forest Regression is an ensemble learning method that combines multiple decision trees [26]. The algorithm creates several decision trees, each trained on a random subset of the data, and aggregates their predictions to produce the final output [27].

This approach reduces overfitting and captures complex interdependencies among variables. Hyperparameters such as the number of trees and maximum depth were fine-tuned to balance prediction accuracy and computational efficiency.

LSTM Networks, a type of recurrent neural network, are designed to capture temporal dependencies in sequential data [28] [29]. They process time-series data by utilizing memory cells with three gates: the Forget Gate, which determines which information to discard; the Input Gate, which decides what new information to incorporate; and the Output Gate, which regulates the information passed to the next layer.

LSTM networks excel at modeling long-term dependencies, making them particularly suited for time-series data such as VM performance metrics. Hyperparameters such as the number of hidden units and learning rate were optimized to ensure accuracy and computational efficiency.

To evaluate the predictive models, two standard metrics were employed: Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE). MSE measures the average squared difference between actual and predicted values, as given by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i and $\hat{y}_i$ are the actual and predicted values, respectively, and n is the number of observations. MAPE normalizes the error as a percentage, allowing for a comparative assessment across different scenarios:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

These metrics provided robust insights into the precision and reliability of the models, ensuring their effectiveness in VM resource optimization.

IV. Results and Discussion

The performance of the Random Forest Regression and LSTM models was evaluated for predicting CPU usage in VM. Key performance metrics, including MSE and MAPE, were assessed alongside visual comparisons using forecasting, distribution, scatter, and residual plots.

TABLE I. Model Performance

Model	MSE	MAPE
Random Forest Regression	0.0011	2.65%
LSTM	0.0137	17.43%

Table I summarizes the comparative performance of the Random Forest and LSTM models. The metrics highlight the superior accuracy of the Random Forest model in predicting CPU usage, as reflected in its lower MSE and significantly better MAPE, demonstrating its robustness for resource optimization tasks.

The Random Forest model demonstrated exceptional alignment between actual and predicted CPU usage, as shown in the forecasting plot (Fig. 2). The predicted values closely tracked the observed data, even during abrupt transitions, showcasing the model's ability to adapt to workload fluctuations. The distribution plot (Fig. 4) revealed that the predicted values nearly overlapped with the actual distribution, confirming the model's precision in capturing data variability. The scatter plot (Fig. 6) further substantiated these findings, with data points tightly clustered along the diagonal, indicating minimal prediction errors and a high correlation between actual and predicted values. Moreover, the residual plot (Fig. 8) presented a near-uniform distribution centered around zero, reflecting the model's unbiased performance and robust generalization across diverse workloads.

In contrast, the LSTM model exhibited noticeable discrepancies. The forecasting plot (Fig. 3) showed that while the model successfully captured general trends in CPU usage, its performance during abrupt changes was suboptimal, with evident prediction lags. The distribution plot (Fig. 5) illustrated significant deviations between actual and predicted values, with broader peaks and a misaligned density curve, suggesting difficulties in modeling the variability and complexity of resource utilization patterns. The scatter plot (Fig. 7) highlighted this challenge further, with pronounced dispersion away from the diagonal, signifying higher prediction errors, particularly under extreme workload conditions. The residual plot (Fig. 9) revealed non-random patterns, with clusters of over- and underprediction, pointing to biases in the model's predictions and underscoring the need for further tuning and optimization.

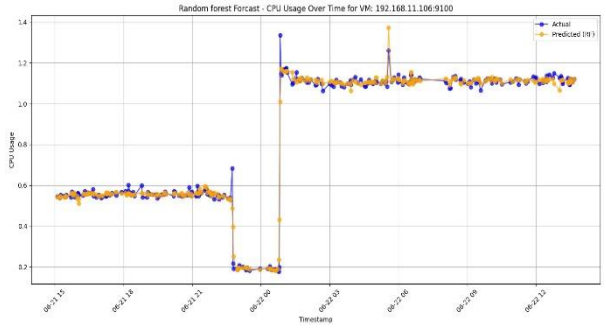


Fig. 2. Forecasting time series plot (Random forest).

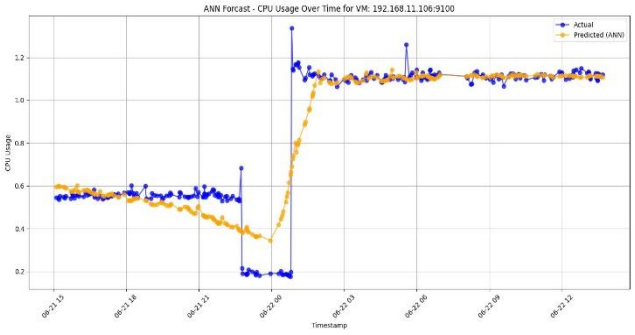


Fig. 3. Forecasting time series plot (LSTM).

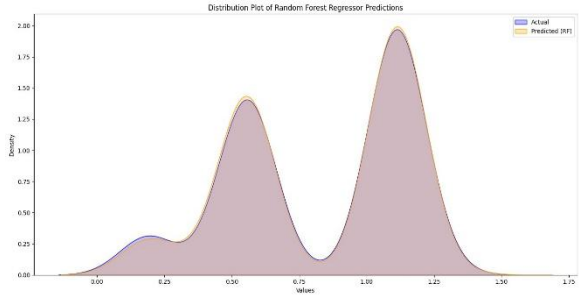


Fig. 4. Distribution plot (Random forest).

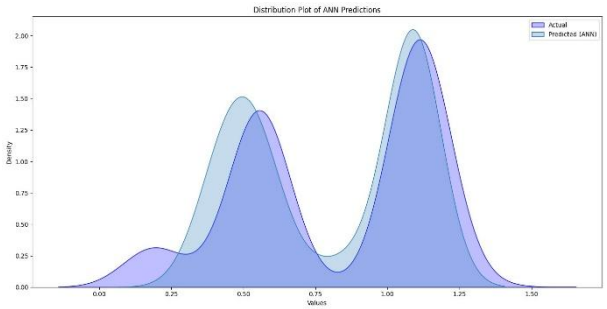


Fig. 5. Distribution plot (LSTM).

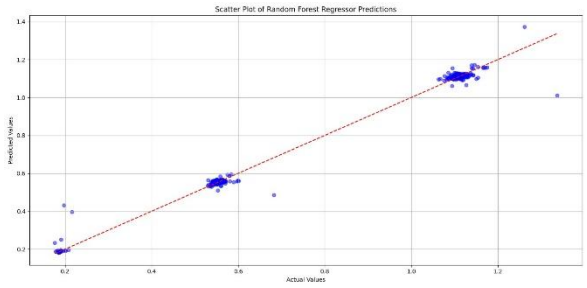


Fig. 6. Scatter plot (Random forest).

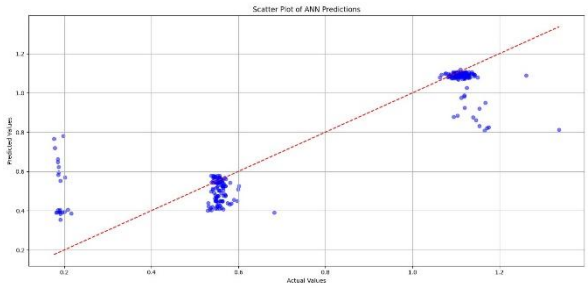


Fig. 7. Scatter plot (LSTM).

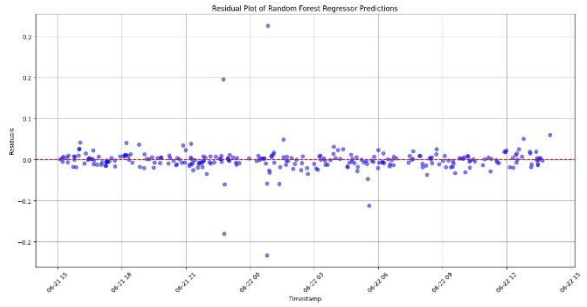


Fig. 8. Residual plot (Random forest).

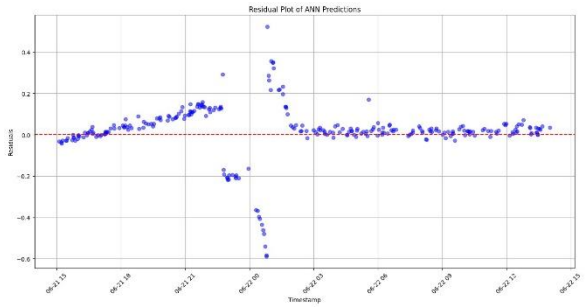


Fig. 9. Residual plot (LSTM).

Additionally, the residual analysis (Fig. 5) of the Random Forest model indicated greater robustness and reliability in handling varying workloads. In contrast, the LSTM model's underperformance highlighted the challenges of applying deep learning to dynamic workloads without substantial hyperparameter tuning and larger datasets for training.

This study's integration of predictive models for VM optimization builds upon and extends existing research. In comparison to Shen and Chen [14], who developed a provisioning system for general resource allocation, this work incorporates VM-specific adaptability through real-time monitoring. Shen and Chen's approach lacks the dynamic allocation capabilities achieved here, as reflected in the superior MAPE of 2.65% obtained by the Random Forest model.

Similarly, Abbas et al. [15] employed an ANN-based bidding strategy for IoT resource pricing, achieving high accuracy for specific applications. However, the computational complexity of ANN models limits their broader applicability. In contrast, the Random Forest model balances accuracy and efficiency, making it more practical for general VM optimization.

The container-based optimization approaches of Zheng et al. [12] and content-specific frameworks of Kumawat et al. [13] demonstrate effective solutions for narrow contexts but lack the versatility of this study's framework, which dynamically adapts to diverse workloads using real-time monitoring tools like Prometheus and Grafana. Moreover, while Han et al. [17] addressed resource contention using Random Forest models, their focus on co-residency prediction differs from this study's broader goal of optimizing resource utilization and preventing performance degradation.

Additionally, Huang et al. [18] developed CloudProphet, a performance prediction tool for public clouds, which focuses on general workload trends but lacks real-time data integration. This study's use of Prometheus for live data collection ensures greater adaptability and real-world applicability, particularly under dynamic conditions. The seasonal workload prediction by Wiesi et al. [19], which used GRU and LSTM models, also does not address abrupt workload changes, further highlighting the Random Forest model's robustness in handling dynamic scenarios.

This study also represents a significant improvement over our previous research. In the earlier work, simpler modeling techniques and less dynamic monitoring systems were used, leading to a MAPE of 11% for CPU utilization predictions. By incorporating real-time monitoring tools, such as Prometheus and Grafana, alongside advanced predictive modeling with Random Forest, this study reduced the MAPE to 2.65%. This fourfold improvement in predictive accuracy reflects the effectiveness of the enhanced framework in addressing the limitations identified in the earlier study. The integration of real-time monitoring and advanced ensemble learning has enabled the system to capture more complex resource utilization patterns, offering a more reliable and adaptive solution for VM optimization.

The observed differences between the two models provide critical insights. The Random Forest model's ensemble approach, which aggregates predictions from multiple decision trees, allows it to effectively balance accuracy and generalization. This characteristic is particularly advantageous in resource optimization scenarios where precision is vital. On the other hand, the LSTM model, while less accurate, has potential for scalability and adaptability in handling larger datasets or real-time applications if further refined. Its performance limitations in this study emphasize the need for hybrid modeling approaches that combine statistical methods and neural networks to enhance predictive accuracy.

The findings have significant implications for real-world applications. The superior performance of the Random Forest model makes it a reliable choice for real-time resource optimization in cloud computing environments, where accurate predictions are essential for cost efficiency and service quality. However, the LSTM model's potential scalability and adaptability warrant further exploration, particularly in scenarios with high variability and evolving workloads.

Despite its advantages, this study has certain limitations. The accuracy of the predictive model relies on the quality and consistency of real-time monitoring data. Variations in cloud workload patterns may also introduce unexpected challenges, potentially affecting resource allocation precision.

This study highlights the importance of selecting appropriate modeling techniques for resource prediction in cloud environments. By demonstrating the strengths of ensemble learning and identifying the limitations of deep learning in this context, the research provides a foundation for future work. Subsequent studies could explore the integration of hybrid models or the application of advanced deep learning architectures to improve predictive performance. Furthermore, real-world deployment of these models in diverse cloud infrastructures will validate their practical utility and scalability, contributing to more efficient resource management and optimization.

V. Conclusion

This research presents a comprehensive framework for optimizing VM performance within cloud computing environments by integrating advanced machine learning methodologies and real-time monitoring tools. The study highlights the exceptional efficacy of the Random Forest Regression model, which achieved a MAPE of 2.65%, significantly outperforming the LSTM model. This reduction in prediction error underscores the model's ability to enable precise resource allocation, leading to substantial improvements in system performance, operational efficiency, and cost-effectiveness.

Compared to traditional approaches and prior studies, this framework represents a critical advancement in cloud resource management. The integration of real-time monitoring tools, such as Prometheus and Grafana, combined with advanced predictive analytics, enables dynamic adaptability to workload changes and more efficient resource utilization. By addressing key limitations of earlier research, such as reliance on less adaptive systems or single-method approaches, this study establishes a new benchmark for VM optimization, particularly by demonstrating the robustness of ensemble learning techniques like Random Forest in handling complex and dynamic resource utilization patterns.

While this research demonstrates significant progress, future work could explore the integration of the Random Forest model into a hybrid framework, building upon the strengths of ensemble learning and deep learning. Such an approach could leverage the Random Forest model's robust accuracy alongside LSTM's ability to handle sequential patterns, creating a scalable and adaptive solution for even more complex cloud environments. Additionally, extending the framework to incorporate metrics like energy consumption and applying it across multi-cloud environments would further enhance its utility and relevance. Real-world deployment and validation in diverse cloud infrastructures will be essential for solidifying its practical impact and scalability. In conclusion, the proposed framework offers a versatile and practical solution for addressing the challenges of modern cloud computing environments, paving the way for more efficient and sustainable cloud operations.

References

[1] O. Obi, S. Dawodu, A. Daraojimba, S. Onwusinkwue, O. Akagha, and I. Ahmad, "REVIEW OF EVOLVING CLOUD COMPUTING PARADIGMS: SECURITY, EFFICIENCY, AND INNOVATIONS," Computer Science & IT Research Journal, vol. 5, pp. 270–292, Feb. 2024, doi: 10.51594/csitrj.v5i2.757.

[2] Y. Wang, Q. Bao, J. Wang, G. Su, and X. Xu, "Cloud Computing for Large-Scale Resource Computation and Storage in Machine Learning," JTPES, vol. 4, no. 03, pp. 163–171, Mar. 2024, doi: 10.53469/jtpes.2024.04(03).14.

[3] M. Attaran, S. Attaran, and D. Kirkland, "Technology and Organizational Change: Harnessing the Power of Digital Workplace," 2019, pp. 383–408. doi: 10.4018/978-1-5225-8933-4.

[4] M. E. E. Alahi et al., "Integration of IoT-Enabled Technologies and Artificial Intelligence (AI) for Smart City Scenario: Recent Advancements and Future Trends," Sensors, vol. 23, no. 11, Art. no. 11, Jan. 2023, doi: 10.3390/s23115206.

[5] P. K. G. Pandian, "Effective Resource Management In Virtualized Environments," vol. 1, no. 7, 2023.

[6] M. Yenugula, S. Sahoo, and S. Goswami, "Cloud computing for sustainable development: An analysis of environmental, economic and social benefits," Journal of Future Sustainability, vol. 4, no. 1, pp. 59–66, 2024.

[7] R. Islam et al., "The Future of Cloud Computing: Benefits and Challenges," International Journal of Communications, Network and System Sciences, vol. 16, no. 4, Art. no. 4, Apr. 2023, doi: 10.4236/ijcns.2023.164004.

[8] H. Ahmed, H. J. Syed, A. Sadiq, A. O. Ibrahim, M. Alohaly, and M. Elsadig, "Exploring Performance Degradation in Virtual Machines Sharing a Cloud Server," Applied Sciences, vol. 13, no. 16, Art. no. 16, Jan. 2023, doi: 10.3390/app13169224.

[9] S. Kraft, G. Casale, D. Krishnamurthy, D. Greer, and P. Kilpatrick, "Performance models of storage contention in cloud environments," Softw Syst Model, vol. 12, no. 4, pp. 681–704, Oct. 2013, doi: 10.1007/s10270-012-0227-2.

[10] Y. Gong, J. Huang, B. Liu, J. Xu, B. Wu, and Y. Zhang, "Dynamic Resource Allocation for Virtual Machine Migration Optimization using Machine Learning," Mar. 20, 2024, arXiv: arXiv:2403.13619. Accessed: Nov. 10, 2024. [Online]. Available: http://arxiv.org/abs/2403.13619

[11] R. Doukha, A. Ez-Zahout, and A. Ndayikengurukiye, "Forecasting virtual machine resource utilization in cloud computing: a hybrid artificial intelligence approach," Indonesian Journal of Electrical Engineering and Computer Science, vol. 37, no. 3, Art. no. 3, Mar. 2025, doi: 10.11591/ijeecs.v37.i3.pp1887-1898.

[12] S. Zheng, F. Huang, C. Li, and H. Wang, "A Cloud Resource Prediction and Migration Method for Container Scheduling," in 2021 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), Dec. 2021, pp. 76–80. doi: 10.1109/TOCS53301.2021.9689034.

[13] N. Kumawat, N. Handa, and A. Kharbanda, "Cloud Computing Resources Utilization and Cost Optimization for Processing Cloud Assets," in 2020 IEEE International Conference on Smart Cloud (SmartCloud), Washington DC, WA, USA: IEEE, Nov. 2020, pp. 41–48. doi: 10.1109/SmartCloud49737.2020.00017.

[14] H. Shen and L. Chen, "A Resource-Efficient Predictive Resource Provisioning System in Cloud Systems," IEEE Transactions on Parallel and Distributed Systems, vol. 33, no. 12, pp. 3886–3900, Dec. 2022, doi: 10.1109/TPDS.2022.3172493.

[15] M. Adeel Abbas, Z. Iqbal, F. Zeeshan Khan, S. Alsubai, A. Binbusayyis, and A. Alqahtani, "An ANN based bidding strategy for resource allocation in cloud computing using IoT double auction algorithm," Sustainable Energy Technologies and Assessments, vol. 52, p. 102358, Aug. 2022, doi: 10.1016/j.seta.2022.102358.

[16] J. Ariza, M. Jimeno, R. Villanueva-Polanco, and J. Capacho, "Provisioning Computational Resources for Cloud-Based e-Learning Platforms Using Deep Learning Techniques," IEEE Access, vol. PP, pp. 1–1, Jun. 2021, doi: 10.1109/ACCESS.2021.3090366.

[17] X. Han, R. Schooley, D. Mackenzie, O. David, and W. J. Lloyd, "Characterizing Public Cloud Resource Contention to Support Virtual Machine Co-residency Prediction," in 2020 IEEE International Conference on Cloud Engineering (IC2E), Apr. 2020, pp. 162–172. doi: 10.1109/IC2E48712.2020.00024.

[18] D. Huang, L. Costero, A. Pahlevan, M. Zapater, and D. Atienza, "CloudProphet: A Machine Learning-Based Performance Prediction for Public Clouds," Sep. 28, 2023, arXiv: arXiv:2309.16333. Accessed: Nov. 10, 2024. [Online]. Available: http://arxiv.org/abs/2309.16333

[19] A. K. Wiesi et al., "Optimizing Cloud Resource Utilization Through Machine Learning Forecasting," . Vol., no. 17.

[20] A. Ndayikengurukiye, R. Doukha, E. Niyukuri, E. Muheto, A. Ez-zahout, and F. Omary, "SOAVMP: Multi-Objective Virtual Machine Placement in Cloud Computing Based on the Seagull Optimization Algorithm," IJCNA, vol. 11, no. 3, p. 375, Jun. 2024, doi: 10.22247/ijcna/2024/24.

[21] J. He, S. Hong, C. Zhang, Y. Liu, F. Deng, and J. Yu, "A Method to Cloud Computing Resources Requirement Prediction on SaaS Application," in 2021 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE), Jul. 2021, pp. 107–116. doi: 10.1109/MLISE54096.2021.00027.

[22] J. Turnbull, Monitoring with Prometheus. Turnbull Press, 2018.

[23] S. A. Algarni, M. R. Ikbal, R. Alroobaea, A. S. Ghiduk, and F. Nadeem, "Performance Evaluation of Xen, KVM, and Proxmox Hypervisors," IJOSSP, vol. 9, no. 2, pp. 39–54, Apr. 2018, doi: 10.4018/IJOSSP.2018040103.

[24] M. Chakraborty and A. P. Kundan, "Grafana," in Monitoring Cloud-Native Applications: Lead Agile Operations Confidently Using Open Source Software, M. Chakraborty and A. P. Kundan, Eds., Berkeley, CA: Apress, 2021, pp. 187–240. doi: 10.1007/978-1-4842-6888-9_6.

[25] S. Gulabani, Amazon S3 Essentials. Packt Publishing Ltd, 2015.

[26] R. M. Schulte, M. D. Lebsock, J. M. Haynes, and Y. Hu, "A random forest algorithm for the prediction of cloud liquid water content from combined CloudSat–CALIPSO observations," Atmospheric Measurement Techniques, vol. 17, no. 11, pp. 3583–3596, Jun. 2024, doi: 10.5194/amt-17-3583-2024.

[27] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," Expert Systems with Applications, vol. 134, pp. 93–101, Nov. 2019, doi: 10.1016/j.eswa.2019.05.028.

[28] A. Gozuoglu, O. Ozgonenel, and C. Gezegin, "CNN-LSTM based deep learning application on Jetson Nano: Estimating electrical energy consumption for future smart homes," Internet of Things, vol. 26, p. 101148, Jul. 2024, doi: 10.1016/j.iot.2024.101148.

[29] J. Wang, S. Hong, Y. Dong, Z. Li, and J. Hu, "Predicting Stock Market Trends Using LSTM Networks: Overcoming RNN Limitations for Improved Financial Forecasting," Journal of Computer Science and Software Applications, vol. 4, no. 3, pp. 1–7, Jul. 2024, doi: 10.5281/zenodo.12200708.

Traffic Safety in Mixed Environments by Predicting Lane Merging and Adaptive Control

Aigerim Amantay, Shyryn Akan, Nurlybek Kenes, Amandyk Kartbayev
School of Information Technology and Engineering, Kazakh-British Technical University, Almaty, Kazakhstan

***Abstract*—Autonomous driving technology is primarily developed to enhance traffic safety through advancements in motion prediction and adaptive control mechanisms. Highway lane merging remains a high-risk scenario, accounting for approximately 7% of highway collisions globally due to misjudged vehicle interactions, according to international statistics. This paper proposes a two-stage deep learning framework for autonomous lane merging in mixed traffic. Using the Argoverse dataset, which contains over 300,000 vehicle trajectories mapped to high-definition road networks, we first predict vehicle trajectories using a Seq2Seq model with LSTM layers, achieving a 21% improvement in prediction accuracy over a baseline Multi-layer Perceptron model. In the second stage, reinforcement learning is employed for maneuver generation, where a Dueling Deep Q-Network outperforms a standard DQN by 8% in collision avoidance. Experimental results indicate that the combined trajectory prediction and RL-based framework significantly reduces merging delays, enhances data-driven decision-making in mixed traffic environments, and provides a scalable solution for safer autonomous highway merging.**

***Keywords*—Autonomous driving; lane merging; traffic safety; trajectory prediction; deep learning; LiDAR; LSTM**

I. Introduction

The rise of artificial intelligence (AI) has transformed various industries, with machine learning and deep learning playing an increasingly significant role in automating complex tasks. In the automotive sector, AI-driven technologies have enabled the development of semi-autonomous vehicles equipped with features such as adaptive cruise control, lane-keeping assistance, automated parking, and collision avoidance systems. While these advancements have improved driving safety and convenience, fully autonomous driving remains a formidable challenge due to the unpredictability of human behavior, dynamic traffic conditions, and infrastructure limitations. One particularly complex and accident-prone scenario that requires further refinement in autonomous driving systems is highway lane merging.

Merging from on-ramps onto high-speed highways presents unique difficulties due to the necessity of balancing speed, gap selection, and interaction with human-driven vehicles. Unlike controlled environments such as intersections with traffic lights, highway merging involves continuous decision-making in real-time, with vehicles needing to adapt to fast-changing conditions. Poor merging decisions can lead to sudden braking, abrupt lane changes, or even multi-vehicle collisions. Statistics indicate that lane-changing and merging maneuvers account for a significant portion of highway accidents, often due to misjudgment of vehicle speeds, miscommunication between drivers, and insufficient gap acceptance. Traditional rule-based autonomous driving systems, which rely on pre-set rules and simple heuristics, struggle to handle the complexity of these situations. The limitations of such approaches necessitate more sophisticated decision-making models.

Trajectory prediction is a crucial component of autonomous driving safety, enabling vehicles to make informed decisions. Traditional approaches have relied on kinematic and physics-based models, which assume that vehicle movement follows deterministic patterns. However, such models fail to capture the stochastic nature of real-world traffic, particularly in scenarios where multiple agents influence each other's actions. In recent years, interaction-aware models have gained traction, leveraging deep learning techniques to encode spatial and temporal dependencies in vehicle behavior. Studies such as those by Zhang et al. [1] and Karle et al. [2] have demonstrated that attention-based trajectory prediction models can significantly improve inference time and prediction accuracy by focusing on relevant surrounding vehicles.

Despite these advances, several challenges remain. Imbalanced trajectory datasets, where lane-keeping data vastly outnumbers lane-changing instances, hinder model training and may lead to poor generalization in merging scenarios. Additionally, most existing models rely solely on position data, neglecting critical information such as vehicle velocity, acceleration patterns, and road geometry. While deep learning-based models can enhance trajectory prediction, they do not directly translate into maneuver execution, requiring additional mechanisms to translate predictions into safe context-aware driving actions.

To address these challenges, this paper proposes a two-stage approach to highway merging automation. In the first stage, we leverage real-world LiDAR data to train a deep learning-based trajectory prediction model capable of capturing vehicle interactions. The baseline model employs a Multi-layer Perceptron (MLP) to establish feasibility, which is later extended to a Seq2Seq model with Long Short-Term Memory (LSTM) layers to improve prediction precision. The second stage integrates deep reinforcement learning (RL) to optimize autonomous vehicle maneuvers. By training RL agents on predicted trajectory data, the system learns to make adaptive decisions that balance safety and traffic flow continuity.

A key hypothesis of this study is that by combining sequence-based trajectory prediction with reinforcement learning, an autonomous vehicles (AVs) can anticipate and react to merging scenarios more effectively than conventional

rule-based systems. The hypothesis is based on the premise that motion forecasting alone is insufficient—vehicles must also be able to evaluate merging feasibility and adjust their actions accordingly. This research aims to improve merging efficiency by ensuring that AVs merge at appropriate speeds, and minimize disruptions to surrounding traffic.

The implications of this work extend beyond highway merging. As AVs gradually transition from human-supervised automation (Level 3 autonomy) to full self-driving (Level 5 autonomy), complex interactions with human drivers will remain a challenge. AVs must not only predict future vehicle positions but also infer driver intent and adapt to cooperative or adversarial driving behaviors. Ensuring safe interactions between autonomous and human-driven vehicles is critical for gaining regulatory approval for autonomous solutions.

Additionally, while reinforcement learning has demonstrated promise in driving applications, challenges such as sample efficiency, reward design, and real-world generalization remain significant barriers. Unlike games and simulations where RL agents can train for millions of iterations in a controlled environment, real-world driving data is limited, and deploying untested policies on public roads carries risks. To mitigate these concerns, simulated environments and digital twin systems can be leveraged to refine RL policies before real-world testing.

Another concern is computational feasibility. Training deep learning models on large-scale trajectory datasets is resource-intensive, often requiring high-performance GPUs and extensive tuning of hyperparameters. Computational constraints limit the ability to explore more complex architectures, such as Transformer-based motion prediction models, which may offer further improvements in prediction accuracy. Additionally, deploying computationally expensive models in real-time AV systems presents challenges, as in-vehicle processors must balance inference speed with energy efficiency.

This study addresses these concerns by proposing an adaptable framework that integrates deep learning techniques for trajectory prediction and optimization strategies for maneuver execution. The findings contribute to existing research by evaluating how Seq2Seq models and reinforcement learning can be combined to improve lane merging performance, reduce collision risks, and ensure smoother highway traffic integration.

The remainder of this paper is structured as follows: Section II reviews related works on trajectory prediction in autonomous vehicles. Section III details the methodology, including dataset preprocessing, the Seq2Seq model, and reinforcement learning integration. Section IV presents experimental results comparing the performance of models like MLP, Seq2Seq, and Dueling DQN. Section V discusses findings, addressing challenges and limitations. Section VI concludes with a summary and future research directions.

II. Related Works

A. Interaction-Aware Trajectory Prediction

In the field of autonomous driving, trajectory prediction is a critical component for ensuring vehicle safety and efficiency. Many studies have focused on interaction-aware approaches to improve prediction accuracy by accounting for the behaviors and intentions of surrounding vehicles. Notably, recent advancements leverage attention-based models, recurrent neural networks, and graph-based architectures to model interactions and improve trajectory accuracy in diverse traffic conditions. For instance, Yan et al. [3] applied spatial-attention mechanisms to handle inter-vehicular interactions, achieving accurate predictions on the HD dataset with minimal computational resources.

Many researchers have proposed intention-driven models that differentiate between short- and long-term intentions for more explicable trajectory predictions [4]. The study in [5] emphasized road constraints in prediction by developing a road-aware model that uses high-definition maps, demonstrating improved data efficiency and realism in trajectory prediction. In addition to model-specific developments, surveys by study [6] outline the evolution of trajectory prediction methods, highlighting physics-based, machine learning, and deep learning approaches. The research in [7] introduced a neural network-based motion planner integrated with model predictive control, balancing conservative planning with interaction-based optimization.

B. On-Demand Approach for Trajectory Prediction

A graph and recurrent neural network (GNN-RNN) based framework has been proposed to capture inter-vehicular interactions on highways using historical vehicle data for future path prediction. This model utilizes directed graphs to represent dynamic traffic interactions and demonstrates the capability to predict multi-vehicle trajectories for high-density traffic environments [8]. Similarly, an attention-based approach enhances trajectory prediction by focusing on the significance of neighboring vehicles through a multi-layer attention mechanism. The model incorporates both local and global attention components, enabling consideration of diverse driving goals and improving accuracy, especially in long-range highway scenarios [9].

An on-demand model for rapid vehicle path prediction with minimal observation windows has also been explored. This method probabilistically extends traditional car-following models, adapting to new traffic configurations with limited input data and improving reaction times for autonomous vehicles [10]. Further developments include multi-attention mechanisms for both spatial and temporal interactions. For instance, a Transformer-based architecture predicts multimodal vehicle trajectories by accommodating complex interaction patterns [11]-[12]. Another attention-based approach focuses on interaction regions and adapts predictions based on the relative positions of surrounding vehicles [13].

Graph-based deep learning frameworks have been integrated with trajectory prediction models to enable proactive longitudinal control. By combining LSTM networks with graph convolutional networks, this method predicts lane-aware behavior and captures inter-vehicle interactions, improving prediction accuracy and passenger ride quality [14]. Additionally, a structural-LSTM network assigns individual

LSTM networks to each vehicle, allowing real-time spatial information exchange. This architecture models fine-grained interactions effectively, enhancing predictive accuracy in mixed-traffic environments [15].

C. Driving Dynamics and Computational Efficiency

To address real-world challenges, one study emphasizes aligning predictive models with real driving dynamics and computational efficiency. The research critiques dataset-based evaluations and advocates for a task-driven approach that reflects the model's impact on downstream driving behavior, highlighting the interaction between autonomous vehicles and other road users as critical to trajectory model accuracy [16].

In their study, [17] presented a novel trajectory prediction framework designed to improve the reliability of autonomous driving systems. The key contribution of their work lies in the introduction of an awareness module that dynamically evaluates the performance of the trajectory prediction model during operation. This self-assessment capability enables the system to identify and respond to potential prediction inaccuracies. An intention-aware transformer model has been developed to adapt to social and temporal learning requirements in trajectory prediction. Using a multi-head self-attention mechanism, this model captures intricate social dependencies and driving behaviors across timestamps, improving its ability to manage complex driving scenarios [18]-[19]. Similarly, contextual cues, such as actor-actor and actor-scene interactions, have been incorporated into prediction frameworks. Attention-based graph modules and convolutional networks integrate spatial-temporal data, enhancing reliability in mixed-traffic conditions [20]-[21].

AVs must excel at predicting future events, a capability human drivers perform instinctively. Imagine an AV preparing to turn right at an intersection while a pedestrian approaches from the crosswalk on the right and another vehicle waits to proceed from the opposite direction. For the AV to navigate this scenario safely, it must anticipate whether the pedestrian will stop or continue crossing and whether the opposing vehicle will yield or attempt to proceed simultaneously.

This complex interplay of actions is central to motion prediction, enabling AVs to understand their surroundings and make proactive decisions. Sensors like gyroscopes, cameras, and etc. provide the necessary environmental data to inform these predictions. While rule-based systems have traditionally been used for such tasks, they falter under uncertainty and complexity, especially as the number of interacting agents increases. A data-driven approach using supervised machine learning offers a more scalable solution [22].

By tracking the movements of nearby objects over a 5-second horizon using their previous 1-second trajectories, AVs can transform motion prediction, planning, and simulation into data-centric problems [23]. Furthermore, the model must be versatile enough to handle scenarios, such as intersections, congested urban streets, and highways. The choice of neural network architecture is pretty obvious to achieve a balance between prediction speed and adaptability. Despite advancements in trajectory prediction and motion planning, the reviewed studies still have several challenges that remain unaddressed. Addressing these gaps requires hybrid applied approaches that combine behavior modeling and uncertainty quantification, as demonstrated in our research.

III. Methodology

A. Dataset

The Argoverse dataset [24] is a publicly available resource designed to advance research in autonomous driving by providing diverse real-world data for perception, trajectory forecasting, and motion planning. Collected from vehicles equipped with high-resolution LiDAR sensors, multiple RGB cameras, and detailed high-definition maps, the dataset enables a broad range of self-driving tasks. It comprises over 300k trajectories from more than 1,000 hours of driving, encompassing urban and suburban scenarios.

The dataset includes two key components: the 3D Tracking Dataset, focused on object detection and tracking, and the Motion Forecasting Dataset, aimed at predicting the future trajectories of vehicles and other traffic participants. With detailed annotations for traffic participants and trajectory data, along with HD maps containing lane geometry and traffic controls, the dataset facilitates accurate motion planning and interaction modeling. Its temporal sequences and multimodal data structure allow for advanced applications such as trajectory prediction, behavior forecasting, and real-time motion planning. As a well-annotated dataset among others, as shown in Fig. 1, Argoverse provides benchmarks for evaluating models, making it a best choice for developing interaction-aware systems in complex traffic environments.

	TIMESTAMP	TRACK_ID	OBJECT_TYPE	X	Y	CITY_NAME
0	3.159682e+08	00000000-0000-0000-0000-000000000000	AV	419.354578	1125.928065	MIA
1	3.159682e+08	00000000-0000-0000-0000-000000023470	OTHERS	404.729217	1253.006591	MIA
2	3.159682e+08	00000000-0000-0000-0000-000000023463	OTHERS	491.967704	1147.286581	MIA
3	3.159682e+08	00000000-0000-0000-0000-000000023476	OTHERS	473.827482	1146.672473	MIA
4	3.159682e+08	00000000-0000-0000-0000-000000023478	OTHERS	419.641337	1252.034538	MIA

Fig. 1. A sample of raw dataset.

To match the data's quality for subsequent modeling tasks, we began by extracting all x-y coordinate data associated with each timestamp and then organized the data by vehicle type to account for behavioral differences among various traffic participants. The coordinates were normalized to the range 0 to 1, representing their relative position to the data-collection vehicle. This normalization step ensures scale invariance and allows for consistent interpretation of spatial relationships. To maintain numerical precision, we retained up to six decimal places during this transformation.

The processed data was then split into five second intervals for each vehicle trajectory, reflecting the temporal progression of the observed environment. Each interval was further divided into two parts: the first three seconds served as training data, capturing historical movements, while the final seconds were used for testing, simulating future trajectory prediction. To improve the reliability of the input data, we filtered out incomplete trajectories. Specifically, any vehicle missing sufficient information defined as fewer than 51 rows of data, equivalent to 4 and 5+ seconds at a 5-10Hz sampling rate was excluded. This careful filtering process minimized signal noise and inconsistencies, ensuring that only reasonable data

informed the training and testing stages (see details in Table I). This pipeline is the foundation for next stages of the trajectory prediction model.

TABLE I. OVERVIEW OF THE DATASET

Field	Argoverse dataset	
	Description	***Most Frequent Value***
TIMESTAMP	Timestamp of the recorded data point	3.16E+08
TRACK_ID	Unique identifier for each tracked vehicle	00000000-0000-0000-0000-000000000000
OBJECT_TYPE	Type of object (e.g., 'AV' for autonomous vehicle, 'OTHERS' for surrounding vehicles)	OTHERS
X	Longitudinal position of the vehicle	402.8939
Y	Lateral position of the vehicle	1253.103
CITY_NAME	City where the data was collected	MIA

In addition to position coordinates (X, Y), velocity components (Vx, Vy) were derived from consecutive position differences over time. The acceleration components were further computed to capture variations in vehicle speed and potential braking or acceleration events. This allowed the model to distinguish between stable lane-following behavior and unexpected maneuvers, such as lane changes or emergency stops. These computed features were normalized within a [0, 1] range to ensure numerical stability during training.

Another crucial aspect was categorical encoding of object types. Since the dataset includes both AVs and surrounding human-driven vehicles (OTHERS), a one-hot encoding scheme was applied to differentiate between the two. This distinction was necessary because human drivers exhibit complex behavior, requiring adaptive mechanisms that consider their almost unpredictable decisions.

The city identifier (CITY_NAME) provided contextual information regarding the driving environment. Data collected from MIA (Miami) was analyzed separately to identify any city-specific driving patterns, such as differences in traffic density, intersection layouts, or lane configurations. While the dataset predominantly focuses on highway scenarios, environmental factors such as road curvature, lane widths, and merging configurations were later incorporated as additional inputs in the models.

To account for lane-aware trajectory dependencies, road geometry data from the high-definition map layers of the Lyft Level 5 dataset [25] was extracted for augmentation. This data included lane centerlines, traffic sign positions, and speed limits, enabling trajectory predictions that align with road constraints rather than purely data-driven extrapolations.

The processed dataset was divided into training (70%), validation (20%), and test (10%) subsets. Given the temporal nature of the data, a sliding window approach was implemented to segment continuous vehicle trajectories into overlapping time frames of 5-second windows.

B. Mathematical Model

The goal is to predict the trajectory of a target vehicle

$$\mathbf{\Theta}_\tau = \left[\boldsymbol{\theta}_\tau^{N_h+1}, \dots, \boldsymbol{\theta}_\tau^{N_h+N_f}\right] \quad (1)$$

where $\boldsymbol{\theta}_\tau^t = [\chi_\tau^t, \psi_\tau^t]$ denotes its longitudinal and lateral positions over the future $T_f = 4$ s ($N_f = 24$ time steps). The input consists of historical observations over $T_h = 3$ s ($N_h = 18$ time steps) for the target and nine surrounding vehicles, represented as,

$$\xi_t^i = \left[\chi_t^i, \psi_t^i, v_{\chi,t}^i, v_{\psi,t}^i\right], \quad (2)$$

where (χ, ψ) are the positions and (v_χ, v_ψ) the velocities. To focus on relative dynamics, the state of each surrounding vehicle is expressed relative to the target as $\Delta\xi_t^i = \xi_t^i - \xi_t^\tau$. The complete input sequence is $\mathbf{\Delta\Xi} = [\Delta\xi^1, \xi^\tau, \dots, \Delta\xi^9]$.

An encoder-decoder framework with LSTM is used to capture temporal dependencies. The encoder processes $\mathbf{\Delta\Xi}$ to produce hidden states $\mathbf{h}_t$ for $t = 1, \dots, N_h$, with $\mathbf{h}_{N_h}$ serving as the context vector $\mathbf{C}$. The decoder predicts the future trajectory step-by-step, using the context vector, the hidden state s_t, and the previous output $\boldsymbol{\theta}_\tau^{t-1}$. The prediction is given by

$$\boldsymbol{\theta}_\tau^t = f_{\text{dec}}(s_t, \mathbf{C}) \quad (3)$$

To address sequence representation limitations, we employ two attention mechanisms. Context-aware attention reweights $\mathbf{C} = \mathbf{h}_{N_h}$ by assigning importance α_t^j to its elements, forming

$$\mathbf{C}_t = [\alpha_t^1 \cdot h_{N_h}^1, \dots, \alpha_t^k \cdot h_{N_h}^k] \quad (4)$$

Lane-aware attention divides the surrounding vehicles into four groups (current, left, and right lanes, behind), producing context vectors $\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3, \mathbf{C}_4$. The final context vector is,

$$\mathbf{C}_t = \beta_t^1 \cdot \mathbf{C}_1 + \beta_t^2 \cdot \mathbf{C}_2 + \beta_t^3 \cdot \mathbf{C}_3 + \beta_t^4 \cdot \boldsymbol{C}_4 \quad (5)$$

where β_t^i reflects each lane's relevance. The model is trained using mean squared error:

$$\mathcal{L} = \frac{1}{N_f} \sum_{t=N_h+1}^{N_h+N_f} \|\boldsymbol{\theta}_\tau^t - \Theta_\tau^t\|^2 \quad (6)$$

This approach integrates spatial and temporal dependencies, enabling interpretable trajectory prediction, as shown in Fig. 2.

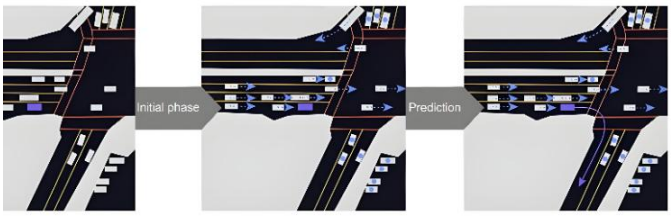


Fig. 2. The trajectory prediction process in a highway merging scenario.

C. Base Model

The baseline model selected for this study is a Multi-Layer Perceptron (MLP), a neural network architecture that uses back-propagation for training. The MLP serves as an initial framework to evaluate the dataset's viability. This model consists of one input layer, two hidden layers, and one output layer, utilizing the Rectified Linear Unit (ReLU) activation

function to ensure non-linearity. During training, the MLP baseline model was first tested to establish feasibility, followed by a Seq2Seq model with LSTM layers. The input to the model comprises 10 frames from the previous second, sampled at 0.1-second intervals, capturing both the agent's and the autonomous vehicle's positions. From these positions, the motion of the surrounding agents is derived and used to predict their trajectories over the next five seconds.

For further refining the predictions, we implement an ensemble approach using a stacking algorithm. Each individual model is trained separately on the data and later combined through additional neural network layers. This ensemble network effectively reduces generalization error and improves prediction accuracy. The approach also allows for trajectory visualization, particularly when incorporating labeled traffic signs into the predictions. The MLP model in this setup has 17k trainable parameters.

The predicted trajectory hypotheses are compared against the ground truth by modeling the likelihood under a mixture of Gaussians [26]. The mean values are set to the predicted trajectories, while the covariance is modeled using an identity matrix, enabling a probabilistic assessment of prediction accuracy. This approach ensures robust trajectory prediction and facilitates interpretable outputs in barely predictable traffic scenarios.

D. Seq2Seq-LSTM Model

To improve trajectory prediction beyond the baseline MLP, we adopted a Seq2Seq model, which as a type of encoder-decoder framework with LSTM layers, is well-suited for translating sequences of one domain into another with different lengths, such as time series positional data to future trajectories. In our case, the model processes 3 seconds (Nh=15 frames) of historical positional data and predicts the next 5 seconds (Nf=25 frames) of motion. The sequence length of 15 frames (for 3-5 seconds at 5 Hz sampling rate) was found to provide the best balance between predictive accuracy and computational efficiency.

The encoder LSTM compresses the input sequence into a latent state, discarding intermediate outputs, while the decoder LSTM uses this latent state to iteratively generate the future trajectory. A dense layer with ReLU activation ensures the output values remain normalized in the range [0,1]. Categorical cross-entropy is used as the loss function, and the model contains 4m trainable parameters.

For integration of the predicted trajectories with a reinforcement learning (RL) module, as shown in Fig. 3, we established positional and velocity mappings. Positional mapping adjusts the vehicle positions to align with a custom highway environment consisting of two main lanes and a merging lane over a 500-meter stretch, divided into four zones: pre-merged, convergence, merge, and post-merged (150m, 100m, 150m, and 100m zones). Vehicle positions (x, y) are transformed relative to the monitored vehicle, mapping lateral placement (x) to lane alignment and longitudinal placement (y) relative to the vehicle's merging zone location. Velocity mapping calculates initial and target velocities based on historical and predicted trajectory data, ensuring consistency with the simulated environment.

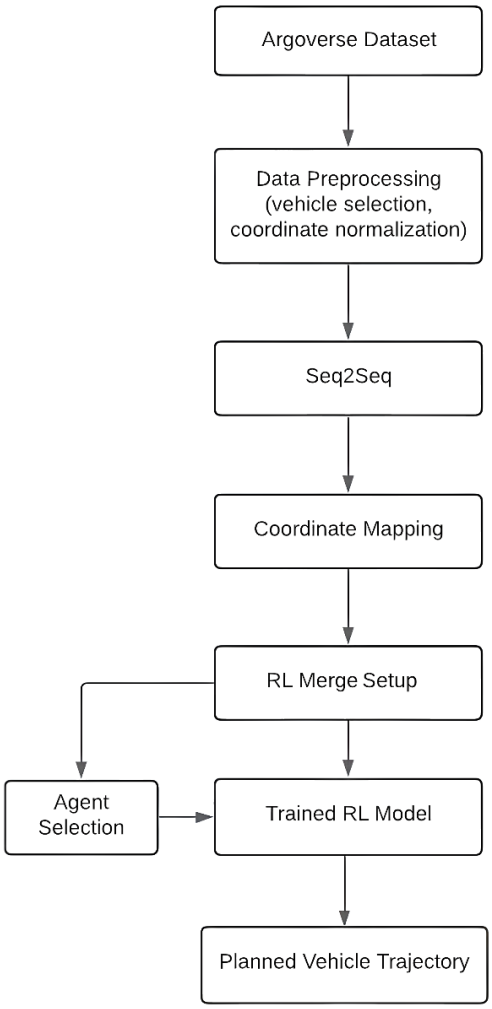


Fig. 3. The general design of the development process.

The RL simulates the merging process by defining six possible actions for the vehicle: maintaining, accelerating, or decelerating speed, either while staying in the current lane or changing lanes. The state space includes the positions (x), velocities (v), and orientations (θ) of all vehicles in the scene. A simplified reward function $\mathcal{R}$ is designed to optimize safe and efficient merging behavior by balancing key factors such as speed, lane preference, and collision avoidance:

$$\mathcal{R} = \lambda_c \cdot \delta_c + \lambda_l \cdot \delta_l + \lambda_s \cdot \delta_s, \quad (7)$$

where:

- $\lambda_c, \lambda_l, \lambda_s, \lambda_m$: tunable hyperparameters, where λ_c is a collision penalty weight, λ_1 - lane preference reward weight, λ_s - speed reward weight,
- δ_c : binary indicator (1 if a collision occurs, 0 otherwise),
- δ_l : binary indicator for being in the desired lane (1 if true, 0 otherwise),

- δ_s : normalized vehicle speed within the desired range.

To address the merging dynamics, an additional penalty is applied if the vehicle's speed deviates from the target speed:

$$\mathcal{R} += \lambda_m \cdot \left(v_\tau^{\text{target}} - v_\tau\right)^2, \quad (8)$$

where:

- λ_m : penalty weight for merging speed deviation,
- $v_\tau^{\text{target.}}$: target speed of the vehicle,
- v_T : current speed of the vehicle,
- $\mathcal{R}$: total reward received by the vehicle.

This RL combines Seq2Seq-based trajectory prediction with a reward mechanism that aligns with adaptability of the model.

IV. Results

A. Implementation of the Model

To address the computational challenges posed by the large dataset, we optimized the training configuration by limiting the number of epochs and batch size. While this approach ensured feasible GPU memory usage, it likely prevented further reduction of the loss function, which could be achieved with extended training iterations. All experiments were conducted using Google Colab with an Nvidia GPU, as the cards on local machines was insufficient for these tasks. Despite the hardware accelerator, training times for individual models varied significantly, often exceeding 11 hours. Models such as DenseNet, ResNet, and others were explored but could not fit within the VRAM limitation of the GPU.

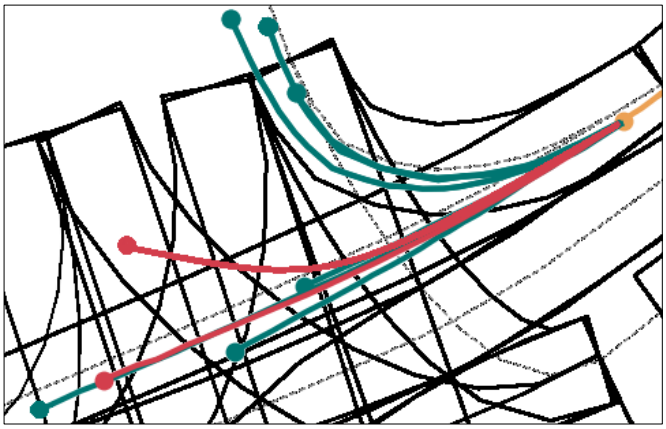


Fig. 4. The schematics of predicted trajectory for an AV.

The primary focus of the trajectory prediction task was on the Seq2Seq model. Training involved mapping historical positions and velocities to future trajectories, visualized using a rasterizer and trajectory drawing functions. For instance, Fig. 4 illustrates the ground truth and predicted trajectory for an AV within a specific map scene. Although the predicted trajectory does not fully align with the ground truth, it remains within the correct lane, demonstrating reasonable accuracy. Additional visualizations also included scenarios with traffic sign labels, where agents responded appropriately to signals—stopping at red lights and proceeding through green lights in the predicted frames.

For implementation, the dataset was preprocessed to load training data, train the model, and then test it on unseen scenes. The trained model generated predictions for both agents and AV trajectories. Using ensemble methods, we improved motion prediction through stacking algorithms, enabling comprehensive visualizations of entire scenes. In maneuver generation, performance comparisons were made across multiple RL agents. As we starting with a baseline MLP agent, we also evaluated the Deep Q-Network (DQN) and Dueling Deep Q-Network (DDQN) agents [27].

Results indicated improved trajectory prediction and decision-making in RL environments using advanced agents. The validation loss, consistently at or below the training loss suggested that the model achieved reliable training accuracy. These experiments shows the efficacy of the ensemble model and the integration of RL for trajectory planning. The visualizations provided insights into the model's accuracy in collision prediction. For example, in Fig. 5, surrounding vehicles exhibited consistent longitudinal motion with risky left side movement, except for one vehicle deviating slightly.

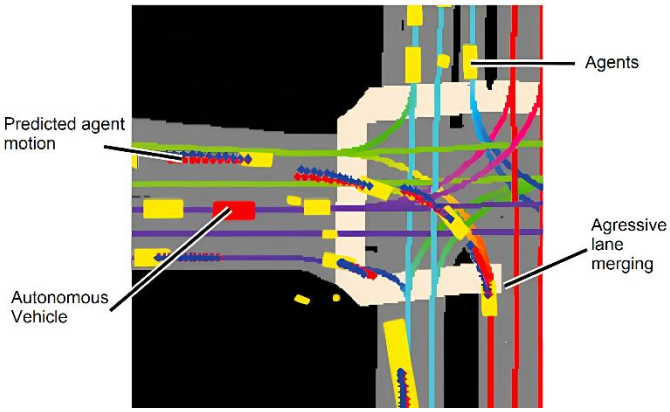


Fig. 5. The risky motion with substantial side movement.

B. Model Evaluation

The experiments begins with preprocessing the Argoverse Dataset, containing over 300k CSV files of vehicular positional data. A subset of 50k samples is selected for training. The data is normalized, and target outputs are quantized to six decimal places for precision. For training, the first 2 to 3 seconds of each 5-second sequence are fed into a Seq2Seq model, and the trained model is saved for inference as well. During inference, the model predicts positional trajectories for the next remained seconds. These predictions are mapped into initial position, initial and target velocity parameters, which serve as inputs to the RL environment.

However, discrepancies such as a significant gap between training and validation losses may require additional regularization techniques, such as dropout, L2 weight regularization, or early stopping. Incorporating domain-specific features like initial and target velocities into the trajectory prediction process, as well as mapping predictions into RL, enhances the practical utility of the model. Overall, a combination of balanced data, hyperparameter optimization,

and careful monitoring of loss metrics could significantly improve the model's accuracy in some complex scenarios [28].

For real-time applicability, inference speed was also analyzed. The Seq2Seq model achieved inference latency of approximately 9.2 milliseconds per trajectory prediction, making it feasible for integration into autonomous vehicle planning pipelines. However, more computationally intensive attention-based models were explored in subsequent experiments to improve prediction robustness while maintaining acceptable inference times.

As shown in Fig. 6 and Fig. 7, the training and validation losses for the Seq2Seq model remain consistently low, with validation loss equal to or slightly lower than training loss, indicating effective generalization in trajectory prediction for autonomous driving.

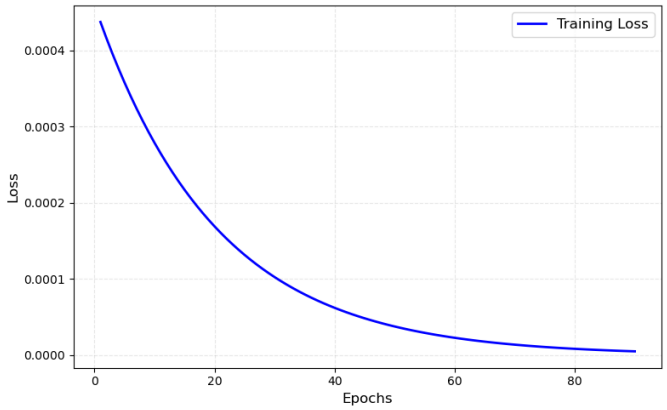


Fig. 6. Training loss metric for the model.

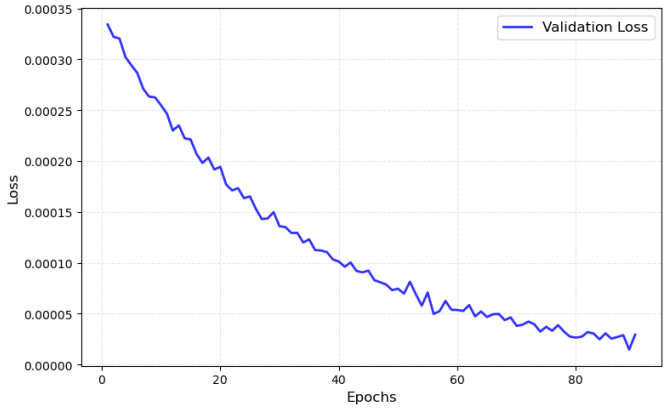


Fig. 7. Validation loss metric for the model.

In the testing phase, a separate 5-second dataset simulates sensory data typically acquired through LiDARs in real-world AV systems. Using this data, the trained Seq2Seq model predicts the positions of surrounding vehicles for the next 3 seconds. These predictions are again mapped into the RL environment, where the trained RL model generates optimized maneuver decisions for the vehicle with certain accuracy, as shown in Table II. We know that current setup increases the risk of overfitting, especially with noisy data, but hyperparameter tuning certainly can mitigate these risks. Gradient checkpointing was applied to reduce GPU memory consumption without significant trade-offs in convergence speed. As an experimental measure, early stopping was implemented to prevent overfitting, terminating training after 10 consecutive epochs without validation loss improvement.

TABLE II. ACCURACY SCORES FOR THE MODELS

Metric	*Baseline MLP*	*Seq2Seq*	*DQN*	*Dueling DQN*
Accuracy (%)	69.4000	83.1000	78.500	87.5000
MSE	0.0028	0.0014	0.002	0.0011
MAE	0.0440	0.0270	0.032	0.0220
R^2	0.6800	0.8300	0.790	0.8800

C. Simulation

The DQN enhances the classical Q-Learning algorithm by approximating the Q-function, defined as:

$$Q^*(s,a) = \mathbb{E}\left[t + \gamma \max_{a'} Q^*(s',a')\right] \quad (9)$$

where $Q^*(s,a)$ represents the maximum expected return starting from state s, taking action a, and following the optimal policy. The discount factor γ controls the importance of future rewards. Instead of storing $Q^*(s,a)$ for all state-action pairs, which is infeasible in large state spaces, DQN uses a neural network with parameters θ to approximate Q -values by minimizing the loss:

$$L_i(\theta_i) = \mathbb{E}_{s,a,r,s'}\left[\left(y_i - Q(s,a;\theta_i)\right)^2\right] \quad (10)$$

where the temporal difference target y_i is given by:

$$y_i = t + \gamma \max_{a'} Q(s',a';\theta_{i-1}) \quad (11)$$

Dueling DQN further refines this by splitting the network into two execution path: one estimating the state and the other estimating action-related advantages. These are combined in the final layer to produce the Q-value, improving performance in scenarios where distinguishing between actions is not immediately necessary. The combination of these two streams in the final layer, while effective for many scenarios, relies heavily on the assumption that the advantage function can be effectively separated from the state value function. In complex environments where actions and states are interdependent, this separation may lead to suboptimal policy learning, as the model might underestimate or overestimate the advantage of specific actions.

In simulation experiments, agents based on DQN and Dueling DQN were evaluated with varying discount factors (γ) over 1000 training epochs. Qualitative results show that dynamic reward settings lead to faster, riskier merges, while less dynamic result in cautious behavior (see Fig. 8 and Fig. 9). We could also explore the impact of reward structures on merging strategies. However, as noted in certain models, a drop in prediction accuracy implies that these algorithms often fail to truly understand the behavior of individual drivers.

While deep learning methods can capture correlations between vehicle trajectories, they may struggle with causal inference, leading to errors in scenarios where driver intent

significantly deviates from learned patterns. As a result, these models often generalize poorly in unpredictable scenarios, where one driver may aggressively accelerate into a merging lane while another might yield prematurely.

We assume, as Dueling DQN excels in scenarios where distinguishing between actions is less critical (e.g., when multiple actions lead to similar outcomes), it may struggle in some fine-grained decision-making cases, such as those involving continuous action spaces or rapid maneuvering. The architecture assumes that state-value estimation can guide the policy sufficiently when action advantages are less distinct, which might not hold true in edge cases requiring much precise action-value differentiation [29].

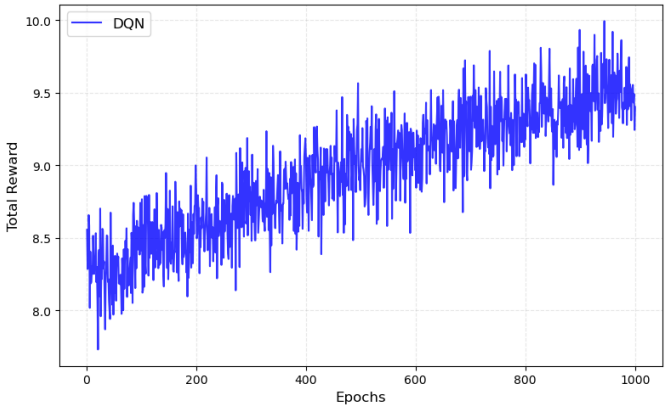


Fig. 8. Performance results of DQN agent.

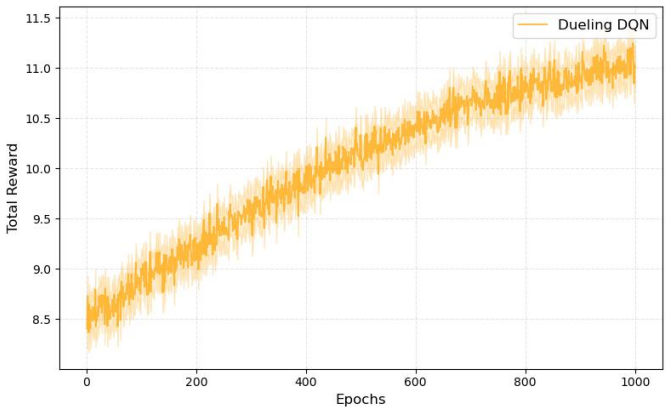


Fig. 9. Performance results of dueling DQN agent.

Nonlinear driver behaviors significantly influenced by sensor noise and capability of interpreting accidents, such as sudden braking or unexpected lane changes. For instance, human drivers exhibit a wide range of behaviors influenced by factors such as aggressiveness, reaction time, and external conditions (e.g., weather, road layout, or traffic density). Traditional deep learning models, such as RNNs or encoder-decoder architectures, may fail to distinguish between cautious and aggressive drivers, treating all vehicle trajectories as homogeneous. Results indicate that Dueling DQN outperformed standard DQN, particularly in collision-avoidance scenarios, as it accounts for delayed decisions when a collision is imminent, with the vehicle approaching to surrounding cars, as depicted in Fig. 10 and Fig. 11.

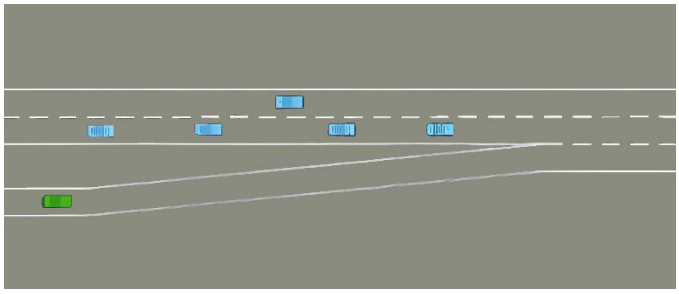


Fig. 10. Collision scenarios before merging.

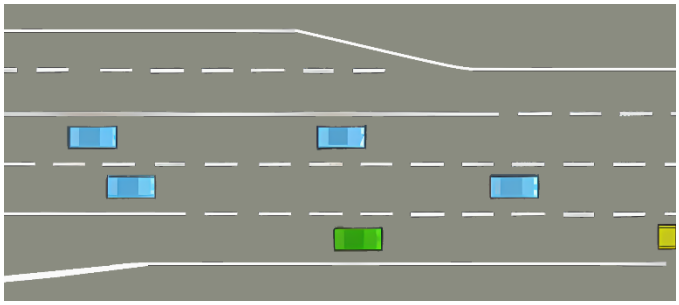


Fig. 11. The vehicle approaching to surrounding cars at the merging moment.

Prediction error generally increases with vehicle density, as shown in Fig. 12. This heatmap reflects the growing complexity of motion prediction in congested traffic, where interactions between multiple agents introduce uncertainty. Addressing this challenge requires enhanced contextual awareness to mitigate errors in high-density scenarios. Additionally, the need for uncertainty-aware models is evident, ensuring robustness across diverse urban environments and enhancing multi-agent interaction strategies, especially for likelihood of emergency maneuvers.

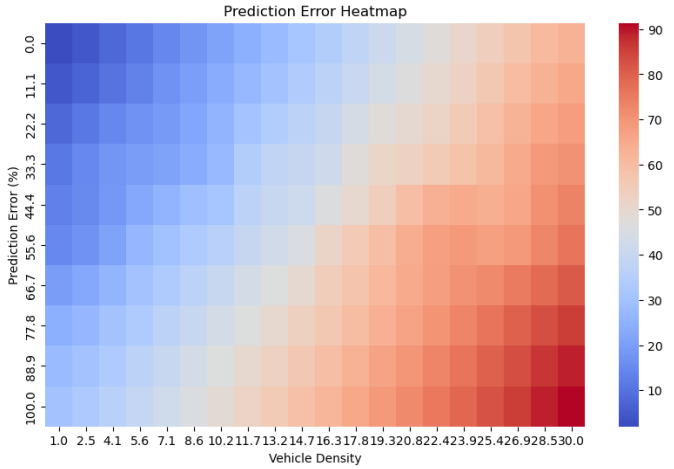


Fig. 12. The relationship between vehicle density and prediction error.

These inaccuracies become more pronounced in long-term predictions, where small errors in trajectory forecasts accumulate over time, resulting in substantial deviations from actual vehicle behavior. Given that AVs must make real-time decisions that depend on both immediate and extended motion forecasting, inconsistencies in predictions can disrupt maneuver planning, leading to suboptimal gap selection,

unnecessary braking, or unsafe merging. To mitigate these issues, a more robust approach is required that combines explicit driver behavior modeling with trajectory prediction. As demonstrated in our application (see Fig. 13), integrating behavior-aware prediction techniques improves the system's ability to anticipate diverse driving patterns, leading to more reliable and adaptive merging strategies.

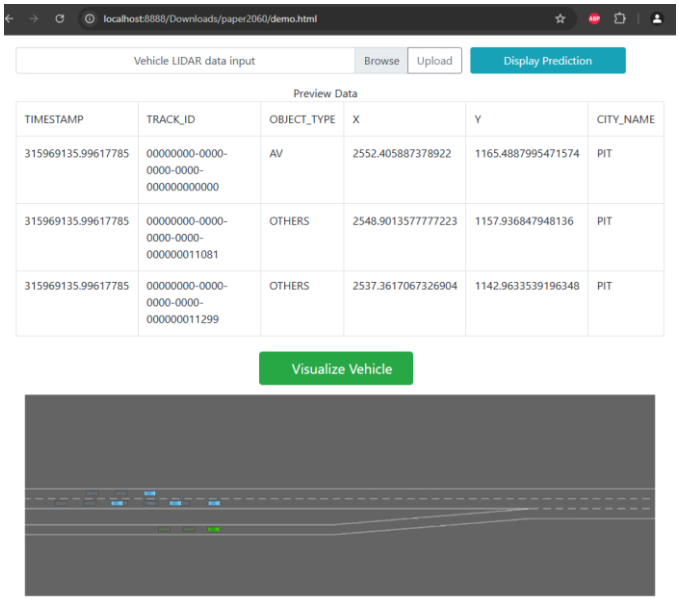


Fig. 13. Demo application for explicit driver behavior modeling.

The experimental results demonstrate several notable advancements in maneuver prediction:

- The removal of incomplete and inconsistent trajectories significantly enhanced prediction stability by decreasing noisy data.
- Incorporating velocity, acceleration, and road geometry features contributed to motion representation quality. The Seq2Seq model with LSTM layers outperformed baseline approaches.
- Optimized batch processing and gradient checkpointing effectively managed GPU memory constraints.
- The inference latency remained within operational thresholds of the model's applicability for real-time deployment.

V. Discussion

The experimental results reveal significant advancements in trajectory prediction and decision-making for AVs. However, these results also highlight key challenges that must be addressed to achieve full autonomy. The trajectories on highways, often categorized into lane-keeping and lane-changing, present distinct challenges due to their unbalanced data distribution. Lane-changing events are relatively rare compared to lane-keeping, leading to difficulties in model generalization. To mitigate this, we increased the penalty for lateral position errors by a factor of three while keeping the longitudinal penalty unchanged. This adjustment improved performance on lane-changing trajectories but further methods such as data augmentation and rebalancing may enhance outcomes.

Our Seq2Seq-based trajectory prediction framework performed reliably in predicting short-term trajectories. The low and stable training and validation losses (Fig. 6 and Fig. 7) indicate effective generalization, especially in scenarios where vehicle movements are smooth and continuous. However, when applied to long-term predictions, the model struggled to account for sudden changes influenced by dynamic factors like merging traffic, traffic signals, and unexpected obstacles. This suggests that while Seq2Seq excels in capturing immediate patterns, it may benefit from additional context for long-term predictions.

Existing studies, such as [30], use a similar encoder-decoder LSTM architecture but focus on fewer surrounding vehicles and rely solely on positional data. Our approach, incorporating both position and velocity information, provided richer dynamics. However, as our experiments show, more information does not always equate to better performance. For short-term predictions, focusing on vehicles immediately behind or adjacent to the target vehicle yielded more accurate results. For longer horizons, integrating data about road geometry, traffic density, and environmental signals could further enhance predictions.

The computational challenges of training on large datasets also posed limitations. Processing 300k files from the Argoverse dataset required significant resources, even after selecting a subset of 50k samples. Models were trained using an Nvidia GPU on Google Colab, as local GPUs lacked sufficient memory. Despite hardware acceleration, training times for Seq2Seq often exceeded tens of hours, and memory-intensive models could not be tested due to their high requirements. These constraints limited our ability to experiment with more complex models and hyperparameter configurations.

The RL module demonstrated promising results for maneuver generation. Starting with the baseline MLP, we evaluated DQN and Dueling DQN agents. The results showed that Dueling DQN consistently outperformed DQN in high-stakes scenarios such as collision avoidance. This can be attributed to Dueling DQN's ability to separate state-value and action-advantage estimations, allowing the model to prioritize critical decisions. As illustrated in Fig. 11, Dueling DQN achieved smoother merges under less aggressive reward settings and faster merges under more dynamic rewards.

Our experiments also highlight the importance of reward design in RL-based trajectory generation. Aggressive reward functions encouraged riskier behavior, with the vehicle accelerating into gaps during merges. Conversely, conservative rewards resulted in cautious behavior, where the vehicle yielded to surrounding vehicles before merging. These findings underscore the need for careful tuning of reward structures to balance safety and efficiency of the model.

Despite these advancements, our system has limitations when applied to urban environments. Highways typically exhibit predictable traffic patterns with fewer obstacles and interactions. Urban settings, by contrast, involve complex

intersections, pedestrian interactions, and diverse traffic actors that require models to generalize across a broader range of objects [31]. Extending our framework to handle such environments will require incorporating additional sensory inputs (e.g., pedestrian detections, stop signs) and more adaptive models capable of responding to unpredictable events [32].

Another limitation of our system lies in its scalability. The ensemble methods used for trajectory prediction and maneuver generation, while effective, are computationally expensive. Real-world deployment of such systems would require significant optimization to achieve real-time performance. Furthermore, our reliance on high-quality datasets like Argoverse means that the system may struggle in environments with less structured data or sensor inaccuracies, such as occlusions and noisy GPS signals.

The imbalance between lane-keeping and lane-changing trajectories also poses broader implications for AV development. While our increased penalties for lateral errors improved lane-changing predictions, the system may still fail in edge cases, such as rapid lane changes or merging under high traffic density. Addressing these scenarios will require not only better modeling but also real-world testing to understand how AVs interact with human drivers in such situations.

Lastly, achieving full autonomy involves challenges beyond technical performance. Ethical considerations, such as decision-making during unavoidable collisions, remain unresolved. Regulatory frameworks for AVs are still evolving, and infrastructure, such as high-definition mapping and vehicle-to-everything (V2X) communication, must be developed to support these systems [33]. As our experiments demonstrate, while trajectory prediction and RL-based planning offer promising solutions, achieving Level 5 autonomy will require a holistic approach that integrates technology, policy, and infrastructure.

A key area of improvement is the integration of multi-agent prediction models, which will enable AVs to better anticipate and respond to interactions in dense and heterogeneous traffic. Expanding RL to continuous action spaces will enable smoother, more natural driving behaviors. Uncertainty-aware trajectory prediction can improve robustness by integrating Bayesian deep learning and Monte Carlo dropout, allowing AVs to quantify uncertainty and adjust decisions dynamically.

A promising approach is hybrid models that integrate physics-based and deep learning techniques for dynamic adaptation. Real-world validation through diverse urban and highway testing, along with sensor fusion (LiDAR, radar, and cameras), will improve perception accuracy. Large-scale simulations and real-world trials will bridge the gap between theoretical performance and practical deployment, ensuring AVs operate with greater computational efficiency.

We suggest, this work demonstrates the feasibility of combining Seq2Seq trajectory prediction and reinforcement learning for autonomous driving. While the results are promising, achieving fully autonomous driving will require addressing significant gaps in model generalization, computational scalability, and adaptation to diverse environments. Ultimately, the path to full autonomy is not just a technological challenge but a multidimensional problem requiring collaboration across domains.

VI. Conclusion

This study presented a framework for trajectory prediction and maneuver generation in autonomous vehicles, combining Seq2Seq-based prediction models with reinforcement learning. The model effectively predicted short-term trajectories by mapping a few seconds of historical position and velocity data to the next seconds of future trajectories. The model demonstrated consistent performance, with validation losses equal to or slightly lower than training losses, suggesting good generalization within the dataset's constraints.

Reinforcement learning was employed to optimize maneuver decisions, with agents such as DQN and Dueling DQN evaluated. Dueling DQN exhibited superior performance in collision-avoidance scenarios due to its separation of state-value and action-advantage estimations, which allowed for better handling of scenarios requiring delayed decision-making. However, the performance of the RL agents was sensitive to reward function design, highlighting the importance of parameterizing rewards to balance safety and efficiency in varying scenarios.

Several limitations were observed. First, the imbalance in trajectory types, such as lane-keeping versus lane-changing, negatively impacted model accuracy despite efforts to mitigate this through weighted loss functions. Second, the computational constraints of training deep learning models on large datasets posed scalability challenges, particularly for real-time applications. Third, while the models performed well on structured datasets, the transition to real-world scenarios, involving dynamic interactions and noisy sensor data, remains an open challenge.

Future research should aim to overcome existing limitations by enhancing data balancing techniques, optimizing computational frameworks, and integrating models with real-world sensory inputs. Refining the underlying algorithms and addressing these challenges will contribute to more reliable trajectory prediction and maneuver planning. Extending the framework to urban driving environments, where traffic patterns are a way more complex, will require incorporating richer environmental features and more adaptive strategies.

References

[1] K. Zhang, L. Zhao, C. Dong, L. Wu and L. Zheng, "AI-TP: Attention-Based Interaction-Aware Trajectory Prediction for Autonomous Driving," in *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 73-83, Jan. 2023, doi: 10.1109/TIV.2022.3155236.

[2] P. Karle, L. Furtner, and M. Lienkamp, "Self-Evaluation of Trajectory Predictors for Autonomous Driving," *Electronics*, vol. 13, no. 5, p. 946, 2024.doi: 10.3390/electronics13050946.

[3] J. Yan, Z. Peng, H. Yin, J. Wang, X. Wang, Y. Shen, W. Stechele, and D. Cremers, "Trajectory prediction for intelligent vehicles using spatial-attention mechanism," *IET Intelligent Transport Systems*, vol. 14, no. 13, pp. 1855–1863, doi: 10.1049/iet-its.2020.0274.

[4] S. Fan, X. Li, and F. Li, "Intention-Driven Trajectory Prediction for Autonomous Driving," in *Proc. IEEE Intelligent Vehicles Symposium (IV)*, 2021, pp. 107–113, doi:10.1109/IV48863.2021.9575253.

[5] Y. Yoon, T. Kim, H. Lee, and J.-H. Park, "Road-Aware Trajectory Prediction for Autonomous Driving on Highways," *Sensors (Basel, Switzerland)*, vol. 20, 2020, doi: 10.3390/s20174703.

[6] Y. Huang, J. Du, Z. Yang, Z. Zhou, L. Zhang and H. Chen, "A Survey on Trajectory-Prediction Methods for Autonomous Driving," in *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 3, pp. 652-674, Sept. 2022, doi: 10.1109/TIV.2022.3167103.

[7] P. Gupta, D. Isele, D. Lee and S. Bae, "Interaction-Aware Trajectory Planning for Autonomous Vehicles with Analytic Integration of Neural Networks into Model Predictive Control," *2023 IEEE International Conference on Robotics and Automation (ICRA)*, London, 2023, pp. 7794-7800, doi: 10.1109/ICRA48891.2023.10160890.

[8] X. Mo, Y. Xing and C. Lv, "Graph and Recurrent Neural Network-based Vehicle Trajectory Prediction For Highway Driving," *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, Indianapolis, IN, USA, 2021, pp. 1934-1939, doi: 10.1109/ITSC48978.2021.9564929.

[9] K. Messaoud, I. Yahiaoui, A. Verroust-Blondet and F. Nashashibi, "Attention Based Vehicle Trajectory Prediction," in *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 1, pp. 175-185, March 2021, doi: 10.1109/TIV.2020.2991952.

[10] B. Kim, C. M. Kang, J. Kim, S. H. Lee, C. C. Chung, and J. W. Choi, "Probabilistic vehicle trajectory prediction over occupancy grid map via recurrent neural network," in *Proc. 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, Yokohama, Japan, 2017, pp. 399–404. doi:10.1109/ITSC.2017.8317943.

[11] C. Anderson, R. Vasudevan and M. Johnson-Roberson, "Low Latency Trajectory Predictions for Interaction Aware Highway Driving," in *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5456-5463, Oct. 2020, doi: 10.1109/LRA.2020.3009068.

[12] X. Chen, H. Zhang, F. Zhao, Y. Cai, H. Wang and Q. Ye, "Vehicle Trajectory Prediction Based on Intention-Aware Non-Autoregressive Transformer With Multi-Attention Learning for Internet of Vehicles," in *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1-12, 2022, Art no. 2513912, doi: 10.1109/TIM.2022.3192056.

[13] B. Khelfa and A. Tordeux, "Lane-changing prediction in highway: Comparing empirically rule-based model MOBIL and a naïve Bayes algorithm," in *Proc. 2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, Indianapolis, IN, USA, 2021, pp. 1598–1603.doi: 10.1109/ITSC48978.2021.9564927.

[14] Y. Yoon and K. Yi, "Trajectory Prediction Using Graph-Based Deep Learning for Longitudinal Control of Autonomous Vehicles: A Proactive Approach for Autonomous Driving in Urban Dynamic Traffic Environments," in *IEEE Vehicular Technology Magazine*, vol. 17, no. 4, pp. 18-27, Dec. 2022, doi: 10.1109/MVT.2022.3207305.

[15] L. Hou, L. Xin, S. E. Li, B. Cheng and W. Wang, "Interactive Trajectory Prediction of Surrounding Road Users for Autonomous Driving Using Structural-LSTM Network," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 11, pp. 4615-4625, Nov. 2020, doi: 10.1109/TITS.2019.2942089.

[16] P. Tran, H. Wu, C. Yu, P. Cai, S. Zheng, and D. Hsu, "What truly matters in trajectory prediction for autonomous driving?" in *Proc. 37th International Conference on Neural Information Processing Systems (NIPS '23)*, Red Hook, NY, USA: Curran Associates Inc., 2024, 3123, pp. 71327–71339, doi: 10.48550/arXiv.2306.15136.

[17] W. Shao, J. Li and H. Wang, "Self-Aware Trajectory Prediction for Safe Autonomous Driving," *2023 IEEE Intelligent Vehicles Symposium (IV)*, Anchorage, AK, USA, 2023, pp. 1-8, doi: 10.1109/IV55152.2023.10186629.

[18] D. Cheng, X. Gu, C. Qian, C. Du and J. Wang, "Vehicle Trajectory Prediction With Interaction Regions and Spatial–Temporal Attention," in *IEEE Access*, vol. 11, pp. 130850-130859, 2023, doi: 10.1109/ACCESS.2023.3335091.

[19] Y. Hu and X. Chen, "Intention-aware Transformer with Adaptive Social and Temporal Learning for Vehicle Trajectory Prediction," *2022 26th International Conference on Pattern Recognition (ICPR)*, Montreal, QC, Canada, 2022, pp. 3721-3727, doi: 10.1109/ICPR56361.2022.9956216.

[20] L. Wang, T. Wu, H. Fu, L. Xiao, Z. Wang and B. Dai, "Multiple Contextual Cues Integrated Trajectory Prediction for Autonomous Driving," in *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6844-6851, Oct. 2021, doi: 10.1109/LRA.2021.3094564.

[21] X. Li, J. Xia, X. Chen, Y. Tan, and J. Chen, "SIT: A Spatial Interaction-Aware Transformer-Based Model for Freeway Trajectory Prediction," *ISPRS Int. J. Geo Inf.*, vol. 11, no. 2, p. 79, 2022, doi: 10.3390/ijgi11020079.

[22] J. Wiest, M. Höffken, U. Kreßel and K. Dietmayer, "Probabilistic trajectory prediction with Gaussian mixture models," *2012 IEEE Intelligent Vehicles Symposium*, Madrid, Spain, 2012, pp. 141-146, doi: 10.1109/IVS.2012.6232277.

[23] N. Assymkhan and A. Kartbayev, "Advanced IoT-Enabled Indoor Thermal Comfort Prediction Using SVM and Random Forest Models" International Journal of Advanced Computer Science and Applications (IJACSA), 15(8), 2024. doi:10.14569/IJACSA.2024.01508102.

[24] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, D. Ramanan, P. Carr, and J. Hays, "Argoverse 2: Next Generation Datasets for Self-driving Perception and Forecasting," in *Proc. Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks)*, 2021. doi**:** 10.48550/ARXIV.2301.00493.

[25] G. Li, Y. Jiao, V. Knoop, S. Calvert, and J. W. C. Lint, "Large car-following data based on Lyft Level-5 Open Dataset: Following autonomous vehicles vs. human-driven vehicles," *arXiv preprint*, arXiv:2305.18921, 2023. https://doi.org/10.48550/arXiv.2305.18921.

[26] A. James and E. Bakolas, "Gaussian Mixture Based Motion Prediction for Cluster Groups of Mobile Agents," *IFAC-PapersOnLine*, vol. 55, no. 37, pp. 408–413, 2022. doi: 10.1016/j.ifacol.2022.11.217.

[27] A. Sharma, D. Pantola, S. Kumar Gupta and D. Kumari, "Performance Evaluation of DQN, DDQN and Dueling DQN in Heart Disease Prediction," *2023 Second International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, Singapore, Singapore, 2023, pp. 5-11, doi: 10.1109/SmartTechCon57526.2023.10391350.

[28] N. Smatov, R. Kalashnikov, and A. Kartbayev, "Development of context-based sentiment classification for intelligent stock market prediction," *Big Data Cogn. Comput.*, vol. 8, 51, 2024. doi: 10.3390/bdcc8060051.

[29] Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas, "Dueling network architectures for deep reinforcement learning," in *Proc. 33rd Int. Conf. Machine Learning (ICML'16)*, JMLR.org, 2016, pp. 1995–2003, doi: 10.48550/arXiv.1511.06581 .

[30] S. H. Park, B. Kim, C. M. Kang, C. C. Chung and J. W. Choi, "Sequence-to-Sequence Prediction of Vehicle Trajectory via LSTM Encoder-Decoder Architecture," *2018 IEEE Intelligent Vehicles Symposium (IV)*, Changshu, China, 2018, pp. 1672-1678, doi: 10.1109/IVS.2018.8500658.

[31] S. Rezwana and N. Lownes, "Interactions and Behaviors of Pedestrians with Autonomous Vehicles: A Synthesis," *Future Transportation*, vol. 4, pp. 722–745, 2024. doi: 10.3390/futuretransp4030034.

[32] S. A. Bagloee, M. Tavana, M. Asadi, and T. Oliver, "Autonomous vehicles: challenges, opportunities, and future implications for transportation policies," *Journal of Modern Transportation*, vol. 24, pp. 284–303, 2016. doi: 10.1007/s40534-016-0117-3.

[33] S. A. Yusuf, A. Khan, and R. Souissi, "Vehicle-to-everything (V2X) in the autonomous vehicles domain – A technical review of communication, sensor, and AI technologies for road user safety," *Transportation Research Interdisciplinary Perspectives*, vol. 23, 2024. doi: 10.1016/j.trip.2023.100980.

Modular Analysis of Complex Products Based on Hybrid Genetic Ant Colony Optimization in the Context of Industry 4.0

Yichun Shi[1]*, Qinhe Shi[2]
Department of Information Engineering, Jiangsu Union Technical Institute, Nanjing, 210000, China[1]
School of Accounting, Nanjing University of Finance and Economics, Nanjing, 210000, China[2]

***Abstract*—With the development of science and technology, industrial construction has entered the era of 4.0 intelligent construction, and various algorithms have been widely applied in the modularization of production products. This study focuses on the modular optimization problem of complex products and establishes a hybrid genetic algorithm based on the ant colony algorithm framework. The new algorithm incorporates visibility analysis of the genetic algorithm, using the obtained solution as the pheromone source for the new algorithm to quickly obtain the optimal solution. The results showed that the algorithm could quickly achieve modularization of complex industrial products, adapt to products with a large number of parts and complex compositions, and obtain the optimal solution. The new algorithm reduced the running time of modular complex products by 35.06% compared to the particle swarm optimization algorithm. The new algorithm optimized the product design process for core components, reducing production costs by 23.46% and increasing production efficiency by 39.20%. Consequently, the novel algorithm modularizes complex products, thereby enhancing production efficiency and providing a novel intelligent method for the design process of complex products.**

Keywords*—Industry 4.0; genetic algorithm; ant colony; complex products; modularization; production efficiency

I. Introduction

In recent years, technological advancements have made traditional manufacturing unable to meet the current society's production needs for complex products [1]. Industry 4.0 is not only an upgrade to traditional supply chain automation and monitoring, but also builds a highly interconnected ecosystem and intelligent driven products through deep integration of intelligent technology [2]. The core concept is to utilize advanced information and communication technology to promote the transformation of the manufacturing industry towards a more flexible, efficient, and personalized intelligent manufacturing model [3]. In this context, the hybrid Genetic Algorithm (GA) has become a powerful tool for solving multidimensional and nonlinear optimization problems in the design of Complex Products Modularization (CPM) due to its unique advantages of strong global search capability, flexible genetic mechanism, and compatibility [4]. The objective of modular design is to disaggregate complex products into a series of relatively autonomous and functionally distinct modules. Through the combination and reconstruction of diverse modules, the design process can rapidly adapt to market demands, enhancing design efficiency and product flexibility [5].

In response to the demand from various sectors of society for the construction of industrial responsibility products in the context of Industry 4.0, a large number of scholars have conducted extensive research on CPM. Li Y et al. proposed a CPM method that combines modularity and design change propagation scope to reduce the impact of design change propagation. This method constructed the adjacency matrix of a weighted directed network model and solved the model using non-dominated sorting GA. The results of verifying the modularity of the driver's cab of a specific electronic sanitation vehicle showed that this method was practical and effective [6]. Wang X et al. proposed a moderated mediation model to address the issue of product modularization in R&D outsourcing practices. Based on survey data from 273 Chinese manufacturing enterprises, hypotheses were tested using hierarchical regression and PROCESS macro-models. The role of product modularization in R&D outsourcing practice was more effective when the trust level of R&D outsourcing partners was high [7]. Wang S et al. proposed a sub-item method oriented towards core components to address the challenges of structural modeling difficulties and unreasonable modular solutions in CPM for complex product sub-item problems. The new method simplified the structural model of complex products, further reduced the difficulty of modeling, and improved the efficiency of solving module partitioning schemes [8]. Forti A W et al. proposed a new structural matrix modularization method to solve the problem of difficult integration of multi-component products in the automotive industry, which combines the use of quality function deployment and design attribute matrix indication matrix. This systematic modular process has effectively played a role, making cross-functional teamwork easier [9].

The application value of hybrid algorithms as an efficient optimization strategy in CPM design is becoming increasingly prominent. Zhao J et al. proposed a multi-ACA approach that combines community relationship networks to address the challenge of balancing solution accuracy and convergence speed in large-scale TSP for the Ant Colony Algorithm (ACA). It improved the accuracy of the solution by collecting the route information of all ants and constructing a route relationship network. The performance of the new algorithm in large-scale TSP was significantly better than other improved algorithms

[10]. Arasteh B et al. proposed a hybrid method based on a grey wolf optimization algorithm to solve the problems of poor CPM quality, low success rate, and limited stability in existing methods, to achieve sub-items of complex products. The new algorithm improved the clustering quality and outperformed other heuristic algorithms in terms of modularity and convergence speed [11]. Liu C et al. proposed an improved discrete imperialist competitive hybrid algorithm to address product design compatibility and quality issues, considering a nonlinear programming approach that maximizes the per capita contribution margin of reliability loss. This hybrid algorithm has improved the solution quality by 6%~17% and 5%~14% compared to GA and simulated annealing algorithms [12].

In summary, the combination of modularity and the scope of design change propagation effectively reduces the impact of design changes. The trust-based product modularization method has improved the modularization efficiency of R&D outsourcing. These methods provide multidimensional ideas for solving CPM. In terms of CPM, most algorithms have only implemented modularity, and there has not been much research on achieving high-efficiency production. Based on this, this study proposes a new hybrid genetic ACA. To adapt to the rapid modularization with fewer components, this study innovatively integrates operations such as mixing, crossover, and mutation into ACA to quickly obtain visibility. To better obtain the optimal solution, the genetic part of the solution is used as a pheromone to enable the ant colony to quickly obtain the optimal solution. Pheromones are chemical substances that are synthesized by ants in ACAs and serve to guide other ants in navigating their environment. The issue of information redundancy in multi-part problems has been addressed by the adoption of ant colony solving, which has supplanted genetic solving. This development has enabled the implementation of a novel algorithm capable of adapting to CPM in multi-part scenarios, thereby facilitating the enhancement of production efficiency for complex products. The article is divided into five sections in total. The first section introduces the research status and importance of CPM under the background of Industry 4.0. The second section elaborates on the framework of Hybrid Genetic Ant Colony Optimization (HGACO) algorithm and its application in CPM. The third section verifies the performance of HGACO algorithm through experiments and compares it with other algorithms for analysis. The fourth section discusses the article and provides personal insights and opinions. The fifth section summarizes the research conclusions and proposes future research directions.

II. METHODS

A. Establishment of Hybrid Genetic Ant Colony Algorithm

In the era of Industry 4.0, traditional design methods cannot meet modern needs. This study designs an innovative HGACO method based on the ACA framework. Through a unique encoding and decoding mechanism, HGACO optimizes the decomposition process of product components. Simplifying complex products will be beneficial for industrial production. GA solving product modularization problems requires the use of various genetic components on chromosomes. The dimensions of each component are represented by chromosome length, forming an initial group [13]. The size of the fitness function determines the quality of an individual. The fitness function is transformed from the objective minimum function, as shown in Eq. (1).

$$F(t) = \frac{1}{Q_m} \tag{1}$$

In Eq. (1), $F(t)$ is the fitness function. Q_m is the shortest path. The formula for selecting arithmetic factors using the turntable method is shown in Eq. (2).

$$F(t) = \begin{cases} f(t) - C_{\max}, f(t) > C_{\max} \\ 0, f(t) = C_{\max} \\ C_{\max} - f(t), f(t) < C_{\max} \end{cases} \tag{2}$$

In Eq. (2), $f(t)$ and $C_{\max}$ are the values and maximum values of the fitness function. The initial population uses a random method to generate multiple chromosomes. The size of chromosomes represents the size of a population. Genetic individuals are screened, and the selected excellent individuals are subjected to subsequent crossover mutations. The selection method adopts the turntable method, and the probability of being selected is determined by the ratio size of individuals, as shown in Eq. (3).

$$P_a = \frac{f_a}{\sum_{a=1}^{N_z} f_a} \tag{3}$$

In Eq. (3), P_a is the probability of being selected, N_z is the individual fitness value, and $\sum_{a=1}^{N_z} f_a$ is the population fitness value. The selected excellent individuals are subjected to crossover operations to generate new offspring individuals. The crossover probability is shown in Eq. (4).

$$P_b = \begin{cases} \dfrac{\alpha(f_d - f_b^d)}{f_d - f_p}, f_b^d \geq f_p \\ \beta, f_b^d < f_p \end{cases} \tag{4}$$

In Eq. (4), P_b is the crossover probability. f_d and f_p are the maximum and average fitness values of the population. f_b^d is the party with a higher fitness value participating in the crossover process. α and β are constants. After selection and crossover, to enhance the comprehensive performance of the algorithm's retrieval ability, mutation is also required. The mutation probability is shown in Eq. (5).

$$P_c = \begin{cases} \dfrac{\gamma(f_d - f_b)}{f_d - f_p}, f_b \geq f_p \\ \lambda, f_b < f_p \end{cases} \tag{5}$$

In Eq. (5), P_c is the mutation probability, f_b is the fitness value of the mutated individual, and γ and λ are constants. The calculation process of GA is shown in Fig. 1.

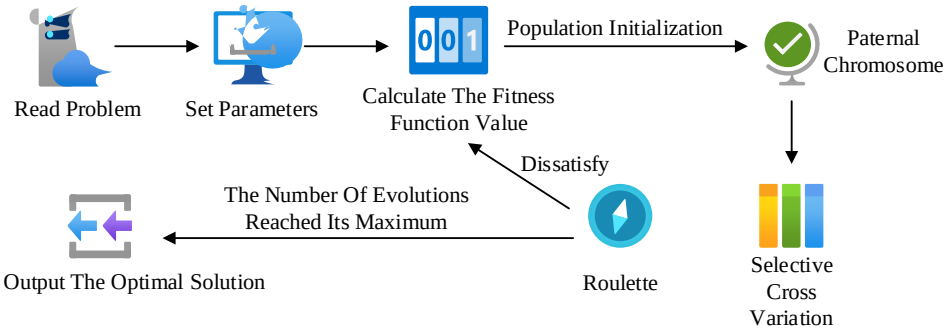


Fig. 1. Ga solution flow chart.

In Fig. 1, in the overall process of GA solving, step 1 is to read the problem being solved and set the corresponding parameters. Step 2 is to set the fitness function to calculate the fitness value, and initialize the population to produce parental chromosome individuals. The next step is to operate on individuals based on the selection, crossover, and mutation probability formulas, using the turntable method to select the transformed individuals. If the individual has reached the maximum number of evolutions, the solution value is directly output, otherwise it returns to the mutation step. GA is integrated into ACA, which is an exploratory algorithm with features of group cooperation, simultaneous computation, and forward selection. Each ant moves according to a fixed rule [14]. The number of ants is shown in Eq. (6).

$$N = \sum_{a=1}^{n} M_a(t) \tag{6}$$

In Eq. (6), N is the number of ants. $M_a(t)$ is the number of ants at time t. The information of each path of ants during the transfer process varies, and the probability calculation method of ant movement is shown in Eq. (7).

$$P_{ab}^{x} = \begin{cases} \dfrac{\tau_{ab}^{p}\eta_{ab}^{q}}{\sum_{S \in xl_x} \tau_{as}^{p}\eta_{as}^{q}}, b \in xl_x \\ 0, b \notin xl_x \end{cases} \tag{7}$$

In Eq. (7), P_{ab}^{x} is the probability that ant x moves from point t to point b at time a. τ_{ab}^{p} is the number of pheromones from point a to point b under the information heuristic factor p. η_{ab}^{q} is the visibility from point a to point b. xl_x is a point that ants have not passed through. After the ant completes a path selection, the pheromones on other paths are shown in Eq. (8) [15].

$$\tau_{ab}t = \varphi \cdot \tau_{ab}t + \Delta\tau_{ab}t \tag{8}$$

In Eq. (8), $\tau_{ab}t$ is the number of pheromones from point a to point b after the t-th cycle. φ is the residual coefficient of pheromones. $\Delta\tau_{ab}t$ is the residual value of pheromones. The pheromone changes of Ant x from point a to point b during a path iteration are shown in Eq. (9).

$$\Delta\tau_{ab}^{x} = \begin{cases} \dfrac{k}{H(C)}, x \text{ participation loop} \\ 0, \text{else} \end{cases} \tag{9}$$

In Eq. (9), $\Delta\tau_{ab}^{x}$ is the pheromone change of ant x from point a to point b in one path iteration. k is the intensity of pheromones. $H(C)$ is the sum of ant journeys after multiple iterations. Fig. 2 shows the calculation process of ACO.

In Fig. 2, the system first initializes the data by randomly placing ants at various points and then begins the loop. According to the formula for calculating the probability of ant movement, the ants are moved. The points that ants pass through are marked. The ant colony cycle model is used to determine whether ants have passed through all points. If they have not passed through all points, the number of cycles is increased by 1. If the ant passes through all points, the pheromone is updated according to the pheromone calculation formula. If the calculated path is the shortest path, the path is updated; otherwise, the ant is asked to select a new point and output the shortest path. ACA lacks pheromones in the early stage, so the process of searching for the optimal solution is relatively slow. GA has high adaptability and fast running speed, but it is prone to generating redundant junk information. By integrating two algorithms, GA can solve the problem of the slow operation of ACA in the early stage. In the later stage, the accumulation of pheromones in ACA reaches a certain level, which increases the running speed and avoids the situation where GA continues to run and generate a large amount of junk information. The key to mixing GA and ACA is to find the time point for mixing. The appropriate timing can enable HGACO to achieve optimal performance. The process of the HGACO algorithm is shown in Fig. 3.

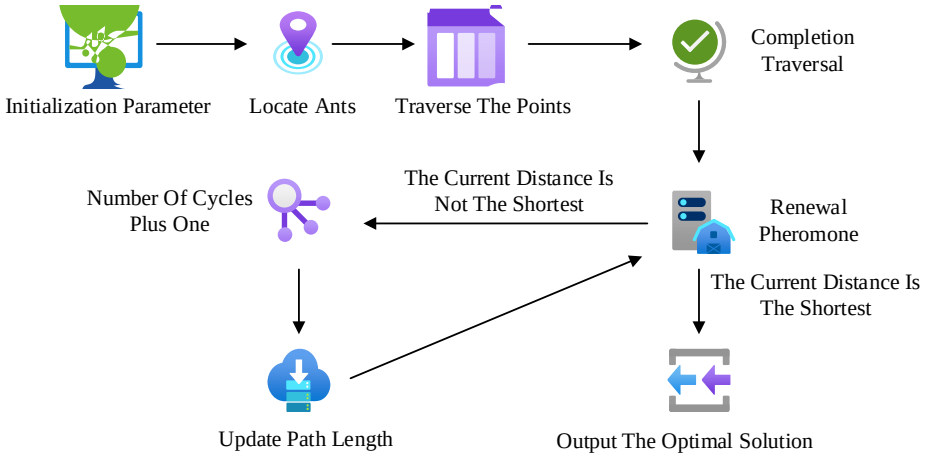


Fig. 2. Flow chart of ACA solution.

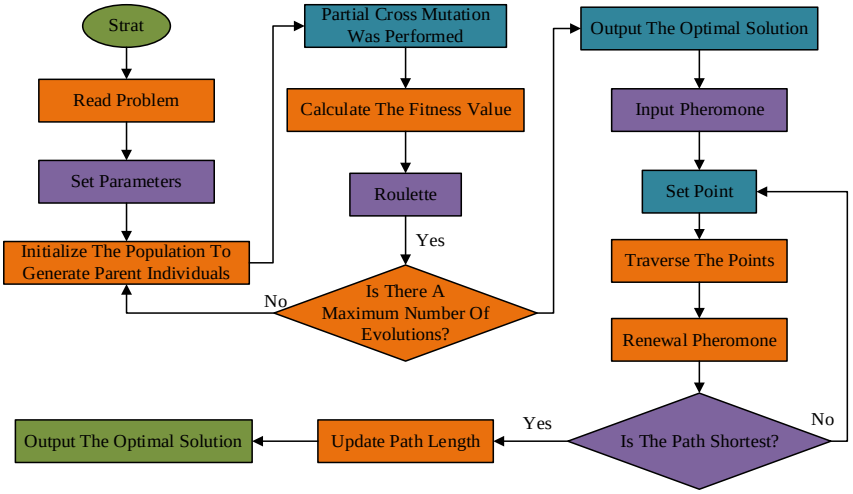


Fig. 3. Flow chart of HGACO.

In Fig. 3, the HGACO algorithm first calculates the optimal solution through the GA part, converting the optimal solution into pheromones while leaving some visibility information. Visibility information refers to the quality information of solutions in GAs, which is used to guide ACAs to quickly find the optimal solution. The visibility data are numbered to represent the spatial numbering in the CPM problem-solving process. The pheromones transformed into the optimal solution are transmitted to the ACA section, and the optimal solution is obtained through the mechanism of simultaneous calculation in the ACA section. Finally, the optimal solution obtained by the HGACO algorithm is outputted. By utilizing the HGACO algorithm, modular processing of complex products can assist in production design. Complex products often have a large number of parts. Modular processing can simplify these parts and reduce the complexity of encoding and decoding.

B. CPM Design

The use of HGACO algorithm for CPM can solve the production efficiency problem in the context of Industry 4.0. The main factor affecting production efficiency is that product quality requirements are often high, and most products are customized with complex structures [16]. These reasons make product production relatively slow in the design process. After disassembling the information on these customized products, the indicators of each part can be obtained. The various components are interrelated. Fig. 4 is a complex product decomposition structure diagram.

In Fig. 4, the internal relationships within the structure of complex products are complex. Various large devices are composed of different components. Components are composed of different components, and components are composed of multiple part modules, which contain multiple small parts. The components at different levels are connected in a complex manner through serial and parallel connections. The logistics list is used to organize components at different levels, determine their mutual demand relationships, and thus form integration. The adaptation relationship between different components will affect the quality performance of the product. There are various indicators of product quality. Reliability is the most important quality indicator. Fig. 5 shows the serial structure between the parts.

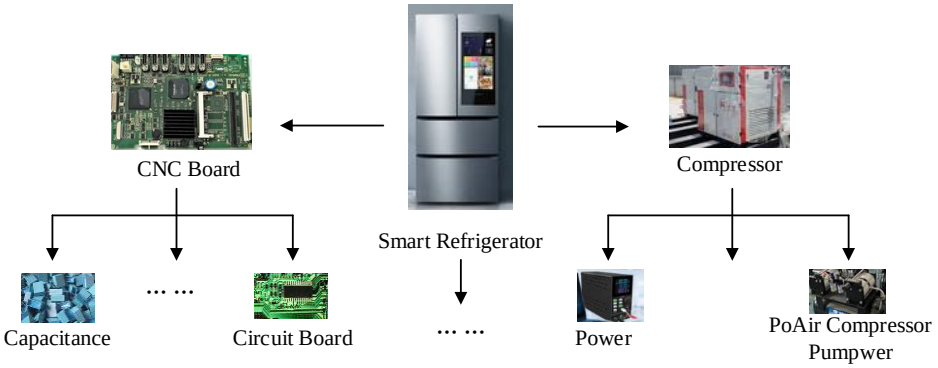


Fig. 4. Complex product breakdown structure diagram.

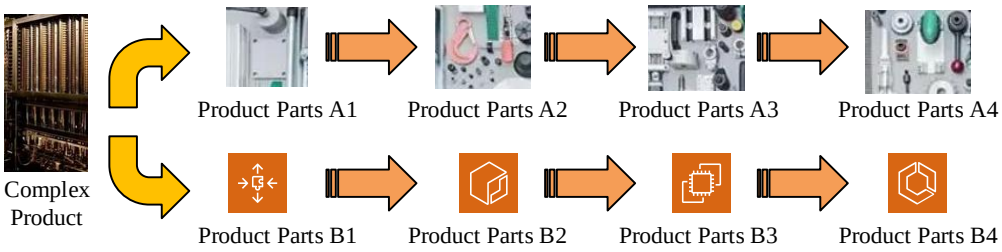


Fig. 5. Series system result model diagram.

In Fig. 5, A1, A2, A3, A4, B1, B2, B3, and B4 respectively represent the component numbers. In a serial system, each component module is sequentially connected to form a linear sequence. The reliability of each component module directly affects the performance of its subsequent modules. Consequently, the overall reliability of the system depends on the reliability of each component module. The characteristic of this structure is that the failure of any component module may lead to the failure of the entire system. Multiple serial components together form a component module. In actual production, these component modules form a propagation mode of parallel connections. Parallel connected components further form components, and most of the components are interleaved in series and parallel. Among them, the reliability between component modules affects each other. The overall reliability of complex products is shown in Eq. (10).

$$D(x) = K(X_1 > x, X_2 > x, ...X_m > x) = \prod_{a=1}^{m} K(X_a > x) = \prod_{a=1}^{m} D_a(x) \quad (10)$$

In Eq. (10), $D(x)$ represents the overall reliability of the product system. K is the reliability coefficient. X_m is the reliability prediction of the m-th component. x is the reliability standard. $\prod$ is a quadrature sum operation. The reliability of the system decreases with the increase of the number of components, and the decrease in system reliability is reflected in the failure rate [17]. The lower the failure rate of components, the higher the reliability of the system. The calculation method for failure rate is shown in Eq. (11).

$$\eta_m = \sum_{a=1}^{m} \eta a \quad (11)$$

In Eq. (11), η_m is the overall failure rate of the system. ηa is the failure rate of each component. Based on the calculation method of failure rate, the average time for system failure is shown in Eq. (12).

$$T_{bf} = \frac{1}{\eta_m} = \frac{1}{\sum_{a=1}^{m} \eta a} \quad (12)$$

In Eq. (12), T_{bf} is the average time between failures, which is inversely proportional to the overall failure rate of the system. If one of the components is adjusted and integrated into parallel mode, the overall reliability calculation method of the system will be adjusted as shown in Eq. (13).

$$D = D_1 \times [1-(1-D_2)^n] \times \prod_{a=3}^{m} D_a, n \in (1, \infty) \quad (13)$$

In Eq. (13), D is the system reliability with parallel features, and $D \in (0,1)$. D_a is the reliability of the a-th component. n is the total number of parts. The calculation method of the first derivative of the reliability of the system as a whole based on the number of parts is given by Eq. (14).

$$\frac{dD}{dn} = -(1-D_2)^n \ln(1-D_2) D_1 \prod_{a=3}^{m} D_a > 0 \quad (14)$$

In Eq. (14), $\frac{dD}{dn}$ is the first derivative of system reliability. With parallel connections in the system, as the number of parts increases, the system becomes significantly more reliable. The expression for the second derivative is given by Eq. (15).

$$\frac{d^2D}{dn^2} = -(1-D_2)^n \ln^2(1-D_2) D_1 \prod_{a=3}^{m} D_a < 0 \quad (15)$$

Complex products often contain a large number of

component modules. Different component modules form various complex system structures. Serial and parallel modes are often mixed together, making the factors affecting system reliability more complex [18]. There is interaction between the components of the product. The impact of interaction is called the influencing factor. There is no interactive influence relationship between the components that make up the parallel state, and there is no subsequent performance impact. The influence relationship between the components that form a mixed state of serial and parallel is relatively complex, and when improving the system, it is greatly affected by the interference variables of the components. The coupling between different levels can improve the overall quality of the system by selecting the quality, quantity, and shape of parts [19]. Based on the analysis of serial and parallel quality performance of parts, this study uses the HGACO algorithm to integrate component modules in a hybrid structure and analyze the impact mechanisms of multiple component systems in the hybrid structure. By reallocating different components and configuring quality performance parameters, more precise and improved system modules are obtained [20]. Fig. 6 shows the technical process of using the HGACO algorithm for CPM.

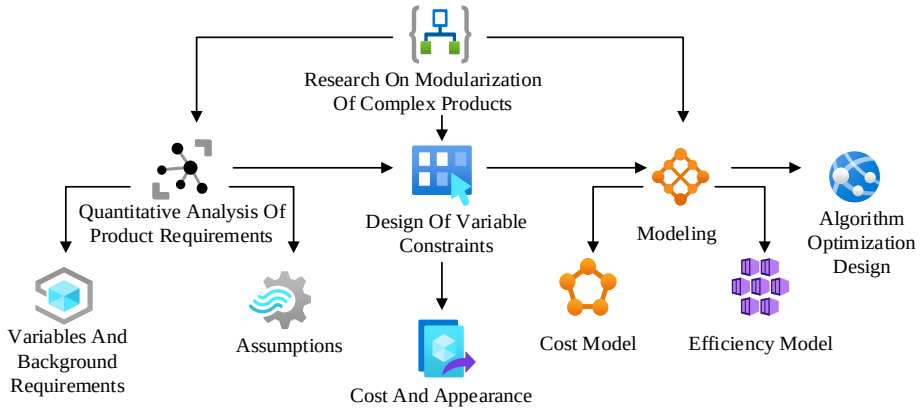


Fig. 6. Technical flow chart of modular design of complex products.

In Fig. 6, the first step is to conduct a quantitative analysis of product requirements, including functional description, size, and cost requirements. Subsequently, design constraints for variables such as cost and appearance. During the design process, it is also necessary to explain the constraints. After obtaining a series of required parameters, a basic cost and profit model for complex products is constructed. The study is predicated on the fundamental model to optimize the design of the HGACO algorithm. Ultimately, the HGACO algorithm is utilized for calculation to obtain the optimal functional module configuration scheme.

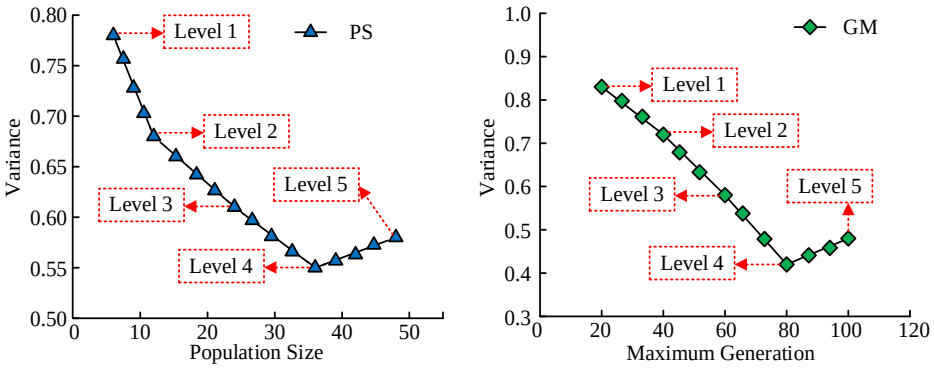
III. Results

A. Performance Analysis of Mixed Genetic ACA

Simulation experiments are designed to analyze the performance of the HGACO algorithm and solve the CPM analysis problem. The experimental language is Java, and the model solving is Matlab. The operating environment is Windows 10, the CPU is Intel Core i7-2600 3.40 Ghz, and the RAM is 4GB. The performance of an algorithm mainly depends on its control parameters. The control parameters are mainly divided into Iterations (GM), Mutation Rate (PM), Population Numbers (PS), and Crossover Rate (PC). The Analysis of Variance (ANOVA) of the algorithm under different parameter controls reflects its performance, with smaller values indicating better performance and greater stability. Fig. 7 shows the variance of PS and GM corresponding to the HGACO algorithm at different levels.

The statistical significance of the variance data in Fig. 7 is shown in Table I. In Fig. 7 (a), PS increases with the increase of the number of levels. When PS is 36, the variance has its minimum value, and as PS continues to expand, there is no significant change in variance. In Fig. 7 (b), when GM is less than 80, the variance shows a stable downward trend, but when GM is greater than 80, the variance suddenly increases. This is because when the iteration parameters are too high, the HGACO algorithm is prone to lagging due to the concentration of pheromones, leading to an increase in ANOVA. Therefore, the optimal PS is 36 and the optimal GM is 80. There are crossover and mutation processes in the HGACO algorithm, and the corresponding PC and PM also have an impact on the variance. Fig. 8 shows the corresponding variances of PC and PM at different levels.

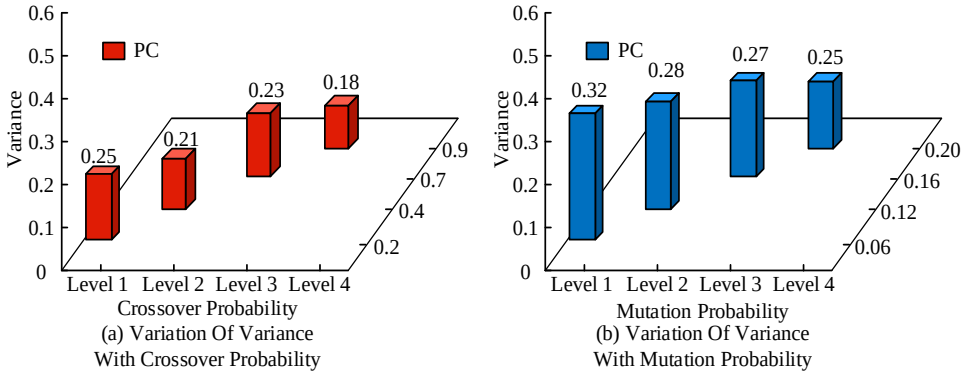
In Fig. 8 (a), the ANOVA transformation is less affected by PC, and the variance is minimized when PC is 0.90. In Fig. 8 (b), the variance is minimized when PM is 0.20. This is because cross-selection makes it easier to generate offspring with significant differences from the parent, while mutation operations have relatively less impact on offspring, so the mutation probability is less affected by hierarchy and has little effect on variance. The PC of the HGACO algorithm is 0.90, and the PM is 0.20. This condition has the best effect on generating the optimal population for the algorithm. Fig. 9 shows the initial values of the independent variables and the distribution of the independent variables after multiple iterations in the HGACO algorithm.



(a) Variation Of Variance With Population Size

(b) Variation Of Variance With Maximum Generation

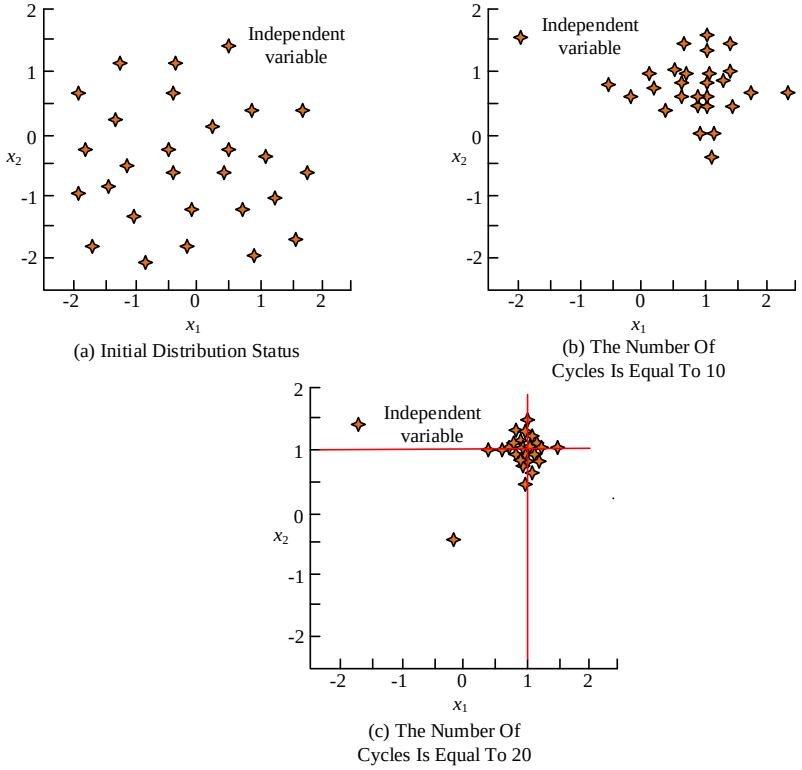
Fig. 7. The Variance of PS and GM of the algorithm at different levels.



(a) Variation Of Variance With Crossover Probability

(b) Variation Of Variance With Mutation Probability

Fig. 8. The Variance of PC and PM of the algorithm at different levels.



(a) Initial Distribution Status

(b) The Number Of Cycles Is Equal To 10

(c) The Number Of Cycles Is Equal To 20

Fig. 9. Distribution of independent variables under different cycles.

TABLE I STATISTICAL ANALYSIS OF VARIANCE DATA

PS		**GM**	
Population Size	variance	Population Size	variance
5	0.78	20	0.83
7	0.75	27	0.8
10	0.71	33	0.75
13	0.67	40	0.71
18	0.65	47	0.68
21	0.63	53	0.62
24	0.60	60	0.58
28	0.58	66	0.55
31	0.56	70	0.48
36	0.54	75	0.45
38	0.55	80	0.41
41	0.56	87	0.44
44	0.57	92	0.46
48	0.58	100	0.48
P	0.0001	P	0.00005
F	12.34	F	15.67
95% Confidence intervals	[0.59,0.67]	95% Confidence intervals	[0.50,0.72]
Standard deviations	0.07	Standard deviations	0.19

In Fig. 9 (a), the initial value distribution of the ant colony is quite scattered and needs to undergo a self changing cycle. In Fig. 9 (b), when the number of iterations reaches 10, the independent variable distribution of the HGACO algorithm begins to converge towards the vicinity of the first independent variable $x_1 = 1$ and the second independent variable $x_2 = 1$. In Fig. 9 (c), when the number of self-variable loops reaches 20, the independent variables of the algorithm converge well. The independent variables of the algorithm gradually converge and reach a convergence state. Continuing to increase the number of loops will cause an unnecessary burden on the computational part of the algorithm. Therefore, the optimal number of iterations for the HGACO algorithm is 20, which results in the best convergence. After determining that the HGACO algorithm can achieve optimal performance under the above parameter conditions, it is necessary to analyze other performance indicators of the algorithm. The most intuitive way to determine whether an algorithm is optimal is to compare its performance with other algorithms under the same conditions.

B. Performance Comparison of Modularization of Complex Products using different Algorithms

In the previous section, various optimal parameters are determined through performance analysis of the HGACO algorithm. To validate the performance of the HGACO algorithm, the Swarm Behavior Heuristic Algorithm (SBH) and the traditional Particle Swarm Optimization (PSO) algorithm are compared with the proposed HGACO algorithm. This experiment compares the algorithm performance under CPM using algorithms. The modularization process is to use different algorithms to modularize complex products with m parts and n associations, where $m \in \{10, 20, 30, 40, 50\}$ corresponds to $n \in \{20, 40, 60, 80, 100\}$ and the problem scale is represented as x_{mn}. The training and validation running times of each algorithm on the DSM dataset are displayed in Fig. 10.

In Fig. 10 (a), during the training process, the initial running time of the HGACO algorithm is relatively long. However, as the number of parts increases, the running time of HGACO is shorter than that of PSO, while the running time of SBH always remains linear. This is because HGACO has insufficient pheromones at the beginning, and the process of cross-selection and mutation takes a long time. However, with the extension of training and the accumulation of pheromones, the speed at which HGCAO seeks the optimal solution increases. In Fig. 10 (b), due to the training of HGACO, it can maintain a relatively short running time throughout the validation process. The average running time of HGACO has been reduced by 35.06% compared to PSO. Table I shows the maximum modular value standard deviation and mean of HGACO, SBH, and PSO during the modular process.

In Table II, when there are few parts, all algorithms can obtain reasonable solutions, with an average value generally within 10. As the data size increases, the average values obtained by PSO and SBH show significant discrepancies, with some solutions even reaching 48900, which is clearly unreasonable. Due to the different solving rules of the algorithms, the model solutions obtained by PSO and SBH do not meet the requirements when there are too many parts. Therefore, only when the optimal solution of HGACO is within a reasonable range in all cases, can it meet the requirements. The production efficiency, duration, and cost of modularizing four complex products C_1, C_2, C_3, and C_4 using three algorithms are shown in Fig. 11.

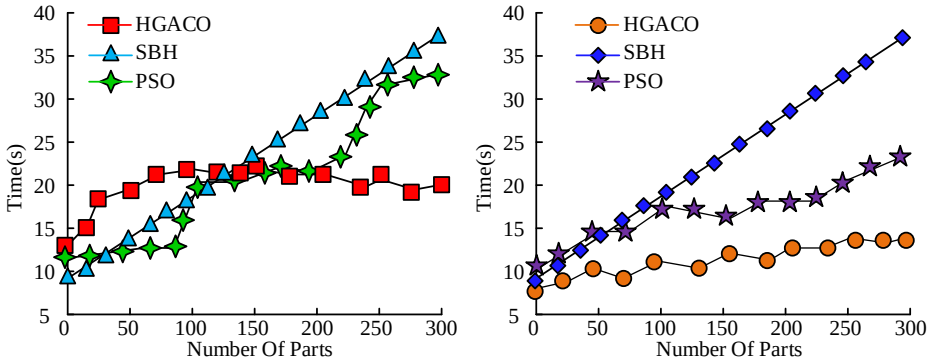


Fig. 10. Algorithm training and validation run time on DSM data set.

TABLE II COMPARISON OF DIFFERENT ALGORITHMS ON MODULARITY VALUES

	HGACO		**PSO**		**SBH**	
Problem Scale	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
x_{1020}	5.87	4.52	4.52	6.25	5.65	6.58
x_{2040}	7.82	7.16	7.16	8.02	2158	1589
x_{3060}	6.21	5.55	145	165	23.1	18.5
x_{4080}	4.59	6.21	3.54	3.51	256	298
x_{50100}	6.51	8.51	48900	36580	10.1	9.99

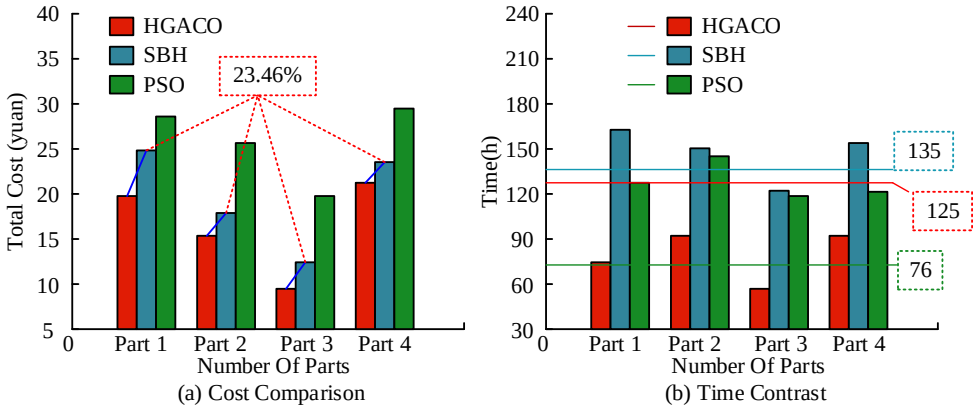


Fig. 11. Production data after modularization of complex products by algorithms.

In Fig. 11 (a), the production costs of modularizing different products using the HGACO algorithm are 19.98 yuan, 15.84 yuan, 9.21 yuan, and 21.02 yuan, respectively. The production costs after modularization using the SBH algorithm are 24.89 yuan, 17.94 yuan, 13.25 yuan, and 23.55 yuan, respectively. Compared with the SBH algorithm, the HGACO algorithm reduces the production costs of different products by 24.69%, 13.25%, 43.87%, and 12.03%, respectively, saving an average of 23.46% of production costs. In Fig. 11 (b), the optimal solution obtained by HGACO can significantly improve efficiency. After modularizing complex production, the average production time is 125 hours, while the average production time under the PSO algorithm is 76 hours. Compared with the ABC algorithm, the HGACO algorithm reduces production time by 39.20%, resulting in a 39.20% increase in production efficiency. This is because HGACO has made product design simpler and more efficient after CPM. Enterprises can carry out intelligent production based on simpler design solutions. Therefore, by modularizing complex production and solving it, HGACO can significantly improve the production efficiency of complex products and reduce production costs.

IV. DISCUSSION

With the advancement of Industry 4.0, CPM design has become a key means to improve production efficiency and reduce costs. The conventional manufacturing paradigm proves challenging in meeting the demands of intricate product development. Conversely, CPM exhibits a capacity to expeditiously adapt to market fluctuations, enhancing design efficiency and product adaptability. In recent years, numerous scholars have devoted themselves to studying CPM and proposed various algorithms, such as GA, ACA, and PSO. While these methods have proven advantageous in addressing modular problems, they are not without limitations. These

limitations include but are not limited to, insufficient applicability and low operational efficiency in multi-component complex products. To improve the operational efficiency of CPM, a CPM optimization method based on HGACO has been proposed. This method combines the global search capability of GA and the pheromone optimization mechanism of ACA. The transformation of the solution of GA into a pheromone, followed by its transmission to ACA, results in the rapid convergence and efficient optimization of the system. In the experimental section, the superiority of the HGACO algorithm is verified by comparing its performance with other algorithms. With respect to the duration of execution, the HGACO algorithm exhibits a longer execution time during the initial training stages. However, as the number of components increases, the execution time of the HGACO algorithm experiences a gradual decrease in comparison to the PSO algorithm. This is primarily due to the necessity for HGACO to amass a sufficient quantity of pheromones during the initial training phases to facilitate ant colony navigation. As pheromones are accumulated, the efficacy of the algorithm for search purposes undergoes a substantial enhancement. A comparison of modular values reveals that the HGACO algorithm can obtain reasonable modular values under different problem scales. In contrast, the PSO and SBH algorithms demonstrate significant deviations in the standard deviation and mean of modular values when there are a large number of components. This indicates that the HGACO algorithm has higher stability and adaptability when dealing with complex products. The underlying rationale pertains to the efficacy of the HGACO algorithm in circumventing local optima through the integration of crossover and mutation operations of GAs. This is complemented by the utilization of the pheromone mechanism of ACA, which facilitates the acceleration of global search, thereby ensuring the maintenance of optimal performance in complex environments.

In summary, the HGACO algorithm has shown significant advantages in the modular design of complex products. It not only outperforms traditional algorithms in terms of running time but also demonstrates excellent stability and adaptability in modular values. These results indicate that the HGACO algorithm can effectively address the challenges in CPM design. Despite the demonstrated efficacy of the HGACO algorithm in experimental and practical test results, its performance may be constrained by the increasing complexity of products and the scale of production systems that accompany the advancement of Industry 4.0. The performance of the HGACO algorithm is crucial in handling large-scale complex products. Future research needs to further enhance the scalability of algorithms, enabling them to efficiently handle large-scale complex products. By using distributed computing and parallel processing techniques, the computational tasks of algorithms can be allocated to multiple processors or computing nodes, significantly improving the execution efficiency of algorithms to meet more complex industrial needs.

V. Conclusion

This study mainly focused on the modular processing and analysis of complex products using algorithms in the context of Industry 4.0. A new HGACO algorithm has been proposed to further improve production efficiency. This study first extracted the problem, then initialized the parameters, and solved it through GA selection, crossover, and mutation. The obtained solution was utilized as a pheromone and visibility information to input into the ACA part to quickly obtain the optimal solution. Finally, experimental verification and comparison were conducted on the optimal solutions obtained by different algorithms. When the PS of the HGACO algorithm was 36, GM was 80, PC was 0.90, and PM was 0.20, the variance was minimized, indicating that the algorithm has the best performance at this time. When the number of cycles reached 20, HGACO just converged and could adapt to complex products with different numbers of parts, and could obtain reasonable optimal solutions. By using HGACO to modularize the responsible products, the running time was reduced by 35.06% compared to the PSO algorithm. HGACO has improved the production efficiency of complex products by 39.20% while reducing production costs by 23.46%. In summary, the comprehensive performance of HGACO is superior to other traditional algorithms, providing a theoretical basis for manufacturing enterprises to solve configuration problems and improve production efficiency for complex products. Although this study has solved the optimal solution problem of CPM, there are still some issues, such as the need to consider weighting for the requirements of different complex parts. Therefore, further research should be conducted on the multi-condition constraint weighting of the algorithm in the future.

References

[1] Khan A S. Multi-objective optimization of a cost-effective modular reconfigurable manufacturing system: An integration of product quality and vehicle routing problem. IEEE Access, 2021, 10(1): 5304-5326.

[2] Zuefle M, Krause D. Multi-Disciplinary Product Design and Modularization–Concept Introduction of the Module Harmonization Chart (MHC). Procedia CIRP, 2023, 119(1): 938-943.

[3] Mertens K G, Rennpferdt C, Greve E, Krause D, Meyer M. Reviewing the intellectual structure of product modularization: Toward a common view and future research agenda. Journal of Product Innovation Management, 2023, 40(1): 86-119.

[4] Lima M B, Kubota F I. A modular product design framework for the home appliance industry. The International Journal of Advanced Manufacturing Technology, 2022, 120(3): 2311-2330.

[5] Ameer M, Dahane M. NSGA-III-based multi-objective approach for reconfigurable manufacturing system design considering single-spindle and multi-spindle modular reconfigurable machines. The International Journal of Advanced Manufacturing Technology, 2023, 128(5-6): 2499-2524.

[6] Li Y, Ni Y, Zhang N, Liu Z. Modularization for the complex product considering the design change requirements. Research in Engineering Design, 2021, 32(4): 507-522.

[7] Wang X, Lee H, Park K, Lee G. The strategic role of R&D outsourcing practices and partners in the relationship between product modularization and new product development efficiency. Journal of Manufacturing Technology Management, 2024, 35(1): 185-202.

[8] Wang S, Li Z, He C, Liu D, Zou G Y. Core components-oriented modularisation methodology for complex products. Journal of Engineering Design, 2022, 33(10): 691-715.

[9] Forti A W, Ramos C C, Muniz Jr J. Integration of design structure matrix and modular function deployment for mass customization and product modularization: a case study on heavy vehicles. The International Journal of Advanced Manufacturing Technology, 2023, 125(3): 1987-2002.

[10] Zhao J, You X, Duan Q, Liu S. Multiple ant colony algorithm combining community relationship network. Arabian Journal for Science and Engineering, 2022, 47(8): 10531-10546.

[11] Arasteh B, Abdi M, Bouyer A. Program source code comprehension by module clustering using combination of discretized gray wolf and genetic algorithms. Advances in Engineering Software, 2022, 173: 103252.

[12] Liu C, Yang X, Wang J. Optimization of product line considering compatibility and reliability via discrete imperialist competitive algorithm. RAIRO-Operations Research, 2021, 55(6): 3773-3795.

[13] Hao J, Gao X, Liu Y, Han Z. Module division method of complex products for responding to user's requirements. Alexandria Engineering Journal, 2023, 82(1): 404-413.

[14] Arasteh B. Clustered design-model generation from a program source code using chaos-based metaheuristic algorithms. Neural Computing and Applications, 2023, 35(4): 3283-3305.

[15] Zhang Z, Lu B, Xu X, Shen X, Feng J, Brunauer G. CN-MgMP: a multi-granularity module partition approach for complex mechanical products based on complex network. Applied Intelligence, 2023, 53(14): 17679-17692.

[16] Lammers T, Guertler M, Skirde H. Can product modularization approaches help address challenges in technical project portfolio management? –Laying the foundations for a methodology transfer. International Journal of Information Systems and Project Management, 2022, 10(2): 26-42.

[17] Silva T, Santos C. Challenges of Product Modularization Methods in SMEs: Lessons Learned from a Manufacturer of Rigid Inflatable Boats. Proceedings of the Design Society, 2023, 3(1): 847-856.

[18] Persson M, Hsuan J, Hansen P K. Improving decision making in product modularization by game-based management training. Proceedings of the Design Society, 2021, 1(1): 1837-1846.

[19] Aryavalli S N G, Kumar G H. Futuristic Vigilance: Empowering Chipko Movement with Cyber-Savvy IoT to Safeguard Forests. Archives of Advanced Engineering Science, 2023, 1(8): 1-16.

[20] Monetti F M, Maffei A. Towards the definition of assembly-oriented modular product architectures: a systematic review. Research in Engineering Design, 2024, 35(2): 137-169.

Detection and Prediction of Polycystic Ovary Syndrome Using Attention-Based CNN-RNN Classification Model

Siji Jose Pulluparambil[1], Subrahmanya Bhat B[2]
Research Scholar, Institute of Computer Science and Information Science, Srinivas University, Mangalore 574146, India[1]
Assistant Professor, Department of Artificial Intelligence and Data Science, Adi Shankara Institute of Engineering and Technology, Kalady, Ernakulam ,683574, Kerala India[1]
Professor, Institute of Computer Science and Information Science, Srinivas University, Mangalore 574146, India[2]

***Abstract*—Polycystic Ovary Syndrome (PCOS) has many challenges when it comes to its diagnosis and treatment due to the diversity of presentation and potential long-term consequences for health. For this reason, sophisticated data pre-processing and classification methods are implemented to enhance the accuracy of PCOS diagnosis. A number of innovative techniques are employed in the process to enhance the accuracy and reliability of PCOS diagnosis. To identify ovarian cysts, real-time ultrasound images are pre-processed initially with the Contrast-Limited Adaptive Histogram Equalization (CLAHE) model to improve image contrast and sharpness. The ultrasound images are segmented with the K-means clustering algorithm, Particle Swarm Optimization (PSO), and a fuzzy filter, enabling precise analysis of regions of interest. An attention-based Convolutional Neural Network-Recurrent Neural Network (CNN-RNN) model is employed for classification and does so effectively to capture the temporal and spatial characteristics of the segmented data. The proposed model has a very good accuracy rate of 96% and works very well on a variety of evaluation metrics such as accuracy, precision, sensitivity, F1-score, and specificity. The results are evidence of the robustness of the model in minimizing false positives and enhancing PCOS diagnostic accuracy. Nevertheless, it is noted that bigger data sets are required to maximize the precision and generalizability of the model. The aim of subsequent research is to use Explainable AI (XAI) methods to enhance clinical decision-making and establish trust by making the model's predictions clearer and understandable for patients and clinicians. Along with enhancing PCOS detection, this comprehensive approach sets a precedent for integrating explainability into AI-based medical diagnostic devices.**

Keywords*—*Polycystic ovary syndrome; contrast limited adaptive histogram equalization; particle swarm optimization; k- means clustering algorithm; convolutional neural network; recurrent neural network

I. Introduction

Polycystic ovarian syndrome (PCOS) is a common endocrinological disorder that affects one in ten pre-menopausal reproductive women worldwide [1]. According to several studies, women with PCOS are more likely to develop ovarian and endometrial cancer, both of which can be dangerous if not recognized early [2, 3]. The identification of PCOS is a major issue since it is a frequent illness that may endanger women's physical and mental health. Ovarian malfunction and an excess of androgen are the two main signs of PCOS [4]. Many factors are believed to cause this disorder, and the factors include genetics, puberty, physiological changes, mental state, and environmental effects. Patients who have PCOS usually exhibit hirsutism, obesity, insulin resistance, irregular menstruation, and cardiovascular problems. As a result, it is crucial for the accurate diagnosis and treatment of PCOS [5, 6]. However, a growing body of scientific research indicates that PCOS can be promptly diagnosed using a well-standardized diagnostic method and that it can be treated with appropriate, symptom-focused, long-term, and dynamic therapies [7].

The most common imaging technique used in the clinical assessment of a patient with ovarian disease is ultrasound. Compared to other medical imaging techniques like computed tomography (CT) and magnetic resonance imaging (MRI), ultrasound provides a number of benefits [8, 9]. This is because ultrasonography is inexpensive, widely available, safe, and delivers results instantly. The use of this imaging approach provides a fantastic chance to create a deep learning model for automatic analysis, improving the objectivity and diagnostic accuracy of the test. Patients are required to undergo ovarian ultrasonography in order to guarantee the correctness of the first PCOS diagnosis [10, 11]. For the purpose of evaluating their metabolism, some of them could even require venous samples. It is a substantial financial burden as the average cost of the initial diagnosis and evaluation of PCOS is estimated to be $740. There has been a lot of interest in using deep learning and machine learning to identify PCOS because artificial intelligence has developed so quickly [12-14].

Currently, a manual process is used to identify the polycystic ovary shape in ultrasound images. It is based on the collective expertise of professionals in identifying the shape and features of ovarian ultrasound images [15]. After reviewing images from the same instance, the specialist's judgment may occasionally be arbitrary and unpredictable. In order to identify PCOM, radiologists have to invest a great deal of time and effort due to the various follicle sizes and their relationship with veins and tissues. Additionally, it causes artefacts and speckle noise in the images. Additionally, this manual framework for diagnosis may increase examination errors, which is inconvenient for the patient. Therefore, it is advised to suggest

clever computer-aided solutions that can provide gynecologists with decision-support tools [16]. Using clinical information and ultrasound images, a deep learning algorithm may be used to determine whether a woman has PCOS. It will also help remove barriers related to the manual review of ultrasound images and the assessment of clinical data for patients [17].

After PCOS was identified using a number of standard techniques, machine learning (ML) techniques were created and applied to clinical data [18]. The ML methods are time consuming, and yields poor detection accuracy results. On the other hand, neural networks are a well-known method for prediction [19]. Once more, some researchers used Convolutional Neural Networks (CNN) to diagnose PCOS from ultrasound images using deep learning techniques [20]. DL is a powerful method used in computer vision and image analysis. Although DL algorithms normally achieve a high degree of accuracy in classifying images. Random forest (RF) classifiers have been developed to classify PCOS and normal samples, yielding an accuracy of 72%. SVM, NB, CNN, and VGG-16 are a few machine learning and deep learning models used to analyze ovarian ultrasound images for diagnostic systems. However, several studies employed clinical data and ultrasound reports in text format rather than ultrasound images to diagnose PCOS [21-23]. PCOS is characterized by oligomenorrhea, anovulation, and biochemical hyperandrogenism. In some cases, it may occur due to the development of ovarian microcysts. Women are learning more about PCOS, which is becoming increasingly widespread. PCOS/PCOD is becoming more common in women and significantly impacts women.

According to a recent study, approximately 18 percent of women in India, especially from the East, suffer from this illness. Infertility, irregular ovulation, and preterm abortions are becoming prevalent issues for women. PCOS, a disorder that affects women of reproductive age, has been found to play a significant role in the cause of infertility. Doctors nowadays diagnose PCOS by manually counting the number of follicular cysts in the ovary, which is used to determine whether or not the condition exists. Variability, reproducibility, and efficiency issues may arise due to manual counting. The main objective of a proposed approach is listed in the following bulleted points,

- To design an effective polycystic ovary syndrome detection method using a machine learning algorithm.
- To eliminate additive noise and enhance the detection process, a dataset is pre-processed using the contrast-limited adaptive histogram equalization method.
- To present a particle swarm optimization (PSO) and K-means clustering algorithm with a fuzzy filter (FF) for the segmentation process.
- To implement an attention-based CNN-RNN deep model for increasing PCOS detection accuracy.

The proposed approach is highly accurate with low false positives and demonstrates impressive performance. The model provides a more accurate and efficient method for the diagnosis of PCOS using advanced deep learning methods, which ultimately leads to improved patient outcomes and clinical decision-making.

II. Related Work

NS. Nilofer et al. [24] developed a hybrid ML model for PCOS detection by extracting the GLCM features. The combination of an artificial neural network (ANN) and an improved fruit fly optimization approach (IFFOA) was developed. The suggested model had three main stages, namely pre-processing, segmentation, feature extraction, and detection. In the first phase, the image resizing and noise removal are executed using a medial filtering approach. Then, the enhanced K-means clustering model was utilized to perform the segmentation process. After attaining a segmented follicle part, the features from the particular portions are extracted using the Grey-Level Co-occurrence Matrix (GLCM). An enhanced optimization model was introduced in the detection model for the optimal two key parameters. The dataset used here was the US image dataset, and the performance measures of precision, recall, F measure, and accuracy are analyzed.

C. Gopalakrishnan et al. [25] presented a detection method of PCOS from the ultra sound image of the ovary. The ovary's size, number of follicles, and location may all be learned by ultrasound imaging. Due to the different follicle sizes and the complex relationship between tissues and blood vessels, diagnosing PCOS in real time can be challenging for radiologists. This frequently leads to incorrect diagnosis. Initially, the pre-processing stage was performed, and in this stage, RGB to Gray conversion, ROI extraction, and speckle noise reduction were executed. The ovary image was binarized for segmentation using the modified Otsu approach. The contour initialization concerns were resolved since the binary mask had foreground and background areas organized properly to segment objects. Additionally, based on the distributed grey level, a thresholding strategy was added to remove the objects from the ovary image. Finally, an improved outcome in terms of accuracy had been achieved. As a result, the modified Otsu method proved effective for follicle extraction.

Recently, medical professionals have been able to identify and categorize illnesses using image recognition techniques. Traditional methods for detecting PCOS have several drawbacks, therefore, Dongyun He et al. [26] developed a probabilistic method. To ensure accurate cue recognition, the training images were first split into several grids of identical sizes. Furthermore, each grid in the supplied image had a quality score that roughly corresponded to its grayscale and texture properties. As a result, one may consider each image to be a scoring matrix. The feature vectors might be provided using the statistically based model while taking into account the score matrix. The probabilistic model was used to train the identified feature vectors, and the learned feature vectors were subsequently transformed into an SVM kernel to identify PCOS.

To identify PCOS from ultrasound images of the ovary, C. Gopalakrishnan et al. [27] developed a model based on scale-invariant feature transform (SIFT) descriptors. Initially, the Canny edge detection method was utilized to enhance image

quality and delineate the margins of follicles in the ultrasound image. This approach involved pre-processing, gradient computation, non-maximum suppression, and thresholding as integral steps of the Canny edge detection process. SIFT descriptors are used to identify the feature descriptors for diagnosing the condition. Then, a support vector machine (SVM) was used to ultimately perform data training and classification. Better accuracy, mean squared error, and normalized absolute error has been achieved for PCOS identification.

In order to identify PCOS using ultrasound images, Untari N. Wisesty et al. [28] suggested a modified back propagation method. Typically, stereology calculations or feature extraction and classification were the key foundations for PCO follicle identification. The extraction and categorization of features served as the foundation for this PCOS detection. The Gabor wavelet was taken into consideration for the feature extractor, while the modified back propagation model was employed as a classifier. Levenberg-Marquardt optimization (LMO) and Conjugate Gradient-Fletcher Reeves (CGFR) were the modified backpropagation algorithms. LMO was used to achieve the highest level of accuracy by considering 33 neurons and 16 vector characteristics.

A PSO model to partition follicles and identify PCOS was reported by E. Setiawati et al. [29]. Here, the follicles were segmented using a novel clustering model created with PSO and a modified non-parametric fitness function. The primary goal of the fitness function would be to detect faults based on pixel values and improve the likeness to human vision. Then, normalized mean square error (NMSE) and the mean of the modified non-parametric fitness function and structural similarity index (MSSIM) approaches were used to create convergent and compact clusters. The PSO fitness function also led to more convergent solutions. The performance of the examined PSO also impacted the extracted follicular size and contrast enhancement.

The chan-vase model and split-Bregman optimization were introduced by H. Prasanna Kumar and S. Srinivasam [30] for quick segmentation of PCOS. A thorough understanding of the size and quantity of follicles might be gained through PCOS diagnosis using ultrasound images. Using an increased active contour without an edge model, the tiny follicles could be identified. In addition, by using the split-Bregman optimization model, the segmented image's accuracy and computation time are enhanced. Results demonstrated that the split-Bregman optimization model produced superior outcomes with less computing time and iteration.

Onyema et al. [31] applied the AI-based Granger panel model approach to provide an empirical analysis of apnea syndrome. The MIT-BIH polysomnographic database (SLPDB) was the source of the data. MATLAB software was utilized for the implementation, and the panel consisted of eighteen patients. The findings indicate a substantial correlation between ECG-blood pressure (BP), ECG-EEG, and EEG-blood pressure (BP) for the eighteen sleep apnea patients.

Tiwari et al. [32] developed a model that diagnoses based on a clinical dataset Kottarathil provided and made available through its Kaggle repository. A variety of machine learning techniques for patient screening without the need for intrusive diagnostics are assessed using non-invasive screening metrics. The experiments demonstrate that the Random Forest (RF) approach outperforms the other well-known machine learning algorithms with an accuracy of 93.25%. Moreover, the model possesses high complexity.

Alamoudi et al. [33] suggested a data set containing an ultrasound image of the ovary and clinical information about a patient classified as either PCOS or non-PCOS is presented. Then, using the Inception model, a deep learning model was built to diagnose the PCOM based on the ultrasound image and achieved 84.81% accuracy. Then, in order to determine whether or not the patient has PCOS, a fusion model that combines clinical data with the ultrasound picture was suggested. By combining clinical features with mobile net architecture to extract image data, the most advanced model to date has achieved 82.46% accuracy.

Lv et al. [34] suggested an automated deep learning method that investigates the possibility of scleral alterations in PCOS identification for auxiliary PCOS detection. After utilizing an enhanced U-Net to separate scleral photos from full-eye images, the method was run on the dataset. From there, deep features were extracted from the scleral images using a Resnet model. In order to accomplish categorization, a multi-instance model was created. A variety of performance metrics, including AUC, F1-score, recall, precision, and accuracy of classification, are used to evaluate the effectiveness of the method. The results demonstrate the high potential of deep learning in PCOS diagnosis, achieving an average AUC of 0.979 and a classification accuracy of 0.929. The comparison with existing methods is shown in Table I.

Gaps Identified in the Literature Review and Contributions of the Work

- Data Scalability and Flexibility: Some of the models mentioned, such as the probabilistic model with SVM, the Canny edge detection method, and the adaptive k-means clustering with GLCM feature extraction and ANN network, struggle with different types of image datasets and large databases.
- Complexity and Parameter Tuning: Support Vector Machine (SVM), Random Forest (RF), and Decision Tree models require high-fidelity parameter tuning and can be computationally intensive, leading to inefficiencies in real- world applications.
- Temporal analysis and feature extraction: Spatial feature extraction is the prime focus of traditional models like active contour with altered Otsu threshold value, Particle Swarm Optimization with a new revised non-parametric fitness function, and LMO (Local Min-Orthogonal) CGFR (Centered Gaussian Fit Regression) algorithms. They cannot capture the temporal dependencies of the data sufficiently.

The CNN-RNN model's strong architecture makes it better placed to deal with large and diverse datasets.

- With the addition of attention mechanisms, the CNN-RNN model enhances model performance and

efficiency while reducing the need for manual parameter tuning.

- The CNN-RNN model excels in this aspect by combining Convolutional Neural Networks for spatial feature extraction and Recurrent Neural Networks for temporal sequence analysis, providing a more comprehensive approach to diagnosing PCOS.
- By combining recurrent neural networks for analysis of temporal sequence and convolutional neural networks for spatial feature learning, the CNN-RNN model excels at this task and provides a more comprehensive approach to PCOS diagnosis.

TABLE I. COMPARISON WITH STATE-OF-THE-ART TECHNIQUES

References	Method	Advantages	Disadvantages
[24]	Adaptive k-means clustering, GLCM feature extraction, and ANN network model	The system is used for alternative medical data, and a number of criteria are used to evaluate its efficacy.	Various algorithms can be employed to optimize the parameters of an artificial neural network.
[25]	Active Contour with modified OTSU threshold value	High accuracy	A large database cannot be applicable
[26]	Probabilistic model and SVM with kernel function	Image quality diagnosis may assess early alterations in endometrial thickness and blood flow and perform early illness diagnosis and therapy.	Accuracy can be improved.
[27]	The Canny Edge detection method	The suggested methods are more suited to extracting and identifying follicles from ovarian images.	Different kinds of image datasets cannot be used.
[28]	LMO and CGFR algorithm	It has high accuracy	The conjugate gradient parameter is not considered
[29]	PSO with a new modified non-parametric fitness function.	The retrieved follicular size may be made to resemble the real follicular size by using contrast enhancement.	The follicles cannot be identified automatically.
[30]	Improved active contour without edge method.	The performance is high	High cost
[31]	AI-based Granger Panel model approach	Valuable statistical source for the analysis of dynamic behaviours	Give less performance in prediction
[32]	RF	Attain accuracy of 93.25%	Highly complex in computation
[44]	Random forest	High accuracy and robustness; handles missing data well.	Can require a great deal of processing power and overfit noisy data.
[45]	Decision tree	It needs minimal data preparation and is easy to analyze and represent.	Vulnerable to overfitting and subject to instability when data is slightly different
[46]	Convolutional neural network (CNN)	Extracts and categorizes features automatically.	High computing power and large data sets are required,
[47]	Transfer learning (e.g., VGG16) [48]	Uses pre-trained models to improve performance.	Require huge processing power.

III. METHODOLOGY

PCOS is caused by a hormonal imbalance that can lead to various illnesses and occurs in one in ten women of reproductive age from 18 to 44. In the proposed approach, machine learning-based techniques are proposed to detect PCOS disorder efficiently. The main novelty of the proposed work lies in the integration of several advanced techniques to address the challenge of diagnosing PCOS using ultrasound images: CLAHE takes it a step further by limiting the amplification of noise in relatively homogeneous regions. This helps in enhancing the contrast of ultrasound images, which is crucial for identifying subtle details like ovarian cysts. This combination of K-means clustering and PSO with fuzzy filtering allows for more accurate segmentation of ultrasound images. K-means clustering partitions the image into clusters based on pixel intensity, while PSO optimizes the clustering process by finding the optimal centroids. Fuzzy filtering further refines the segmentation by considering the uncertainty in pixel classification, improving the regions of interest identification.

CNN-RNN architecture is designed to effectively capture both spatial and temporal features within the segmented ultrasound images. The attention mechanism allows the model to focus on the most relevant regions, while the CNN component extracts spatial features, and the RNN component processes temporal sequences. This holistic approach enables more accurate classification of ultrasound images, aiding in the diagnosis of PCOS. By combining these techniques, the proposed approach offers a comprehensive solution for PCOS diagnosis. The proposed approach comprises three stages for PCOS detection they are;

- Pre-processing
- Segmentation
- Classification.

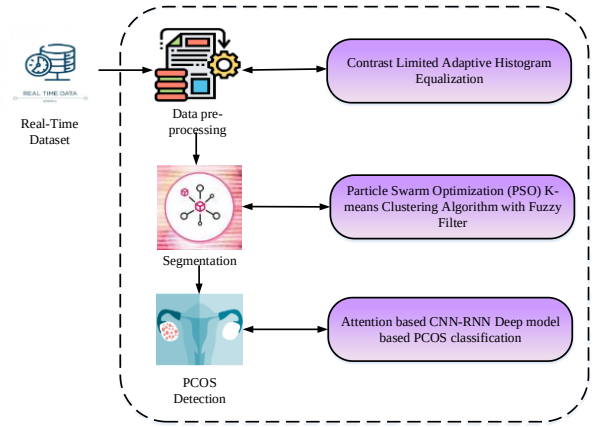


Fig. 1. Workflow for proposed approach.

Fig. 1 represents the step-by-step process involved in the proposed approach. In the first stage of the proposed approach, the real-time dataset is pre-processed to enhance the performance of PCOS detection. After that, the segmentation process is done for pre-processed data using particle swarm optimization (PSO) K-means clustering algorithm with FF. Then, the detection process was done by attention-based CNN-RNN deep model. The proposed classification model applies to many technical problems with reduced complexity. The proposed model provides the best optimal outcome and improves classification accuracy effectively. The step-by-step process of PCOS detection using machine learning techniques will be discussed briefly in subsequent sections.

A. Pre-processing

The initial stage of the proposed work is pre-processing, which is aimed at enhancing the quality and clarity of ultrasound images specifically for the purpose of identifying ovarian cysts, which are a hallmark of PCOS diagnosis. CLAHE is applied to improve image contrast and detail. This enhancement is crucial for subsequent steps in the analysis pipeline, such as segmentation and classification, as it enables more accurate regions of interest identification and characterization within the ultrasound images. Therefore, the pre-processing is focused on preparing the ultrasound images for further analysis and diagnosis of PCOS. The proposed approach pre-processes the real-time dataset using the CLAHE method [35]. The use of CLAHE for pre-processing is attributed to its effectiveness in enhancing image contrast and improving local details, particularly in medical imaging applications like detecting polycystic ovary syndrome. CLAHE is specifically designed to address the limitations of traditional histogram equalization by limiting contrast amplification in regions with high contrast variations, thereby avoiding over-enhancement artifacts.

Additionally, CLAHE's adaptive nature allows it to adjust parameters locally based on the characteristics of different image regions, which is crucial for maintaining diagnostic information integrity in medical images. This flexibility guarantees the preservation and highlighting of pertinent characteristics linked to polycystic ovary syndrome, improving the overall efficacy of later analysis algorithms. Moreover, CLAHE is a better option for pre-processing tasks due to its ease of integration into current pipelines and low computational overhead brought about by its simplicity and efficiency. Image quality consistency between datasets is guaranteed by using CLAHE to standardize pre-processing. Input image in the CLAHE model is classified into non-overlapping contextual regions known as blocks/sub-images/tiles. In the CLAHE model, two main metrics are Clip Limit-CL and Block Size-BS. Using these metrics in the CLAHE model maintains and enhances image quality. When the CL is higher then, the histogram becomes flatter, and the image's brightness is increased due to the low intensity of the input image. In addition, if BS is higher, the image contrast is increased, and the dynamic range becomes higher. An optimal image quality is generated using the image entropy when the two metrics are determined at the point with the huge entropy curvature. The major steps of the CLAHE model are defined below:

Step 1: Classify the intensity of an original input image into non-overlapping contextual regions. In which, $m \times n$ is considered the total number of image blocks and 8×8 is considered the optimal value for preserving the chromatic data in the image.

Step 2: Based on the values of the grey level present in the image of the array, evaluate the contrast-limited histogram for each contextual zone.

Step 3: By using the CL value, the limited contrast histogram for the contextual region is calculated, and the mathematical equation is given below:

$$N(Avg) = \left[\frac{N(U)_r \times N(V)_r}{N(GL)} \right] \quad (1)$$

Here the average number of pixels is represented by $N(Avg)$, the number of gray levels in the contextual region is represented as $N(GL)$, $N(U)_r$ *and* $N(V)_r$ represents the number of pixels in the U dimension, and V dimension for the contextual region. The actual CL is calculated by the following equation;

$$N(CL) = Nor(CL) \times N(Avg) \quad (2)$$

Here $N(CL)$ represents the actual CL, $Nor(CL)$ represents the normalized CL. The total number of clipped pixels is determined as $Nor\sum(CL)$. Finally, the average of the remaining pixels to be distributed into each gray level is defined in the following equation;

$$N(Avg)Nor(CL) = \left[\frac{Nor\sum(CL)}{N(GL)} \right] \quad (3)$$

By using the following condition, the histogram clipping rule is calculated;

If $His_{reg}(x) > N(CL)$ then;

$$His_{reg\_clip}(x) > N(CL)$$

Else if $(His_{reg}(x) + N(Avg)N(CL) > N(CL)$ then

$$His_{reg\_clip}(x) > N(CL)$$

Else $(His_{reg\_clip}(x) = His_{reg\_clip}(x) + N(CL)$

Here the original histogram is represented as $His_{reg}(x)$, and the clipped histogram for each region at x^{th} gray level is represented as $His_{reg\_clip}(x)$.

Step 4: Reallocate the other pixel until all have been allocated. The step for reallocating the pixel is represented in the following equation;

$$S(PI) = \left[\frac{N(GL)}{N(\mathrm{Re}main)} \right] \quad (4)$$

Here $S(PI)$ represents the step of a positive integer of at least one, $N(GL)$ represents the number of gray levels, and $N(\mathrm{Re}main)$ represents the remaining number of clipped pixels. The program begins its search process from the lowest to the highest order of gray level using the above step. The program will allocate one pixel to the gray level if the number of pixels in the gray level is less than $N(CL)$. If the distribution of the pixels is not even performed when the search is complete, the program computes a new step using Eq. (4) and starts a new search round until all of the remaining pixels are dispersed.

Step 5: Each region's intensity values are improved using the Rayleigh transform model. The clipped histogram is distorted to cumulative probability $CP_{inp}(x)$; it is used to generate a transfer function. The PCOS image shows as another original when the Rayleigh distribution is applied. The mathematical equation of using the Rayleigh forward transform is represented in the following equation;

$$RFT(x) = \left[PV_{LB} + \sqrt{2\delta^2 In\left(\frac{1}{1 - CP_{inp}(x)} \right)} \right] \quad (5)$$

Here RFT represents the Rayleigh forward transform, PV_{LB} represents the pixel value of the lower bound, and δ represents the scaling parameter of the Rayleigh distribution. The output of probability density is defined below:

$$CP(RFT(x)) = \left[\frac{RFT(x) - PV_{LB}}{\delta^2} \bullet Exp\left(-\frac{(RFT(x) - PV_{LB})^2}{2\delta^2} \right) \right] \quad for\ RFT(x) \geq PV_{LB} \quad (6)$$

Where a higher δ value results in more effective contrast improvement in PCOS image to enhance the rate of saturate value and reduce the level of noise.

Step 6: Limiting the effect of a sudden change. Linear contrast stretch is used to resize the transfer function output dynamically. The linear contrast stretch is represented in the following equation;

$$RFT(x) = \left[\frac{TF(x) - TF_{Min}}{TF_{Max} - TF_{Min}} \right] \quad (7)$$

Here $TF(x)$ represents the transfer function.

Step 7: In order to prevent border artefacts, the new grey level task of pixels inside a sub-matrix contextual region is computed via a bi-linear interpolation between four mappings.

By using CLAHE, the dark area in the images is clearer and more prominent for higher contrast. As a result, the CLAHE model is suggested to improve PCOS detection performance by eliminating boundary artefacts.

B. PSO-Based K-means Model with Fuzzy Filter for PCOS Segmentation

In the proposed approach, PCOS segmentation is considered an essential stage to extract significant objects lying in images and segment the images into nearby semantic regions. The segmentation process of PCOS images is commonly difficult because the images are complicated, diverse and differ from person to person. Various methods are proposed for segmenting the follicles in ultrasound images. In the proposed approach, the segmentation process is performed by a particle swarm optimization (PSO) based K-means clustering algorithm with an FF to enhance the performance [36]. In the proposed approach, FF is used in image processing to enhance the performance of classifiable filters. The main objective of using an FF in the proposed approach is to reduce the unwanted noises in contaminated images when there are many uncertainties. In the fuzzy filtering approach, a pixel is denoted by the membership function and a set of fuzzy rules that consider adjacent information in a limited area or other data to remove the noise with blurry edges of the PCOS image. Recently, the K-means method has become one of the most prominent models in medical technology. The k-means clustering method is considered the most significant algorithm used in centre-pivot clustering techniques in clustering. The initialization of cluster centres and other factors in the K-means clustering method begins with allocating Euclidean distance criteria to one cluster with the shortest distance among the cluster centres. PSO is a population-based search algorithm; each individual in PSO is considered a particle and "flown" via hyper-dimensional space. Individuals' social psychological behaviour causes particles to shift their position in the search space to replicate the success of others. The neighbours' knowledge and experience impact the changes in the particles within the same swarm. The search behaviour of a particle is influenced by the behaviour of other particles in the equivalent swarm.

K-means Clustering: K-means clustering is a form of unsupervised learning used when no groupings or categories exist in the data. The main objective of a K-Learning algorithm is to detect groups in data, where the variable K indicates the number of groups. In the proposed approach, every PCOS

image waits to be segmented while the data points set is represented in the following equation;

$$X = \left[x_1, x_2, \ldots, x_n \right] \tag{8}$$

Here n represents the dimension vector. Using the K-means segmentation process in the proposed approach, PCOS images are divided into K-cluster. The normal method detects the subset $S = \{Cl_1, Cl_2, \ldots, Cl_k\}$ in a set x to reduce the target function $TF = \sum_{i=1}^{k} \sum x_j \in s_j D_{ij}(x_j, Cl_i)$. In which $D_{ij}(x_j, Cl_i)$ represents the Euclidean distance from the data point x_j to a clustering centre Cl_i. The target function is represented in the following equation;

$$TF = \left[\sum_{i-1}^{k} \sum_{x_j \in Cl_i} \left\| x_j - Cl_i \right\|^2 \right] \tag{9}$$

The target function TF is closer to clustering with the clustering effect. As a result, to get the optimum clustering effect, the target function value and all the clustering centres should be set as zero, and the mathematical equation is given below:

$$Cl_i = \left[\frac{1}{N_i} \sum_{j=1}^{n_i} x_j \quad (i = 1,2,\ldots,k) \right] \tag{10}$$

Here N_i represents the number of data points in the cluster i. The clustering centre of K-Means is constantly updated by reducing several iterations of the target function. The similarity of data points in the same clustering rapidly enhances as the clustering centre is updated, whereas the similarity of data points in separate clustering gradually decreases. The clustering process is over when no new information exists in the clustering centre.

PSO: In the proposed approach, PSO-based K-means clustering is used for segmentation to detect the approximate solutions for PCOS detection. PSO is an artificial intelligence (AI) based technique used to find approximate solutions to numerical maximizing and minimization problems that are exceedingly difficult or impossible to solve. In PSO, the best solution for each problem may be computed into a single particle with no quality or volume in N-dimensional space. The step-by-step process involved in the PSO-based K-means clustering algorithm is as follows;

Let's consider the particle P_i present in the dimension space D; then the position is represented in the following equation;

$$P_i = \left[P_{i1}, P_{i2}, \ldots, P_{iD} \right] \tag{11}$$

Then, the speed is represented in the following equation;

$$SD_i = \left[SD_{i1}, SD_{i2}, \ldots, SD_{iD} \right] \tag{12}$$

During the update process, the best position of a particle is detected, and the iteration is represented in the following equation;

$$P_i = \left[P_{i1}, P_{i2}, \ldots, P_{iD} \right] \tag{13}$$

Then, its best position is;

$$P_B = \left[P_{B1}, P_{B2}, \ldots, P_{BD} \right] \tag{14}$$

Then, the speed calculation of each particle is represented in the following equation;

$$SD_{ij}(T+1) = \left[WSD_{ij}(T) + LF_1 R_1 \left(P_{ij} - X_{ij}(T) \right) + LF_2 R_2 \left(P_{Bi} - X_{ij}(T) \right) \right] \tag{15}$$

The calculation of particle position is represented in the following equation;

$$X_{ij}(T+1) = \left[X_{ij}(T) + SD_{ij}(T+1) \right] \quad j = 1,2,\ldots,D \tag{16}$$

Here, the inertia coefficient is represented as W, learning factors are denoted as $LF_1 R_1$ and $LF_2 R_2$ in that order, denoting the capacity for self-learning and learning from an excellent group of particles. In a PSOK-based algorithm, the linear regression model is used to assist the inertia coefficient and enhance the convergence and optimization of the quality speed rate.

PSO-based K-means Clustering with FF for PCOS Segmentation: In the proposed approach, PSO-based K-means clustering (PSOK) is proposed to enhance the segmentation process of PCOS detection. The cluster centres in PSOK are recognized as the particle positions, and the weakest particle is removed by searching for the optimal solution using the PSO algorithm to enhance the computation. The proposed approach proposes the K-means clustering algorithm to update the particle positions. In PSOK, n particles are initialized, with their positions and velocities updated as needed, and the fitness values evaluated and arranged in decreasing order in a list. The iteration process continues until the maximum number of iterations is met or the minimal error condition is reached.

Step 1: Choose m particles known as the initial population number, and feed the particles into the initial swarm called IS is represented in the following equation;

$$IS_1 = \left[P(1), P(2), \ldots, P(m) \right] \tag{17}$$

Then, initialize the position x_{id} for the swarm S by utilizing the K-means clustering algorithm.

Step 2: Arbitrarily begins the velocities v_{id}

Step 3: Fitness evaluation for each particle by using $\left(x_{id}(T) \right)$.

Step 4: The exploration stage is given below:

$$p_{id}(T+1)=\begin{cases} p_{id}(T) & FitVal(p_{id}(T))>FitVal(x_{id}(T)) \\ x_{id}(T) & FitVal(p_{id}(T))\le FitVal(x_{id}(T)) \end{cases} \quad (18)$$

Step 5: Determine the global best $p_{GB}(T+1)$ for the particle position by the best fitness value computed in the swarm.

Step 6: Based on the K-means clustering algorithm, the position of every new particle is optimized in the $S(T+1)$ new swarm.

Step 7: Based on the below equation, the velocity vector $(v_{id}(T+1))$ for each particle is varied.

$$\begin{pmatrix} v_{id}(T)=wv_{id}(T-1)+LF_1Rand()[p_{id}(T-1)-x_{ii}(T-1)] \\ +LF_2Rand()[p_{id}(T-1)-x_i(T-1)x_{id}(T)] \\ =x_{id}(T-1)+v_{id}(T)+v_{id}(T-1) \end{pmatrix} \quad (19)$$

Step 8: In which, i^{th} particle in a D dimensional space is represented by $(x_{id}(T))$ at the time step T velocity v_{id} of $p_{id}(T)$. Here LF_1 and LF_2 represents the learning factor, w represents the inertia weight, and $Rand()$ represents the random function.

Step 9: Then update each particle in $S(T+1)$.

Step 10: Stop the process until the maximum number of iterations reached of the minimum error condition is satisfied, or else repeat the process.

Therefore, the proposed approach uses a PSOK clustering-based fuzzy filter model for partitioning the ultrasonic image of PCOS into several segments. It uses an appropriate suppression factor for perfect segmentation of PCOS to enhance detection performance.

C. Attention-Based CNN-RNN Deep Model for PCOS Classification

In the classification stage, the segmented PCOS features are fed into the detection model for PCOS detection. An attention-based CNN-RNN deep model is proposed for PCOS detection. The proposed approach is a combination of a CNN and an RNN [37]. The attention-based CNN-RNN deep model is essential for classifying PCOS because it can capture complex temporal dynamics and spatial patterns in medical image sequences. The model gets a thorough knowledge of PCOS-related features by incorporating RNN for modeling temporal dependencies and CNN for extracting spatial information. The inclusion of attention mechanisms concentrates on pertinent areas or time steps in the data, which improves the interpretability of the model even more. By ensuring resilience to the individual variations in PCOS symptoms, this hybrid architecture enhances generalization performance. Furthermore, the interpretability of the model helps physicians comprehend the logic underlying its predictions, enabling well-informed decision-making. All things considered, the Attention-based CNN-RNN deep model makes a substantial contribution to medical image analysis and diagnosis by providing a strong and adaptable framework for precise, comprehensible, and clinically useful PCOS classification.

Using a combination of cutting-edge NN components, the attention-based CNN-RNN deep model is a sophisticated architecture designed for the complex problem of PCOS categorization. Its purpose is to extract valuable information from medical image sequences. CNNs, well-known for their ability to extract spatial characteristics from images, serve as the model's foundation. CNNs are particularly good at detecting minute patterns and structures in ultrasound scans and other medical images, which is important when it comes to differentiating between ovaries that are damaged and those that are not. CNNs are enhanced by RNNs, who are skilled in simulating the temporal dependencies in sequential data. This skill is essential for documenting dynamic changes in ovarian morphology over time, as these alterations may be markers for diagnosing PCOS. The model can identify temporal patterns and fluctuations that could indicate the onset or progression of PCOS due to RNNs.

The capacity of the model to rank pertinent areas or time steps in the input data is improved by the addition of attention mechanisms. Attention processes enable the model in the setting of PCOS classification to concentrate on particular areas of interest within medical images or sequences, removing unnecessary information and highlighting important diagnostic signals. The model obtains a comprehensive grasp of PCOS-related factors by merging temporal dynamics recorded by RNNs with spatial data recovered by CNNs. This combination of temporal and spatial data improves diagnostic accuracy by allowing the model to identify intricate patterns and changes in both domains.

Attention techniques in the hybrid CNN-RNN architecture give the model robustness and allow it to generalize well to various PCOS symptoms seen in clinical practice. This resilience guarantees the model's strong performance in various patient demographics and imaging situations, improving its clinical usefulness and dependability. Beyond its potential for classification, the attention-based CNN-RNN deep model is interpretable, giving physicians an understanding of the machine's decision-making process. The clinical applicability and acceptance of the model are eventually increased by this openness, which also promotes trust and cooperative decision-making between clinicians and AI systems. Using the synergy of CNNs, RNNs, and attention mechanisms to extract, integrate, and interpret complex information from medical image sequences, the Attention-based CNN-RNN deep model advances state-of-the-art PCOS diagnosis and improves patient care.

Here, CNN consists of three layers: convolution, fully connected, and pooling layers. The convolution layer is first used in CNN to extract the required features from input PCOS images. The edges and corner information about the images are extracted using a feature map. The second fully connected layer is used for classification, which uses the convolution layer output to detect the image class effectively. The pooling layer reduces the convolved feature map size to minimize computational costs. RNN is a neural network (NN) type in

which the previous output is considered the current input. The input layer contains the initial data for the NN, the hidden layer acts as an intermediate layer among input and output layers, and the output layer generates the final results. In the proposed approach, the CNN model contains seven layers; the initial two are convolution layers with 64x3x3 kernels. The local features of a PCOS image are extracted using 64x1x1 kernels of a locally connected layer. Each RNN unit in the sequence modelling stage has a probability of 0.5 and a dropout of 512 hidden units, followed by a P-way fully connected layer and a softmax classifier.

The integration of CNN and RNN with attention processes may elevate the possibility of overfitting, particularly in scenarios involving small dataset sizes. In order to mitigate this risk, the suggested model incorporates dropout layers. P represents the number of polycystic ovaries to be predicted. The final class of PCOS image is predicted using the average pooling of a softmax output. RNN consists of encoding contextual information and feedback loops of a temporal sequence. Let's consider the input sequence $\{S_1, S_2, \ldots, S_T\}$ segmented from the input PCOS, $\{H_T\}$ represents the hidden state and $\{O_T\}$ represents the output state. Therefore, the mathematical equation is defined below:

$$\{H_T = Height(W_{InH}F_T + W_{HH}H_{T-1} + B_H)\} \quad (20)$$

$$\{O_T = W_{HO}H_T + B_0\} \quad (21)$$

The weight matrices of the three layers are represented as $\{W_{InH}, W_{HH}, W_{HO}\}$. In the proposed approach, long short-term memory (LSTM) is used to improve the standard of RNN from the vanishing of gradient issue. In LSTM, each unit contains an input gate, forget gate, cell gate, and output gate, and their equations are given below;

$$\{In_T = \delta(W_{In}[H_{T-1}, F_T] + B_{In})\} \quad (22)$$

$$\{F_T = \delta(W_F[H_{T-1}, F_T] + B_F)\} \quad (23)$$

$$\{O_T = \delta(W_O[H_{T-1}, F_T] + B_O)\} \quad (24)$$

$$\{\hat{C}_T = TanH(W_C[H_{T-1}, F_T] + B_C)\} \quad (25)$$

$$\{C_T = F_T \bullet C_{T-1} + In_T \bullet \hat{C}_T\} \quad (26)$$

$$\{H_{T-1} = O_T \bullet TanH(C_T)\} \quad (27)$$

Here $In, F, O and C$ represents the input, forget, output, and cell gate activation and δ represents the logistic sigmoid function. In the proposed approach, an attention layer is proposed to improve the hybrid CNN-RNN performance is expressed in the following equation;

$$A_T = TanH(W_H H_T) \quad (28)$$

$$\alpha_T = SM(W^T M_T) \quad (29)$$

$$R = \sum_{T=1}^{T} \alpha_T H_T \quad (30)$$

Here H_T represents the output of T^{th} hidden layer in the RNN module, α_T represents the T^{th} attention weight, W_H and W^T represents the weighted matrices, and the attention module is denoted as R. The output R is followed by a P-way fully connected layer and softmax classifier. The loss function of the detection model is represented by the following equation;

$$LF = [\alpha.L_{Att} + \beta.L_{Att} + \lambda.\|W\|]^2 \quad (31)$$

Here LF represents the attention loss, L_{Att} represents the target replication loss, and W represents the regularization term. α, β, λ represents the three weight parameters.

$$L_{Att} = \left[\frac{1}{TS}.l(G_1(x), y)\right] \quad (32)$$

$$l[P_1(x), y] = \left[-\sum_{i=1}^{P} 1_i(y) Log\, P_1(x)^i\right] \quad (33)$$

$$P_1(x) = [F_S(F_a(F_H(x_1), F_H(x_2), \ldots, F_H(x_T)))] \quad (34)$$

Here x represents the PCOS detected, T represents the number of time steps of RNN, P represents the number of polycystic ovaries to detect, $P_1(x)^i$ represents the i^{th} dimension of $P_1(x)$ and $P_i()$ is the indicator function. F_S, F_h and F_a represents the hybrid CNN-RNN framework, attention module, and the last softmax layer, respectively.

$$L_{Tar} = \left[\frac{1}{TS}\sum_{T=1}^{TS} l(P_2(x_T), y)\right] \quad (35)$$

$$l[P_2(x_T), y] = \left[-\sum_{i=1}^{P} 1_i(y) Log\, P_2(x_T)^i\right] \quad (36)$$

$$P_2(x_T) = [F_S(F_H(X_T))] \quad (37)$$

Here x_T represents the T^{th} sub-segment of x and $P_1(x)^i$ represents the i^{th} dimension of $P_1(x)$. F_S and F_h represents the softmax layer.

IV. Result and Discussion

The proposed model is trained using a real-time dataset collected from clinical sources and meticulously augmented to enrich its diversity and enhance training efficacy. The dataset size amounts to 1.32MB, comprising images categorized into two distinct classes: "affected" and "not affected." The dataset's sources encompass a broad spectrum of clinical scenarios, ensuring a representative sample that captures the variability inherent in polycystic ovary syndrome cases. The images were collected from the Kerala hospital containing nearly 20 women's scan reports. The augmentation process involves rotation, scaling, and flipping to expand the dataset's size and improve model generalization. This comprehensive approach to dataset preparation underscores the robustness and reliability of the proposed model in real-world clinical applications. In the result and discussion section, all machine learning methods are compared with a proposed model using a graph. The proposed approach is implemented in the PYTHON platform and assessed using accuracy, precision, sensitivity, and F1-score. The experimental outcomes are evaluated and compared with the earlier methods like SVM, logistic regression (LR), naive Bayes (NB), random forest (RF), and classification and regression tree (CART) [38]. The detailed description is given below;

- SVM is a supervised learning algorithm. It is based on SVM and can handle multiclass. The SVM classifier model is placed near a classifier's margin to enhance the performance.
- The LR model is an ML algorithm that solves categorization issues. Binary values are calculated using classifications and determined by a group of different values. The detection probability is used, and the value remains between zero and one.

A. Performance Analysis

Accuracy: The performance accuracy for a proposed approach is considered a significant metric for computing the efficiency and the improved rate of any proposed classification method compared with existing methods. The following equation calculates the accuracy performance of a proposed approach;

$$Accuracy = \left[\frac{TruePos + TrueNeg}{TruePos + TrueNeg + FlasePos + FlaseNeg} \right] \quad (38)$$

Where Acc represents the accuracy performance, $TruePos$ represents the true positive, $TrueNeg$ represents the true negative, $FalsePos$ represents the false positive and $FalseNeg$ represents the false negative value. The proposed approach considers accuracy a significant task in defining classification performance to detect PCOS.

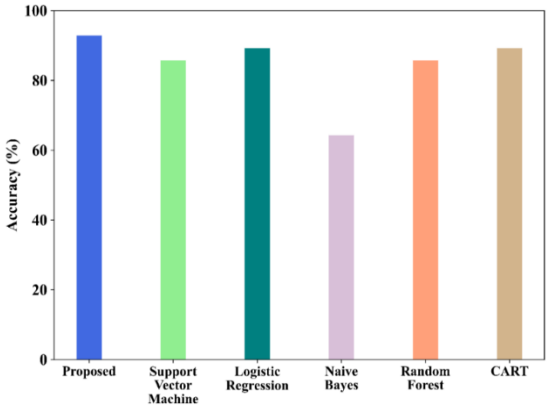


Fig. 2. Accuracy performance.

Fig. 2 illustrates the accuracy analysis. It shows that the proposed approach provides an effective accuracy rate in detecting PCO or non-PCO follicle classes. The proposed approach attains 96% accuracy in detecting PCOS, which is more reliable than existing approaches like support vector machine (85%), logistic regression (89%), naive Bayes (64%), random forest (85%), and CART (89%). Evaluating other methods proves that the defined model provides better results in classifying the PCO or non-PCO follicle class than existing methods.

Precision: Precision performance is considered the most important metric for computing the classification results in the proposed approach. The following equation calculates the performance of precision;

$$\Pr ecision = \left[\frac{TruePos}{TruePos + FalsePos} \right] \quad (39)$$

Here $\Pr e$ represents the precision performance, $TruePos$ represents the true positive, and $FalsePos$ represents the false positive values. Precision performance is used to establish the images that are exactly classified to authenticate the overall accuracy of the detection system of PCOS. The performance result of precision in proposed and existing methods is represented in Fig. 3.

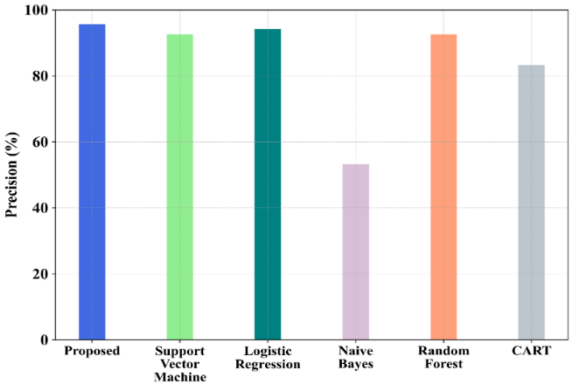


Fig. 3. Precision performance.

Fig. 3 illustrates the precision performance for a proposed model. The proposed approach provides an effective precision rate in detecting PCO or non-PCO follicle class. The proposed approach attains 96% precision in detecting, which is more reliable than existing approaches like SVM (92%), LR (94%), NB (53%), RF (92%), and CART (83%). The proposed method offers exceptional results in identifying the PCO or non-PCO follicle class, as demonstrated by the evaluation of alternative approaches.

Sensitivity: Sensitivity performance is another significant metric for detecting the overall sensitivity of the system model. The following equation calculates the performance of sensitivity;

$$Sensitivity = \left[\frac{TruePos}{TruePos + FalseNeg} \right] \quad (40)$$

Here Rec denotes the performance of sensitivity, $TruePos$ denotes the true positive, and $FalseNeg$ denotes the false negative value. $FalseNeg$ denotes the actual value is true, but the obtained result is false. By using sensitivity performance, classification errors are detected significantly. The sensitivity performance analysis for proposed and existing approaches is represented in the following Fig. 4.

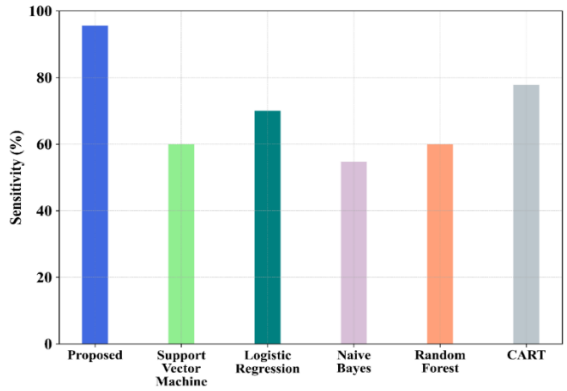


Fig. 4. Sensitivity performance.

Fig. 4 illustrates the sensitivity performance. The proposed approach provides an effective rate of sensitivity performance in detecting. The proposed approach attains 97% sensitivity, which is more reliable than existing approaches like SVM (60%), LR (70%), NB (54%), RF (60%), and CART (77%).

Specificity: It is considered a significant metric for detecting the PCO or non-PCO follicle class in an image without any modifications. Specificity performance is used to detect the negative results, which correctly detect the negative class. The following equation analyzes the performance of a TNR metric;

$$Specificity = \left[\frac{TruePos}{FlasePos + TrueNeg} \right] \quad (41)$$

Here $specificity$ represents specificity performance, $FalsePos$ represents the false positive, $TruePos$ represents the true positive, and $TrueNeg$ represents the true negative value. The following Fig. 5 represents the TNR performance analysis for proposed and existing approaches.

Fig. 5 illustrates the specificity performance for proposed and existing models. The proposed approach provides an effective rate of sensitivity performance. The proposed approach attains (96%) specificity, which is more reliable than existing approaches like SVM (92%) specificity, LR (94%) specificity, NB (53%) specificity, RF (92%) specificity, and CART (83%) of specificity.

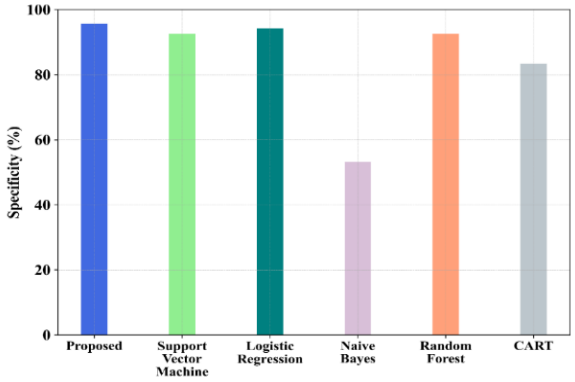


Fig. 5. Specificity performance.

F1-measures: The performance of an F1-measures is calculated between the recall and precision performance. The harmonic mean of the precision and recall performance is called an F-measure. The following equation calculates the F-measure performance;

$$F1-measure = \left[\frac{2 \times sensitivity \times \mathrm{Pr}ecision}{sensitivity + \mathrm{Pr}ecision} \right] \quad (42)$$

Here $F1-measure$ represents the F-measure performance, and $sensitivity \times \mathrm{Pr}ecision$ denotes the harmonic mean. The F-measure performance for proposed and existing approaches is depicted in the following Fig. 6.

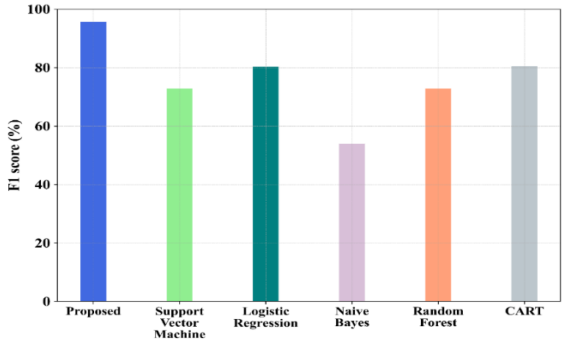


Fig. 6. F1-Measure performance.

Fig. 6 illustrates the F1-measure performance for a proposed and existing model. The proposed approach provides an effective rate of F1-measure performance in detecting PCO or non-PCO follicle class. The proposed approach attains (97%) of F1-measure in detecting PCO or non-PCO follicle class, which is more reliable than existing approaches like support vector machine (72%) F1-measure, logistic regression (80%) F1-measure, naive Bayes (53%) F1-measure, random forest (72%) F1-measure and CART (80%) F1-measure.

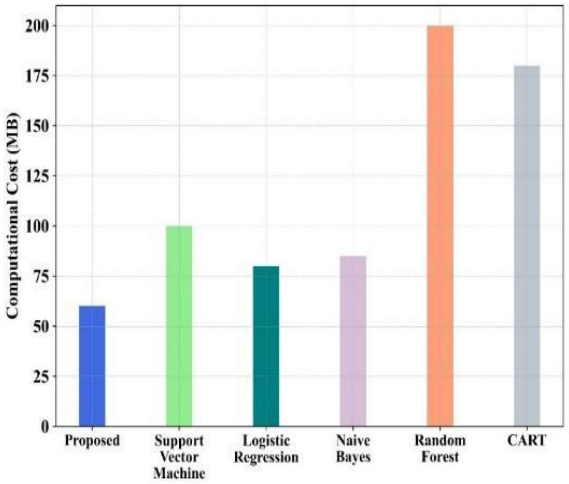


Fig. 7. Comparative analysis of the proposed model in terms of computational cost.

Fig. 7 provides an intricate analysis of the computational costs associated with the proposed model versus existing counterparts, elucidating the efficiency of our approach. The model exhibits a computational requirement of just 60MB, a substantial reduction compared to the other models, which have 100MB, 80MB, 85MB, 200MB, and 180MB of computational cost, respectively. This meticulous examination underscores the superior efficiency of the model, signaling its suitability for adoption in resource-constrained settings without compromising performance. The minimal computational cost underscores the streamlined complexity of the proposed model, making it a compelling choice for widespread implementation. The proposed model is compared with other models with a different dataset, as shown in Table II.

TABLE II. DATASET COMPARISON OF THE PROPOSED MODEL

Dataset	Method	Accuracy [%]
PCOS dataset from Kaggle [39]	HRFLR	87
PCOS dataset from UCI[40]	Red deer algorithm and RF	89.81
PCOS [41][42] [43]	PSO-SVM	90.18
Proposed	CNN-RNN	96

The proposed model outperforms its counterparts, trained on diverse datasets such as UCI and Kaggle. Despite the variations in dataset composition and characteristics, our model consistently demonstrates superior performance. This comparison encompasses thorough accuracy and illustrates the robustness of the approach across different data sources. The comprehensive analysis underscores the adaptability and efficacy of the model across a wide range of datasets, further bolstering its suitability for real-world applications across various domains. The Comparative value analysis of the proposed and other existing ML methods is given in Table III.

The comparative analysis of the proposed model with several performance measures is shown in Table III, and it shows that the proposed model reveals a better outcome.

TABLE III. COMPARATIVE VALUE ANALYSIS OF PROPOSED AND OTHER EXISTING ML METHODS

Methods	Accuracy [%]	Precision [%]	F1-measure [%]	Sensitivity [%]	Specificity [%]
SVM	85	92	72	60	92
Logistic Regression	89	94	80	70	94
Naive Bayes	64	53	53	54	53
Random Forest	85	92	72	60	92
CART	89	83	80	77	83
Proposed	96	96	97	97	96

V. LIMITATIONS OF THE STUDY

Based on the study, the attention-based CNN-RNN classification model is a reliable and valid approach to identifying Polycystic Ovary Syndrome (PCOS). The model is based on the efficiency of the PSO-based K-means clustering algorithm with a fuzzy filter and Contrast-Limited Adaptive Histogram Equalization (CLAHE) in pre-processing and segmenting ultrasound images. It is anticipated that performance metrics such as specificity, F1-score, sensitivity, accuracy, and precision effectively capture the strength of the model. However, there are flaws in the study, primarily the need for a larger dataset to enhance accuracy and minimize misclassification errors. The resilience of the model needs to be enhanced by testing it on smaller datasets. In addition, the ability of the model to be used broadly can be limited to the specific dataset and imaging conditions utilized in the study. Because it is based on ultrasound scans, the model may not account for other PCOS diagnostic criteria.

VI. FUTURE SCOPE

The application of Explainable AI (XAI) methods would significantly enhance the current research in identifying and predicting PCOS based on attention-based CNN-RNN classification models. Ensuring that the predictions of the AI model are understandable, comprehensible, and trustworthy to both doctors and patients would be the primary objective of these advances. By applying XAI methods such as Local Interpretable Model-agnostic Explanations (LIME) or SHapley Additive exPlanations (SHAP), it would be shown what factors or areas of the ultrasound images have the most impact on the conclusions of the model. Simplifying the AI decision-making process, these approaches would help healthcare professionals have more confidence in and trust the predictions of the AI.In addition, in order to further enhance the accuracy of the predicted model, additional research can focus on incorporating multiple data sources of different modalities, including

genomics, hormonal profiles, and clinical data. Explainable AI with these numerous data sources has the potential to give a broader picture of the patient's disease, leading to more personalized and effective treatment regimens. In addition, the model's generalizability and robustness to different populations would be increased through the utilization of a larger and more diverse dataset during training. Aside from enhancing PCOS detection, this approach would set a precedent for the incorporation of explainability in AI-based medical diagnostic systems.

VII. CONCLUSION

The conclusion of this study highlights the significance of the proposed approach in the early detection of PCOS in women. The proposed method employs a real-time dataset, pre-processed using the Contrast-Limited Adaptive Histogram Equalization (CLAHE) model, to improve image quality by reducing noise. Subsequently, segmentation is performed using a Particle Swarm Optimization (PSO)-based K-means clustering algorithm with a Fuzzy Filter (FF) to enhance detection performance before classification. Finally, PCOS detection is carried out using an attention-based Convolutional Neural Network-Recurrent Neural Network (CNN-RNN) model. The approach implemented in Python exhibits superior performance when compared to existing methods. The model achieved impressive accuracy (96%), precision (96%), sensitivity (97%), F1-score (97%), and specificity (96%) metrics. These results underscore the reliability and precision of the proposed approach in PCOS detection, surpassing other models in the field. However, there are some limitations to the proposed methodology. Particularly, the model's performance should be evaluated with larger dataset samples to enhance accuracy and minimize misclassification errors. Furthermore, validating PCOS detection with smaller dataset samples is essential for refining the model's robustness. Although the proposed approach demonstrates better results, continuous refinement and validation efforts are vital to guarantee its effectiveness in real-world PCOS detection scenarios.

REFERENCES

[1] Nazarudin, N. Zulkarnain, A. Hussain, S. S. Mokri, & I. M. "Nordin, Review on automated follicle identification for polycystic ovarian syndrome. *Bulletin of Electrical Engineering and Informatics*, vol. *9*, no. 2, pp.588-593, 2020.

[2] C. V. B. Palm, D. Glintborg, H. B. Kyhl, H. D. McIntyre, R. C. Jensen, T. K. Jensen,... & M. Andersen, "Polycystic ovary syndrome and hyperglycaemia in pregnancy. A narrative review and results from a prospective Danish cohort study. *Diabetes Research and Clinical Practice*, vol. *145*, pp.167-177, 2018.

[3] H. F. Escobar-Morreale, Polycystic ovary syndrome: definition, aetiology, diagnosis and treatment," *Nature Reviews Endocrinology*, vol. *14*, no. 5, pp.270-284, 2018.

[4] S. Behboudi-Gandevani, M. Amiri, R. Bidhendi Yarandi, M. Noroozzadeh, M. Farahmand, M. Rostami Dovom, & F. Ramezani Tehrani, "The risk of metabolic syndrome in polycystic ovary syndrome: A systematic review and meta-analysis," *Clinical endocrinology*, vol. *88*, no. 2, pp.169-184, 2018.

[5] Gibson-Helm, M. Tassone, E. C. Teede, H. J. Dokras, A. & Garad, R. "The needs of women and healthcare providers regarding polycystic ovary syndrome information, resources, and education: a systematic search and narrative review," In *Seminars in reproductive medicine* Vol. 36, No. 01, pp. 035-041 Thieme Medical Publishers.2018.

[6] P. Soni, & S. Vashisht, "Image segmentation for detecting polycystic ovarian disease using deep neural networks," *International Journal of Computer Sciences and Engineering*, vol. *7*, no. 3, pp.534-537, 2019.

[7] S. Patel, "Polycystic ovary syndrome (PCOS), an inflammatory, systemic, lifestyle endocrinopathy," *The Journal of steroid biochemistry and molecular biology*, vol. *182*, pp.27-36, 2018.

[8] Q. Zhang, Z. K. Bao, M. X. Deng, Q. Xu, D. D. Ding, M. M. Pan,... & F. Qu, " Fetal growth, fetal development, and placental features in women with polycystic ovary syndrome: analysis based on fetal and placental magnetic resonance imaging. *Journal of Zhejiang University. Science. B*, vol. *21*, no. 12, pp.977, 2020.

[9] S. M. Y. Farooq, N. Zulfiqar, M. Zulfiqar, S. A. Gilani, Z. Fatima, A. Hanif, & A. Akhter, "Use of Uterine and Ovarian Arteries Doppler Parameters for the Prediction of Infertility in Females," *Exclusive breastfeeding for the first six months after birth: A cross-sectional study in health care centers in Khartoum, Sudan.. page* vol. *67*, no. 10.

[10] K. S. Kumar, V. Nirmala, K. Venkatalakshmi, & K. Karthikeyan, "Analysis of optimization algorithms on follicles segmentation to support polycystic ovarian syndrome detection," *Journal of Computational and Theoretical Nanoscience*, vol. *15*, no. 1, pp.380-391, 2018.

[11] M. Sumathi, P. Chitra, R. S. Prabha, & K. Srilatha, "Study and detection of PCOS related diseases using CNN,"In *IOP Conference Series: Materials Science and Engineering* Vol. 1070, No. 1, p. 012062, 2021. IOP Publishing.

[12] B. Y. Jarrett, H. Vanden Brink, A. L. Oldfield, & M. E. Lujan, Ultrasound characterization of disordered antral follicle development in women with polycystic ovary syndrome," *The Journal of Clinical Endocrinology & Metabolism*, vol. *105*, no. 11, pp. e3847-e3861, 2020.

[13] P. G. Yılmaz, & G. Özmen, "Follicle detection for polycystic ovary syndrome by using image processing methods," *International Journal of Applied Mathematics Electronics and Computers*, vol. *8*, no. 4, pp.203-208, 2020.

[14] K. Aggarwal, M. M. Mijwil, A. H. Al-Mistarehi, S. Alomari M., Gök, A. M. Z. Alaabdin, & S. H. Abdulrhman, "Has the future started? The current growth of artificial intelligence, machine learning, and deep learning," *Iraqi Journal for Computer Science and Mathematics*, vol. *3*, no. 1, pp.115-123, 2022.

[15] S. Srivastava, P. Kumar, V. Chaudhry, & A. Singh, "Detection of ovarian cyst in ultrasound images using fine-tuned VGG-16 deep learning network," *SN Computer Science*, vol. *1*, no. 2, pp. 81,2020.

[16] S. C. Nandipati, & C. X. Ying, "Polycystic Ovarian Syndrome (PCOS) classification and feature selection by machine learning techniques," *Applied Mathematics and Computational Intelligence (AMCI)*, vol. *9*, 65-74, 2020.

[17] V. Deepika, "Applications of artificial intelligence techniques in polycystic ovarian syndrome diagnosis," *J. Adv. Res. Technol. Manag. Sci*, vol. *1*, no. 3, pp.59-63, 2019.

[18] S. Kharya, E. M. Onyema, A. Zafar, M. A. Wajid, R. K. Afriyie, T. Swarnkar, & S. Soni, "Weighted Bayesian belief network: a computational intelligence approach for predictive modeling in clinical datasets," *Computational Intelligence and Neuroscience*, vol. *2022*, no. 1, pp.3813705, 2022.

[19] J. Parashar, V. S. Kushwah, & M. Rai, "Determination human behavior prediction supported by cognitive computing-based neural network," In *Soft Computing: Theories and Applications: Proceedings of SoCTA 2022* pp. 431-441, 2023. Singapore: Springer Nature Singapore.

[20] S. Bharati, P. Podder, & M. R. H. Mondal, "Diagnosis of polycystic ovary syndrome using machine learning algorithms,"In *2020 IEEE region 10 symposium (TENSYMP)* pp. 1486-1489, 2020. IEEE.

[21] A. Denny, A. Raj, A. Ashok, C. M. Ram, & R. George, "i-hope: Detection and prediction system for polycystic ovary syndrome (pcos) using machine learning techniques," In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)* pp. 673-678, 2019. IEEE.

[22] S. Ramamoorthy, & R. Sivasubramaniam, "Monitoring the growth of polycystic ovary syndrome using mono-modal image registration technique: Application of medical big data in healthcare," In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data* pp. 180-187, 2019

[23] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, & X. Yang, "Deep learning in medical image registration: a review," *Physics in Medicine & Biology*, vol. *65*, no. 20, pp. 20TR01, 2020.

[24] N. S. Nilofer, "Follicles classification to detect polycystic ovary syndrome using GLCM and novel hybrid machine learning," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. *12*, no. 7, pp.1062-1073, 2021.

[25] C. Gopalakrishnan, & M. Iyapparaja, "Active contour with modified Otsu method for automatic detection of polycystic ovary syndrome from ultrasound image of ovary," *Multimedia Tools and Applications*, vol. *79*, no. 23, pp. 17169-17192, 2020.

[26] D. He, L. Liu, S. Miao, X. Tong, & M. Sheng, "Probabilistic guided polycystic ovary syndrome recognition using learned quality kernel," *Journal of Visual Communication and Image Representation*, vol. *63*, pp. 102587, 2019.

[27] C. Gopalakrishnan, & M. Iyapparaja, "Detection of polycystic ovary syndrome from ultrasound images using SIFT descriptors," *Bonfring International Journal of Software Engineering and Soft Computing*, vol. *9*, no. 2, pp. 26-30, 2019.

[28] U. N. Wisesty, J. Nasri, & Adiwijaya. "Modified backpropagation algorithm for polycystic ovary syndrome detection based on ultrasound images," In *Recent Advances on Soft Computing and Data Mining: The Second International Conference on Soft Computing and Data Mining (SCDM-2016), Bandung, Indonesia, August 18-20, 2016 Proceedings Second* pp. 141-151, 2017. Springer International Publishing.

[29] E. Setiawati, & A. B. W. Tjokorda, "Particle swarm optimization on follicles segmentation to support PCOS detection," In *2015 3rd international conference on information and communication technology (ICoICT)* pp. 369-374, 2015. IEEE.

[30] H. P. Kumar, & S. Srinivasan, "Segmentation of polycystic ovary in ultrasound images," In *Second International Conference on Current Trends In Engineering and Technology-ICCTET 2014*, pp. 237-240 IEEE.

[31] E. M. Onyema, T. A. Ahanger, G. Samir, M. Shrivastava, M. Maheshwari, G. M. Seghir, & D. Krah, " Empirical Analysis of Apnea Syndrome Using an Artificial Intelligence-Based Granger Panel Model Approach," *Computational Intelligence and Neuroscience*, vol. *2022*, no. 1, pp.7969389, 2022.

[32] S. Tiwari, L. Kane, D. Koundal, A. Jain, A. Alhudhaif, K. Polat,... & S. A. Althubiti, "SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning," *Expert Systems with Applications*, vol. *203*, pp.117592, 2022.

[33] A. Alamoudi, I. U. Khan, N. Aslam, N. Alqahtani, H. S. Alsaif, O. Al Dandan,... & R. Al Bahrani, " A deep learning fusion approach to diagnosis the polycystic ovary syndrome pcos,"*Applied Computational Intelligence and Soft Computing*, vol. *2023*, no. 1, pp.9686697, 2023.

[34] W. Lv, Y. Song, R. Fu, X. Lin, Y. Su, X. Jin,... & G. Huang, "Deep learning algorithm for automated detection of polycystic ovary syndrome using scleral images. *Frontiers in Endocrinology*, vol. *12*, pp.789878, 2022.

[35] J. Ma, X. Fan, S. X. Yang, X. Zhang, & X. Zhu, " Contrast limited adaptive histogram equalization-based fusion in YIQ and HSI color spaces for underwater image enhancement," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. *32*, no. 07, pp.1854018, 2018.

[36] D. Parasar, & V. R. Rathod, "Particle swarm optimisation K-means clustering segmentation of foetus ultrasound image," *International Journal of Signal and Imaging Systems Engineering*, vol. *10*, no. 1-2, pp.95-103, 2017.

[37] Y. Hu, Y. Wong, W. Wei, Y. Du, M. Kankanhalli, & W. Geng, "A novel attention-based hybrid CNN-RNN architecture for sEMG-based gesture recognition," *PloS one*, vol. *13*, no. 10, pp.e0206049, 2018.

[38] M. M. Hassan, & T. Mirza, "Comparative analysis of machine learning algorithms in diagnosis of polycystic ovarian syndrome," *Int. J. Comput. Appl*, vol. *975*, no. 8887, 2020.

[39] S. A. Bhat, "*Detection of polycystic ovary syndrome using machine learning algorithms* ," (Doctoral dissertation, Dublin, National College of Ireland.2021.

[40] S. Sreejith, H. K. Nehemiah, & A. Kannan, "A clinical decision support system for polycystic ovarian syndrome using red deer algorithm and random forest classifier," *Healthcare Analytics*, vol. *2*, pp.100102, 2022.

[41] L. H. Shaufee, H. Jantan, & U. F. M. Bahrin, "Polycystic Ovary Syndrome (PCOS) Prediction System Using PSO-SVM," *Journal of Computing Research and Innovation*, vol. *9*, no. 1, pp. 269-282, 2024.

[42] Kadam, M. (2024). Diagnosis of Polycystic Ovary Syndrome (PCOS) using Deep Learning and Classification Technique's. *Indian Scientific Journal Of Research In Engineering And Management*, *08*(05), 1–5. https://doi.org/10.55041/ijsrem35318

[43] Nurdiawan, O., Susana, H., & Faqih, A. (2024). Deep learning for polycystic ovarian syndrome classification using convolutional neural network. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *9*(2), 218–226. https://doi.org/10.33480/jitk.v9i2.4575

[44] Prathibanandhi, J., & Vimala, G. S. A. G. (2024). *Diagnosis of Polycystic Ovary Syndrome Using Deep Learning Neural Network-based Segmentation Techniques*. 1–7. https://doi.org/10.1109/iceeict61591.2024.10718597

[45] *Classification of Ultrasound PCOS Image using Deep Learning based Hybrid Models*. (2023). https://doi.org/10.1109/icears56392.2023.10085400

[46] Shanmugavadivel, K., Dhar, M., Mahesh, T. R., Al-Shehari, T., Alsadhan, N., & Yimer, T. (2024). Optimized polycystic ovarian disease prognosis and classification using AI based computational approaches on multi-modality data. *BMC Medical Informatics and Decision Making*, *24*(1). https://doi.org/10.1186/s12911-024-02688-9

[47] Prathibanandhi, J., & Vimala, G. S. A. G. (2024). *Diagnosis of Polycystic Ovarian Syndrome with the Implementation of Deep Learning Models*. 1–7. https://doi.org/10.1109/iceeict61591.2024.10718482

[48] Srivastav, S., Guleria, K., & Sharma, S. (2024). *A Transfer Learning-Based Fine Tuned VGG16 Model for PCOS Classification*. 1074–1079. https://doi.org/10.1109/idciot59759.2024.10467747

A Review of AI and IoT Implementation in a Museum's Ecosystem: Benefits, Challenges, and a Novel Conceptual Model

Shinta Puspasari[1]*, Indah Agustien Siradjuddin[2], Rachmansyah[3]
Department of Informatics-Faculty of Computer Science, Universitas Indo Global Mandiri, Palembang, Indonesia[1]
Department of Informatics-Faculty of Engineering, Universitas Trunojoyo, Madura, Indonesia[2]
Department of Computer System-Faculty of Computer Science, Universitas Indo Global Mandiri, Palembang, Indonesia[3]

***Abstract*—The museums need to transform into modern museums by developing a digital ecosystem that integrates all elements in the museum to optimize organizational outcomes and impact people's welfare in the era of Society 5.0. This paper aims to conduct a review of the museum's digital ecosystem based on the implementation of artificial intelligence (AI) and internet of things (IoT). PRISMA methodology for literature review was adopted to search for the answers to the research questions, knowing digital technology trends, challenges, and benefits of a digital museum ecosystem development, and proposed a novel conceptual model of the museum ecosystem based on AI and IoT implementation. The dataset contained metadata from Scopus, Google Scholar, and IEEExplore databases. Several stages were implemented in the literature review process so that it is known that AI and IoT technologies have never been separated in the development of digital museums since 2020, but there has yet to be research on the digital museum ecosystem model that integrates IoT and AI. The museum's digital ecosystem implementation benefits will improve museum resources and increase museum competitiveness. However, there will be challenges related to cybersecurity issues, data integration in multi-media formats, and interface designs to overcome user acceptance challenges of the technology constructed in the digital museum ecosystem. The proposed AI and IoT-based model also require an evaluation for implementation validation at the museum in future works.**

***Keywords*—AI; IoT; digital museum; digital ecosystem**

I. INTRODUCTION

The COVID-19 pandemic that spread throughout the world has had a negative and positive impact. Various sectors of people's lives have transformed quickly to adjust to and overcome the problems during the COVID-19 pandemic. The community has adapted to new habits where community activities are based on Information and Communication Technology (ICT), which allows communication, community services, and business to be held without direct contact and is more efficient than conventional methods. One is in the Museum's public service sector for education, culture, and tourism. The traditional Museum has been transformed into a modern ICT-based museum by presenting interesting technology for visitors to add value to the Museum [1]. The digital Museum's performance during the pandemic was optimal compared to traditional museums. Digital museums provide a variety of applications to support exhibition spaces and services for visitors for educational purposes and museum tours. However, more than a digital museum is needed to provide added value to museums and society, especially, toward Society 5.0. In that era, people demanded convenience in various public services supported by artificial intelligence-based ICT infrastructure that promised smart services to the public [2]. Stakeholder communities for museums will benefit more from the existence of a digital museum ecosystem.

The digital ecosystem is a concept of connectedness between human entities, technology, and organizational elements to provide added value in achieving organizational goals [3]. These entities are connected digitally to communicate and collaborate in the organization's business processes [4], [5]. Modern museums make it possible to develop digital ecosystems because they involve the community and various entities, which, if fully integrated with digital infrastructure, will provide added value in achieving museum tasks for education and tourism after the COVID-19 pandemic. Visitors with diverse backgrounds have different motivations and expectations, so they need an integrated system to exchange information effectively and efficiently to support the right and fast decision-making in supporting the performance of the Museum as a public service organization.

The digital museum ecosystem has been developed in previous research [6]. However, it is not yet known how developments, benefits, and challenges are faced, especially in the era of Society 5.0 where the application of artificial intelligence (AI) models and the Internet of Things (IoT) is a must. IoT offers ways to collaborate and more interconnection among people in an ecosystem [7] and AI delivers a smart way for living [8]. These questions will be answered in this study. This research conducts a literature review study on research on developing a digital museum ecosystem towards Society 5.0. This method makes it possible to carry out an analysis of the literature dataset regarding the digital ecosystem in a structured manner [9]. The research begins with formulating research questions, followed by applying the dataset inclusion and exclusion criteria until the final results are obtained to answer the research questions [10]. Based on the findings, a novel conceptual model of AI and IoT-based museum ecosystem is proposed and presented in this paper. The paper begins with the introduction in Section I discusses the research problem, gap, and related works Section II describes the research methodology, followed by the results and discussion in Section

IV. Finally, the conclusion is presented in Section V at the end of this paper.

II. Research Methodology

The literature review is carried out to extract information systematically [11]. The literature review begins with developing a dataset containing documents related to the research. It is continued with filtering steps based on several criteria so that a dataset containing articles is produced and will be analyzed to get answers to research questions [11], [12]. To find out the current development of the digital museum ecosystem, three research questions are formulated to be answered in this study, *RQ1*, *RQ2*, *RQ3*, and *RQ4*, as follows:

RQ1: What is the trend of digital technology application at the museum?

RQ2: What are the benefits of IoT and AI-based museum's ecosystem development?

RQ3: What are the challenges in IoT and AI-based museum's ecosystem development?

RQ4: What is the novel conceptual model of AI and IoT-based museum's ecosystem?

The methodology of the literature review process for finding this research question's answer adopted PRISMA [12]. Previous studies reviewed articles regarding existing models of museum digital ecosystems based on IoT and AI. The search was continued with article metadata in the developed database. The identification of new studies in databases was conducted using the following steps: database development, keywords for articles search, inclusion and exclusion criteria, and analysis of final search results. One thousand three articles were selected using the criteria in PRISMA, leaving ten articles reviewed in full text after inclusion and exclusion (Fig. 1). The study continued with searching for new studies on the IEEExplore website using the same keyword on database search. The final articles were analyzed to find the answers to research questions".

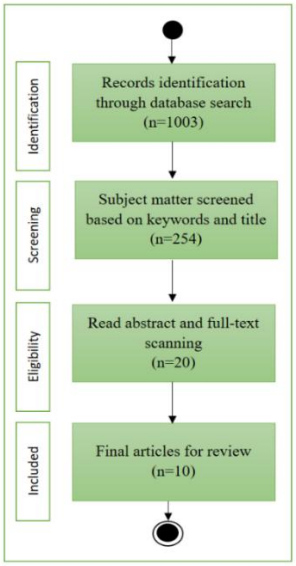


Fig. 1. Literature review methodology.

A. Identification

The development of a database to build a literature dataset that will be reviewed systematically is carried out in the early stages of the research. Articles identification was conducted via metadata of Scopus and Google Scholar databases chosen to develop the dataset because Scopus contains up-to-date, reputable articles that guarantee the latest research on digital ecosystem museums. The Publish and Perish tools collected metadata from the Scopus and Google Scholar databases. The formed dataset is saved. RIS file format for analysis using the VosViewer tool. The article search period for the last five years, 2019-2023.

B. Screening

Article searches in the Scopus and Google Scholar databases are heavily influenced by keywords to produce relevant articles with research problems. Keywords are repeatedly tried so that the best search results are obtained based on the articles found, namely with the keywords "Artificial Intelligence" OR "Internet of Things" Digital Ecosystem" OR "Digital Museum" OR "AI" or "IoT." With these keywords, a dataset was successfully formed from the Scopus and Google Scholar databases for the 2019-2023 period of 1003 articles.

C. Eligibility

The initial search results are corrected to find the most relevant articles to be analyzed to extract information that effectively answers the research problem. The process of inclusion and exclusion of articles in the dataset was carried out by filtering based on keywords in the article's title so that there were only 255 articles in the dataset. Furthermore, the screening process was carried out based on the abstract and full text scanning to obtain the final results of 20 articles to be analyzed in full-text.

D. Inclusion

The dataset analysis was carried out by mapping the bibliography in the dataset with the VosViewer tool [13] and continued with an analysis of 20 full-text articles to find answers to questions *RQ1*, *RQ2*, and *RQ3*. Bibliographical mapping will show the linkages between articles found based on keywords and some information related to the trend of digital research on the museum ecosystem toward Society 5.0. A full-text analysis is needed to find the challenges and benefits of developing digital museum ecosystems, especially during the COVID-19 pandemic in 2019-2023.

III. Results

The digital ecosystem is the concept of connectedness between elements in an organization with digital infrastructure to enable communication and collaboration in achieving goals and providing added value to the organization [4]. Elements contained in the digital ecosystem interact with each other to attentively produce information to support efficient decision-making by actors in the system related to digital technology, data, services, and organizational partnerships [14]. The main entity of the digital ecosystem is people, and the goal is to serve people [15]. The digital ecosystem within an organization [16] consists of a digital economy, digital infrastructure and adoption, and digital society, rights, and governance.

Information and communication technology infrastructure availability is central to forming a digital ecosystem within the organization [3], such as museums.

As a public service organization with roles for education, collection storage, and art, culture, and history tourism, the museum needs to form a digital museum ecosystem. The ecosystem enables visitors, service personnel, collections, exhibition rooms, museum facilities, and all entities and stakeholders in the museum to integrate, exchange information, and collaborate in carrying out museum roles. Additionally, with the digital museum ecosystem, it is possible to add value to society with the presence of ICT. The web-based digital ecosystem of museums has connected people in museums and collections [17]. However, more than a web-driven digital ecosystem for museums is needed for Society 5.0. In that era, people needed the support of digital technology to improve the quality of their lives and provide solutions to various life problems to present sustainable development in various aspects of life, including cultural heritage and museums [18]. IoT integration in digital ecosystem development is a promising solution for data communication and collaboration problems among people in an ecosystem [19], [20].

Meanwhile, AI optimizes data processing for an organization's competitive advantage [21]. IoT and AI impact the ecosystem model for improving its performance [22], [23]. The state of the art of the museum's digital ecosystem model, which integrates IoT and AI, needs to be explored for development. With a digital museum ecosystem, cultural preservation and resilience through museums can be maintained. It is effective in providing social impacts and has an economic impact on society.

A. RQ1: Trend of Digital Technology Application at Museum

The *RQ1* also tried to explore the related works of the digital museum ecosystem model proposed in this study. The finding may describe the gap between the existing model of the digital museum ecosystem. The dataset to be reviewed consists of bibliographies described statistically as in Table I. The average citation of 62.38 is relatively high because there is literature in a type of books cited > 1000. If the literature were removed from the dataset, the average number of citations would be 31.23. The distribution of literature with the number of citations > 1000 is described in Table II. The book was the most cited literature in this study, with citations double the journal articles' citation numbers.

Bibliographic mapping was carried out with the VosViewer tool, and the results of the overlay visualization are illustrated in Fig. 2. Based on the visualization of Fig. 2, which shows mapping based on keywords in the bibliography, it is known that there is no direct relationship between the four keywords "MUSEUM," "AI," "IoT," and "DIGITAL ECOSYSTEM."

The bibliography on the topic of the museum is quite a lot. However, those related to AI, IoT, and Digital Ecosystems have opportunities for further research. The trend of developing museums based on digital technology and artificial intelligence can be seen in publications starting to develop in 2020. For example, there are networks with the keywords "METAVERSE"," "MACHINE LEARNING", and "DIGITAL ECOSYSTEM" after 2020, which are marked by visualization in dominant yellow color. These keywords are also accompanied by the emergence of "COVID", which is connected with the keyword "MUSEUM". It can be concluded that the COVID-19 pandemic has forced the museum's digital implementation to adapt to a pandemic situation by developing museums based on AI and METAVERSE and leading to a digital ecosystem.

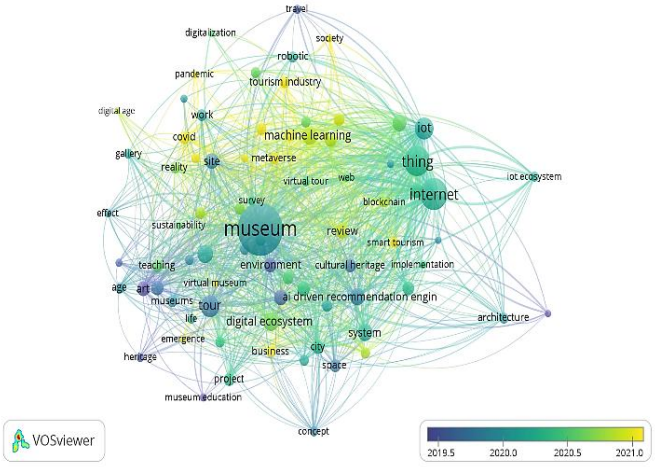


Fig. 2. Overlay visualization of bibliography mapping.

TABLE I. DATASET STATISTICS IN THIS RESEARCH

Information	Results
Document	1033
Citation	62.38
Authors	2417
Multi-authored Documents	657
Single-authored Documents	345
Authors per Document	2.41
Co-Authors	1.41

TABLE II. SUBSET OF DATASET WITH CITATIONS >1000

Literature Type	Frequency	Citations
Journal Articles	5	11227
Book	6	31013
Citation	1	1011

The process was continued with selection, inclusion, and execution, and twenty documents were obtained for overall review. After conducting a literature review of 20 articles in full text, it is known that ten articles contain the development of digital technology to construct digital ecosystems that are applied in the museum domains with IoT and AI content. The digital technology development trend can be seen per year, as illustrated in Table III. The development of a digital museum ecosystem model has been carried out in previous research that is only based on digital technology with AI content. Meanwhile, a digital museum ecosystem model integrating IoT and AI has yet to be developed. The trend of developing digital technology in museums and tourism is described in Fig. 3. The development of digital technology for museum applications began to integrate

AI models in 2019, which were integrated with IoT. The development of AI technology in museums from 2020 to 2022 is relatively high. A pandemic situation requires a solution to the problem of limiting distance and interaction with exhibitions, collections, and visitors. AI and IoT are present as solutions that provide convenience and automation for museum management during a pandemic [24]. This condition has triggered the rapid development of the digital ecosystem at museums based on AI and IoT content since 2020. These results answer research question *RQ1*.

TABLE III. THE TREND OF MUSEUM DIGITAL ECOSYSTEM DEVELOPMENT

References	IoT	AI	Digital Ecosystem Model
[25], [26]	-	✓	NA
[27]	✓	✓	NA
[6]	✓	-	Available
[28], [18]	✓	-	NA
[29], [30]	✓	✓	NA
[25], [26]	-	✓	NA

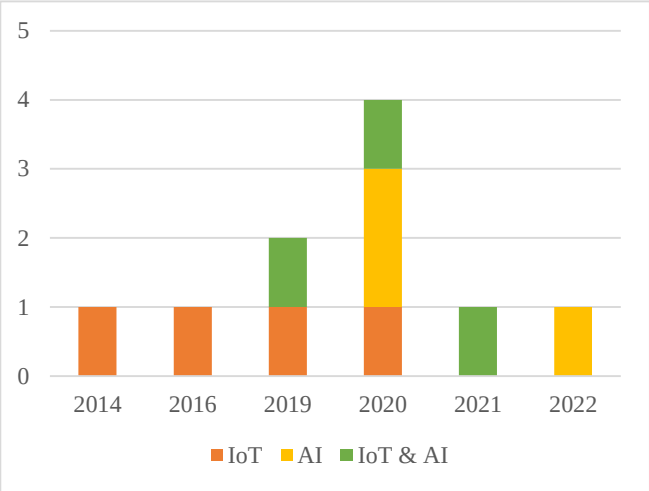


Fig. 3. The trend of IoT and AI implementation in museum.

B. RQ2: Benefits of the Museum's Ecosystem Development

The visitors come to the museum with various motivations, including learning and entertainment. The monotonous presentation of museum showrooms requires a practical touch of technology to enhance the learning experience [31], [32] and the visitor's knowledge, especially students, as dominant visitors [33]. Museum digital transformation is more than just digitizing the museum. It requires designing a digital ecosystem model that benefits society and museum organizations, as described in Table IV. Integrating IoT technology and AI models in an ecosystem will provide added value for all the museum's people. Impact on museum resources, thereby creating the competitive advantages of the museum [29].

The optimization of museum services for people related to its role in education and tourism will be achieved through a digital ecosystem based on IoT and AI integration models. By utilizing digital technology, museums can optimize their performance during the pandemic, which limits museum visits [34]. Its efficiency can be achieved through AI and IoT implementation [30]. These findings answer the research question *RQ2*.

TABLE IV. BENEFITS OF THE AI AND IOT-BASED MUSEUM'S DIGITAL ECOSYSTEM TRANSFORMATION

Benefits	References
Visitor engagement: Visitors interact and connect actively with the content of the exhibit.	[26][29] [6] [27]
A new business process: Enabling a new process in museum management.	[18] [29] [6][35] [27]
Competitive advantage: Enabling a museum to outperform its competitors.	[29] [35]
Resource improvement: Enhanced quality and efficiency of the available museum resources.	[18][6] [35] [30][25] [36]
Service improvement: Enhancing service quantity, quality, and efficiency of the museum.	[18] [6] [35][30] [36]
Visitor satisfaction: Fulfilment and positive experience after interacting with museum resources or services.	[18] [29] [27]
Learning enhancement: Improvement of learning experience to foster a better understanding of the museum.	[29] [6] [26] [27]

C. RQ3: Challenges of the Museum's Ecosystem Development

There are challenges to be faced in developing the digital museum ecosystem. These challenges are described in Table V. Museum visitors who come from various backgrounds with low ability to use digital technology, especially those aged >30, face a challenge in accepting digital museum technology [37], [38]. The issue of cyber security and integration of data stored in various formats due to the implementation of IoT and various multimedia-based technologies is attractive to museum visitors [39][40]. It has become an obstacle that needs to be addressed, preceded by research in designing an effective digital ecosystem model for the museum domain. Interface design that is easily understood by various backgrounds of people's profiles.

TABLE V. BENEFITS OF THE AI AND IOT-BASED MUSEUM'S DIGITAL ECOSYSTEM DEVELOPMENT

Challenges	References
Technology acceptance: Accepting new technologies, reluctance to use the technologies at the museum because of lack of knowledge or refusing to use the technology.	[35] [27][26]
Cyber security and ethical concerns: Protecting the museum system and data from unauthorized access, disruption, or theft, is also behavior that has moral conflict concerning the privacy or well-being of the museum ecosystem.	[35] [30]
Feasible models and interfaces: Designing models and user interfaces suitable for different museums.	[18] [30][27]
Infrastructure improvement: A lot of cost for hardware improvement that supports the museum's digital ecosystem development.	[27] [36]
Data integration: The multimedia format creates a multi-data format that needs to be integrated into a museum's digital ecosystem.	[36][26]

Challenges also arise in the museum ecosystem for the implementation of digital technology by internal museum actors. Staff with inadequate knowledge of digital technology are an obstacle in carrying out the process of digitizing museum

collections. There is a lack of awareness of the benefits of digital transformation to increase competitiveness and optimize the role of museums compared to relying on physical museum collections as the main strength. This requires increased understanding from museum staff. High motivation from staff will accelerate the museum's digital transformation process, and the usage of digital technology to support the museum's role in education and tourism is also a challenge in developing the digital museum ecosystem [41]. These findings answer the research question *RQ3.*

D. RQ4: The Proposed AI and IoT-Based Museum Ecosystem Conceptual Model

In the museum digital ecosystem, the three main elements (people, services, and collection & facilities) collaborate and communicate to conduct the museum's roles by applying digital technology for optimizing the museum's organizational goals achievement. A proposed digital museum model is discussed in this study based on the literature review results presented as the answer of RQ4. Integrating IoT and AI in museum ecosystem models is expected to be effective in providing added value benefits for museum elements, especially visitors. Fig. 4 illustrates the proposed AI and IoT-based museum ecosystem model.

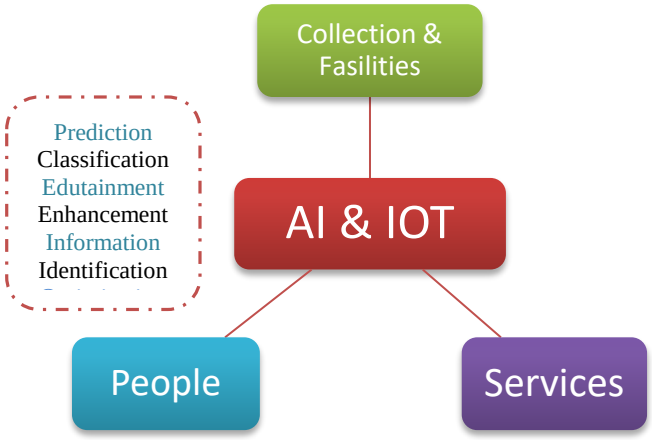


Fig. 4. The proposed AI and IoT-based museum ecosystem conceptual model.

a) Prediction: estimating museum visits to prevent overcrowding, especially during a pandemic which required distance among elements in the museum such as visitors and collections. Predictions try to find out which collections or exhibitions that most interesting, and estimate the resources and facilities [32], [42] needed for the operation and management of the museum.

b) Identification: determining the characteristics of objects due to manage and increase the museum experience. Object identification involves various processes, including computer vision for analyzing visual data, such as images or videos of museum objects as well as object recognition [43] effectively identifying objects presented to visitors to enhance experiences at museums.

c) Classification: Task classification is grouping objects or collections based on collection characteristics to determine effective and efficient handling [44] which impacts the quality and curiosity of the collection. For example, the accurate automated classification of damage types to a museum painting or archive collection will help determine quick and appropriate actions to handle problems based on the type of damage to the collection.

d) Information: AI-based applications in museums enhance information delivery. This feature has not been included in the previous model proposed at the beginning of this study [42] . By using technologies like machine learning and natural language processing, museums can offer personalized experiences, where visitors receive tailored content based on their preferences. AI can do automated tasks like cataloging the collection, and creating real-time updates to exhibits. It also enables seamless information retrieval through smart search systems and multilingual support, making museums more accessible [45]. AI able to analyze data to optimize exhibits and improve accessibility for diverse audiences. These advancements make information more organized, accessible, and interactive, enriching the overall museum experience and management.

e) Enhancement: The implementation of AI at the museum effectively adds value for exhibitions and collections that present different non-traditional new experiences for visitors that impact revisit intention. Digital interactive technology integrated 3D and augmented reality technology in the exhibition presenting virtual and informative collections that are impossible in traditional museums. Especially for the Gen-Z who live in the digital era and are used to all the convenience features and technological sophistication, AI and IoT implementation are necessary for museum visit enhancement. The traditional museums that seem old-fashioned and unattractive without digital technology as an additional tool for visitors and museum management.

f) Edutainment: The museum provides interactive digital technology delivered to the visitor to attract their interest and interactive experiences by using digital technology, such as Augmented reality and virtual reality games integrating AI for smart apps giving added value to increase satisfaction and interest to revisit the museum [31]. The AI-based applications are effective in delivering entertainment at the museum and improving visitors' learning experiences and performance.

g) Optimization: AI and IoT-based applications in museums optimize various aspects of operations and visitor experience. By analyzing visitor behavior, it helps adjust exhibition and content in real-time to enhance engagement. It also optimizes visitor flow, using data to predict crowded areas and suggest alternative routes to reduce congestion. Additionally, AI systems manage energy use by adjusting lighting, and temperature systems based on visitor activity, improving sustainability and reducing costs. Personalized content delivery, such as tailored tours and recommendations based on visitor's profiles, improves the visitor's satisfaction, experience, and learning performance. On the operational side, AI automates routine tasks like inventory management and staff scheduling, improving efficiency and resource allocation. AI and IoT enable museums to streamline their operations,

enhance engagement, and create a more sustainable and efficient environment for museum elements.

IV. DISCUSSION

The proposed model integrates IoT and AI in museum ecosystem models that better than other existing models compared in Table III and discussed in *RQ1*. The models also identify the specific functionality of the museum ecosystem with AI and IoT integration than other models. It contains prediction, classification, education, edutainment, information, identification, and optimization to support the museum's role in education and tourism. The proposed model can be a case line for museum ecosystem development based on IoT and AI to support museum management with the benefits and challenges discussed in *RQ2* and *RQ3*. Further research is necessary for the proposed model validation.

The AI and IoT integration in the museum ecosystem model benefits society and museum organizations, as described in Table IV. It will provide added value for all the museum's elements and create competitive advantages. The strategic planning to overcome the challenges in developing the museum's digital ecosystem is also necessary for the museum's goals achievement and fully utilized by people in the museum ecosystem, especially in Society 5.0.

The proposed model in this study is based on the literature review and observations. The development evaluation of AI and IoT-based museum ecosystem models is required in future research to answer challenges and present valid new concepts, techniques, and innovations to provide a new atmosphere for museums suitable for Society 5.0. The presence of a digital museum ecosystem will not dismiss the existence of museums but will complement and improve performance after the COVID-19 pandemic.

The relationship between elements or entities in the museum's ecosystem connected through the digital technology that integrates AI and IoT will impact the preservation of the historical, cultural, and artistic values stored in museums. The new digital museum ecosystem model based on IoT and AI is expected to create innovation for modern museums. With IoT connectivity capabilities, the speed, and accuracy provided by AI models, developing the research implementation and evaluation stages is challenging for future work.

V. CONCLUSION

The digital transformation of museums was accelerated by the pandemic, which required the museums to find solutions to keep providing educational and tourism services for the public. Towards the era of Society 5.0, people need added value from museums regarding convenience, which impacts their well-being promised by implementing AI and IoT in museums. This study surveys by reviewing the literature dataset of bibliographies with Scopus and Google Scholar metadata sources of research topics about AI and IoT in museum apps. The dataset was processed with PRISMA methodology. The selected 20 documents were thoroughly reviewed. The research questions that will answered are, what are the benefits, challenges, and the novel conceptual model of museum's ecosystem based on literature review conducted in this study.

Bibliography mapping based on keywords relevant to the research topic was conducted using the VosViewer app. It was found that the development trend of digital museum technology started to be based on IoT and AI from 2019 to 2022, which was triggered by the COVID-19 pandemic. AI has become an integrated model in the development of digital museums since 2020 because of its ability to automate, making it easy for visitors and museum managers to deal with a pandemic. IoT integration in the museum's digital ecosystem development is a promising solution for data communication and collaboration problems among people as the main entity of the museum ecosystem. The background of the people involved in the museum, especially museum visitors, presents challenges in user acceptance to take advantage of digital technology in the museum ecosystem, which will have an impact on improving the quantity and quality of museum resources and lead to increasing museum competitiveness as the added value will be obtained, especially in the Society 5.0. However, the absence of a digital museum ecosystem model validation is an opportunity to develop a museum ecosystem based on the proposed model for model validation in future work. The proposed museum ecosystem contains three elements, people, services, and collection & facilities, which are designed to be integrated by AI & IoT-based technology. AI and IoT will facilitate the museum's role and present functionality for prediction, identification, classification, information, enhancement, edutainment, and optimization. AI and IoT implementation may improve museum performance and present added value to museum elements, especially visitors. Further research is necessary for the proposed model effectiveness validation.

ACKNOWLEDGMENT

This study was funded by the Ministry of Education and Culture, Research, and Technology, the Republic of Indonesia, through The Domestic Cooperation Research Grand (PKDN) 2023. Contract ID 178/E5/PG. 02.00.PL/2023.

REFERENCES

[1] H. Chen and C. Ryan, "Transforming the museum and meeting visitor requirements: The case of the Shaanxi History Museum," *J. Destin. Mark. Manag.*, vol. 18, no. April, p. 100483, 2020, doi: 10.1016/j.jdmm.2020.100483.

[2] A. Tlili, R. Huang, and x Kinshuk, "Metaverse for climbing the ladder toward 'Industry 5.0'and 'Society 5.0'?," *Serv. Ind. J.*, 2023, [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/02642069.2023.2178644.

[3] M. Koch, D. Krohmer, M. Naab, D. Rost, and M. Trapp, "A matter of definition: Criteria for digital ecosystems," *Digit. Bus.*, vol. 2, no. 2, p. 100027, 2022, doi: 10.1016/j.digbus.2022.100027.

[4] X. Li, L. Zhang, and J. Cao, "Research on the mechanism of sustainable business model innovation driven by the digital platform ecosystem," *J. Eng. Technol. Manag.*, vol. 68, no. January, p. 101738, 2023, doi: 10.1016/j.jengtecman.2023.101738.

[5] S. Wolfert, C. Verdouw, L. Van Wassenaer, W. Dolfsma, and L. Klerkx, "Digital innovation ecosystems in agri-food : design principles and organizational framework," *Agric. Syst.*, vol. 204, no. November 2022, p. 103558, 2023, doi: 10.1016/j.agsy.2022.103558.

[6] P. M. D. Ivanova, S. J. Stoikov, P. L. Radoslavova, and L. D. Mihailov, "System architecture and intelligent data curation of virtual museum for ancient history," *SPIIRAS Proc.*, vol. 18, no. 2, pp. 444–470, 2019, doi: 10.15622/sp.18.2.444-470.

[7] S. Kubler, J. Robert, A. Hefnawy, K. Främling, C. Cherifi, and A. Bouras, "Open IoT Ecosystem for Sporting Event Management," *IEEE Access*, vol. 5, pp. 7064–7079, 2017, doi: 10.1109/ACCESS.2017.2692247.

[8] R. Qureshi, M. Irfan, S. Ali, A. Shah, T. M. Gondal, and F. Sadak, "Artificial Intelligence and Biosensors in Healthcare and Its Clinical Relevance : A Review," *IEEE Access*, vol. 11, no. June, pp. 61600–61620, 2023, doi: 10.1109/ACCESS.2023.3285596.

[9] S. Suuronen, J. Ukko, R. Eskola, R. S. Semken, and H. Rantanen, "A systematic literature review for digital business ecosystems in the manufacturing industry_ Prerequisites, challenges, and benefits," *CIRP J. Manuf. Sci. Technol.*, vol. 37, pp. 414–426, 2022, doi: 10.1016/j.cirpj.2022.02.016.

[10] M. Inkarbekov, R. Monahan, and B. A. Pearlmutter, "Visualization of AI Systems in Virtual Reality : A Comprehensive Review," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 8, pp. 33–42, 2023.

[11] R. S. Nuzulismah, H. B. Santoso, and P. O. H. Putra, "E-Learning Adoption Readiness in Secondary Education of Developed and Developing Countries: a Systematic Literature Review," *J. Ilm. Kursor*, vol. 11, no. 4, pp. 165–178, 2022, [Online]. Available: http://kursorjournal.org/index.php/kursor/article/view/301.

[12] B. R. Neciosup-bolaños and S. E. Cieza-mostacero, "The Heart of Artificial Intelligence : A Review of Machine Learning for Heart Disease Prediction," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 12, pp. 80–85, 2024.

[13] O. J. Aroba, N. Naicker, T. T. Adeliyi, A. Gupthar, and K. Karodia, "A Review: The Bibliometric Analysis of Emerging Node Localization in Wireless Sensor Network," *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 15, no. 2023, pp. 141–153, 2023.

[14] E. Brea, "A framework for mapping actor roles and their innovation potential in digital ecosystems," *Technovation*, vol. 125, no. December 2021, p. 102783, 2023, doi: 10.1016/j.technovation.2023.102783.

[15] M. J. Anwar and A. Q. Gill, "A review of the seven modelling approaches for digital ecosystem architecture," *Proc. - 21st IEEE Conf. Bus. Informatics, CBI 2019*, vol. 1, pp. 94–103, 2019, doi: 10.1109/CBI.2019.00018.

[16] USAID, "Digital Ecosystem Framework," 2022.

[17] P. Eklund, P. Goodall, and T. Wray, "Virtual museums and Web-based digital ecosystems," *4th IEEE Int. Conf. Digit. Ecosyst. Technol. - Conf. Proc. IEEE-DEST 2010, DEST 2010*, pp. 141–146, 2010, doi: 10.1109/DEST.2010.5610657.

[18] A. Chianese and F. Piccialli, "Designing a smart museum: When cultural heritage joins IoT," *Proceedings - 2014 8th International Conference on Next Generation Mobile Applications, Services and Technologies, NGMAST 2014*. pp. 300–306, 2014, doi: 10.1109/NGMAST.2014.21.

[19] S. A. Kidanu, Y. Cardinale, G. Tekli, and R. Chbeir, "A Multimedia-Oriented Digital Ecosystem: A new collaborative environment," *2015 IEEE/ACIS 14th Int. Conf. Comput. Inf. Sci. ICIS 2015 - Proc.*, pp. 411–416, 2015, doi: 10.1109/ICIS.2015.7166629.

[20] N. Sabry and P. Krause, "A digital ecosystem view on cloud computing," *IEEE Int. Conf. Digit. Ecosyst. Technol.*, 2012, doi: 10.1109/DEST.2012.6227905.

[21] A. Kordon, "Applied Artificial Intelligence-Based Systems as Competitive Advantage," pp. 6–18, 2020.

[22] W. Luther, N. Baloian, D. Biella, and D. Sacher, "Digital Twins and Enabling Technologies in Museums and Cultural Heritage: An Overview," *Sensors*. mdpi.com, 2023, [Online]. Available: https://www.mdpi.com/1424-8220/23/3/1583.

[23] L. Bajenaru, M. Ianculescu, and C. Dobre, "A Holistic Approach for Creating a Digital Ecosystem Enabling Personalized Assistive Care for Elderly," *Proc. - 16th Int. Conf. Embed. Ubiquitous Comput. EUC 2018*, pp. 89–95, 2018, doi: 10.1109/EUC.2018.00020.

[24] F. Bandarin, E. Ciciotti, M. Cremaschi, G. Madera, P. Perulli, and D. Shendrikova, "After Covid-19: A survey on the prospects for cities," *City, Cult. Soc.*, vol. 25, no. July, p. 100400, 2021, doi: 10.1016/j.ccs.2021.100400.

[25] J. S. Goussous, "Artificial intelligence-based restoration: The case of petra," *Civil Engineering and Architecture*, vol. 8, no. 6. researchgate.net, pp. 1350–1358, 2020, doi: 10.13189/cea.2020.080618.

[26] M. Winter, L. Sweeney, K. Mason, and ..., "Low-power Machine Learning for Visitor Engagement in Museums," *Proceedings of the 6th …*. cris.brighton.ac.uk, 2022, [Online]. Available: https://cris.brighton.ac.uk/ws/files/34091316/Winter_et_al_2022_Low_power_Machine_Learning_for_Visitor_Engagement_in_Museums.pdf.

[27] B. Wang, "Digital design of smart museum based on artificial intelligence," *Mobile Information Systems*. hindawi.com, 2021, [Online]. Available: https://www.hindawi.com/journals/misy/2021/4894131/.

[28] A. Renato and P. D. Thesis, "The Internet of Things supporting the Cultural Heritage domain : analysis , design and implementation of a smart framework enhancing the smartness of cultural spaces Declaration of Authorship," *Universita degli Studi di Napoli Federico II*. fedoa.unina.it, 2016, [Online]. Available: http://www.fedoa.unina.it/10799/1/piccialli_upload.pdf.

[29] S. Kontogiannis, G. Kokkonis, I. Kazanidis, M. Dossis, and S. Valsamidis, "Cultural IoT Framework Focusing on Interactive and Personalized Museum Sightseeing," *Internet of Things*, pp. 151–181, 2020, doi: 10.1007/978-3-030-42573-9_10.

[30] K. Anatoly *et al.*, "CHPC: A complex semantic-based secured approach to heritage preservation and secure IoT-based museum processes," *Comput. Commun.*, vol. 148, pp. 240–249, 2019, doi: 10.1016/j.comcom.2019.08.001.

[31] S. Puspasari, N. Suhandi, and J. N. Iman, "Enhancing The Visitors Learning Experience in SMB II Museum Using Augmented Reality Technology," in *International Conference on Electrical Engineering and Informatics (ICEEI)*, 2019, pp. 296–300, doi: https://doi.org/10.1109/ICEEI47359.2019.8988831.

[32] S. Puspasari, Ermatita, and Zulkardi, "Machine Learning for Exhibition Recommendation in a Museum's Virtual Tour Application," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 4, pp. 404–412, 2022, doi: https://doi.org/10.14569/IJACSA.2022.0130448.

[33] E. Ermatita, S. Puspasari, and Z. Zulkardi, "Improving Student ' s Cognitive Performance during the Pandemic through a Machine Learning-Based Virtual Museum," *TEM Journal.*, vol. 12, no. 2, pp. 948–955, 2023, doi: 10.18421/TEM122.

[34] S. Puspasari, Ermatita, and Zulkardi, "Constructing Smart Digital Media for Museum Education Post Pandemic Recovery: A Review and Recommendation," in *2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS*, Oct. 2021, pp. 238–243, doi: https://doi.org/10.1109/ICIMCIS53775.2021.9699345.

[35] R. Kılıçhan and M. Yılmaz, "Artificial intelligence and robotic technologies in tourism and hospitality industry," *Erciyes Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, vol. 3, no. 50. dergipark.org.tr, pp. 353–380, 2020, [Online]. Available: https://dergipark.org.tr/en/pub/erusosbilder/article/838193.

[36] A. Lerario and A. Varasano, "An IoT smart infrastructure for S. Domenico Church in Matera's 'Sassi': A multiscale perspective to built heritage conservation," *Sustain.*, vol. 12, no. 16, 2020, doi: 10.3390/su12166553.

[37] S. Puspasari, N. Suhandi, and J. N. Iman, "Evaluation of Augmented Reality Application Development for Cultural Artefact Education," *Int. J. Comput.*, vol. 20, no. 2, pp. 237–245, 2021, doi: 10.47839/ijc.20.2.2171.

[38] S. Puspasari, E. Ermatita, and Z. Zulkardi, "Assessing an Innovative Virtual Museum Application using Technology Acceptance Model," *Int. J. Informatics Dev.*, vol. 11, no. 1, pp. 212–221, 2022, doi: 10.14421/ijid.2022.3758.

[39] M. Bouzidi, N. Gupta, and S. Member, "A Novel Architectural Framework on IoT Ecosystem , Security Aspects and Mechanisms : A Comprehensive Survey," *IEEE Access*, vol. 10, no. July, pp. 101362–101384, 2022.

[40] F. Ponsignon and M. Derbaix, "The impact of interactive technologies on the social experience: An empirical study in a cultural tourism context," *Tour. Manag. Perspect.*, vol. 35, no. April, p. 100723, 2020, doi: 10.1016/j.tmp.2020.100723.

[41] V. Kamariotou, M. Kamariotou, and F. Kitsios, "Strategic planning for virtual exhibitions and visitors' experience: A multidisciplinary approach for museums in the digital age," *Digit. Appl. Archaeol. Cult. Herit.*, vol. 21, no. December 2020, p. e00183, 2021, doi: 10.1016/j.daach.2021.e00183.

[42] S. Puspasari, I. A. Siradjuddin, Rachmansyah, M. A. F. Rahman, and D. Haversyalapa, "IoT and AI-Driven Conceptual Model of Museum Ecosystem," in 2023 International Conference on Electrical Engineering and Informatics (ICEEI), 2023, pp. 1–6, doi: 10.1109/ICEEI59426.2023.10346949.

[43] G. Ioannakis, L. Bampis, and A. Koutsoudis, "Exploiting artificial intelligence for digitally enriched museum visits," J. Cult. Herit., vol. 42, pp. 171–180, 2020, doi: 10.1016/j.culher.2019.07.019.

[44] C. Sandoval, "Two-Stage Deep Learning Approach to the Classification of Fine-Art Paintings," IEEE Access, vol. 7, pp. 41770–41781, 2019, doi: 10.1109/ACCESS.2019.2907986.

[45] J. Li, X. Zheng, I. Watanabe, and Y. Ochiai, "A systematic review of digital transformation technologies in museum exhibition," Comput. Human Behav., vol. 161, no. September 2023, p. 108407, 2024, doi: 10.1016/j.chb.2024.108407.

Optimizing the GRU-LSTM Hybrid Model for Air Temperature Prediction in Degraded Wetlands and Climate Change Implications

Yuslena Sari[1], Yudi Firmanul Arifin[2], Novitasari Novitasari[3], Samingun Handoyo[4], Andreyan Rizky Baskara[5], Nurul Fathanah Musatamin[6], Muhammad Tommy Maulidyanto[7], Siti Viona Indah Swari[8], Erika Maulidiya[9]
Department of Information Technology, Universitas Lambung Mangkurat, Banjarmasin, Indonesia[1, 5, 6, 8, 9]
Faculty of Forestry, Universitas Lambung Mangkurat, Banjarmasin, Indonesia[2]
Faculty of Engineering, Universitas Lambung Mangkurat, Banjarmasin, Indonesia[3]
Data Science Study Program, Brawijaya University, Malang, Indonesia[4]
Department of Mining Engineering, Politeknik Negeri Banjarmasin, Banjarmasin, Indonesia[7]

***Abstract*—Accurate air temperature prediction is critical, particularly for micro air temperatures. The temperature of micro air changes quickly. Micro and macro air temperatures vary, particularly in degraded wetlands. By predicting air temperature, climate change in a degraded wetland environment can be predicted earlier. Furthermore, micro and macro air temperatures are drought index parameters. Knowing the drought index can help you avoid disasters like fires and floods. However, the right indicators for predicting micro or macro temperatures have yet to be found. LSTM excels at tasks requiring complex long-term memory, whereas GRU excels at tasks requiring rapid processing. We proposed a deep learning strategy based on the GRU-LSTM Hybrid model. Both of these deep learning models are excellent for predicting time series. The performance of this hybrid model is affected by changes in model indicators. The preprocessing stage, the number of input parameters, and the presence or absence of a Dropout Layer in the model architecture are among the most influential indicators of model performance. The best macro temperature prediction performance was obtained using 12 monthly average data to predict the next month's temperature, yielding an RMSE of 0.056807, MAE of 0.046592, and R2 of 0.989371. This model also performed well in predicting daily micro temperature, with an RMSE of 0.227086, MAE of 0.190801, and R2 of 0.981802.**

Keywords*—Predictions; temperature; Gated Recurrent Unit (GRU); Long Short-Term Memory (LSTM); performance; indicators

I. INTRODUCTION

The wetland ecosystem and biological processes are influenced by air temperature. Temperature knowledge can provide insight into environmental changes and aid in effectively managing and conserving wetlands [1], [2]. Prediction of air temperature is critical in identifying and anticipating potential wetlands disasters. Temperature extremes can cause changes in rain patterns, affecting water discharge and potential flooding. Higher temperatures can cause more evaporation and more intense rain in a shorter time, potentially causing flooding in wetlands [3]-[5]. Extreme or unusual temperatures can harm wetland ecosystems. Overheating, for example, can cause a drop in water levels and dry out wetlands, which can disrupt habitats and threaten the survival of species that live there. In wetlands, hot, dry weather can increase the risk of fire. If extreme temperatures continue, dry vegetation and the risk of wildfires will increase, endangering ecosystems and the surrounding environment [6]-[8]. Because of the numerous effects of temperature changes, it is necessary to forecast temperature to mitigate natural disasters that may occur in the wetland environment.

Various temperature prediction methods have been explored, including ARIMA (Autoregressive Integrated Moving Average), SARIMA (Seasonal ARIMA), LSTM (Long Short-Term Memory), and GRU (Gated Recurrent Unit) models [9–13]. Many studies have employed time series data for predictions, often using Recurrent Neural Network (RNN) architectures. Improved RNN variants, such as GRU and LSTM, have been widely adopted due to their superior performance over traditional RNN models. Long input sequences in RNNs often lead to exploding and vanishing gradients, an issue LSTM effectively mitigates through its gating mechanism, which regulates information flow and retains long-term dependencies in sequential data [10], [14]. LSTM can extract historical patterns more efficiently, outperforming conventional time series forecasting methods. GRU, a simplified LSTM variant, has fewer parameters, making it computationally more efficient while maintaining comparable accuracy [15], [16]. The hybrid GRU-LSTM model was chosen because it leverages the strengths of both architectures: LSTM's ability to capture long-term dependencies and GRU's computational efficiency. The proposed approach balances predictive accuracy and processing time by combining these models, making it well-suited for real-time temperature forecasting in wetland ecosystems. This architecture was optimized by tuning hyperparameters such as learning rate, batch size, and the number of hidden units to minimize prediction errors. The GRU-LSTM model has demonstrated higher accuracy than individual GRU and LSTM models and lower RMSE values than other prediction methods [17], [18].

To address the limitations of conventional models and improve prediction accuracy, this study proposes a novel workflow for air temperature prediction in wetland ecosystems by integrating the GRU-LSTM model. The proposed workflow

includes data collection, preprocessing, feature selection, model implementation, and performance evaluation using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) metrics. By combining the strengths of GRU and LSTM, this approach enhances the model's ability to capture complex temperature patterns while maintaining computational efficiency. This model is designed for practical implementation in environmental monitoring systems, enabling real-time temperature predictions that can assist policymakers and conservationists in making informed decisions. Additionally, the model's architecture is flexible and can be adapted to different geographical regions by retraining it with local temperature datasets. The GRU-LSTM model also demonstrates scalability, as it can be deployed on cloud-based platforms or edge computing devices, allowing for efficient processing of large-scale temperature data from multiple sensor networks. This model offers a viable solution for long-term wetland monitoring and disaster risk mitigation by balancing predictive accuracy and computational efficiency. Consequently, this study contributes to developing more accurate and efficient temperature prediction methods, supporting disaster risk mitigation and wetland conservation efforts.

II. MATERIAL

This study processed and analyzed data using a Lenovo Ideapad 330 15-ARR Laptop. The laptop has an AMD Ryzen 7 2700U processor, 8GB of RAM, a 1TB hard drive, and the Windows 10 operating system, providing sufficient data processing power. Table I presents the data sources used in this study.

TABLE I. SOURCE OF THE DATA

No.	Type of the Data	Source of the Data
1.	Data of Micro Air Temperature	Air temperature data from temperature sensors in the Liang Anggang Protected Forest Block 1
2.	Data of Macro Air Temperature	Daily Average Air Temperature Data from the Meteorology, Climatology and Geophysics Agency (BMKG) Online Data Syamsudin Noor Meteorological Station

TABLE II. DATASET

Date	Tavg
01-01-1996	25.60
01-02-1996	25.50
01-03-1996	26.50
01-04-1996	25.70
….	…
12-31-2021	27.20

This study used air temperature data from Liang Anggang Protected Forest Block 1, which included 42 records with 3-hour intervals and 42 daily average temperature values. The data was collected using IoT sensors and manual measurements to ensure comprehensive and reliable temperature monitoring. Additionally, macro-scale air temperature data was incorporated to provide a broader perspective on temperature variations. This dataset included daily average air temperature from the Meteorology, Climatology, and Geophysics Agency (BMKG) Online Data of the Syamsudin Noor Meteorological Station, covering 26 years (January 1, 1996 – December 31, 2021). Weekly and monthly average temperatures were also used to identify long-term trends and seasonal patterns. Table II presents an example of the collected air temperature data.

III. PROPOSED METHOD

The research flow is a series of steps required to conduct research systematically. Fig. 1 shows the research flow of this research.

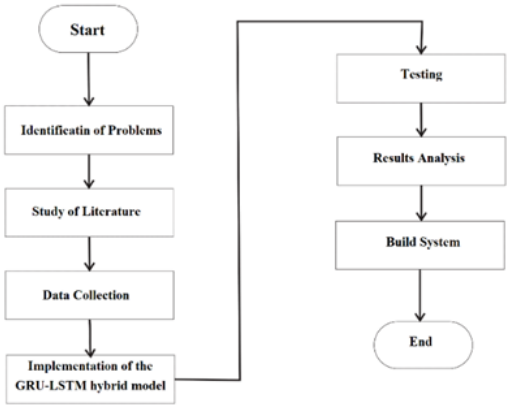


Fig. 1. Flow chart of research process.

The focus of this research shifted from data collection to testing to obtain model performance results that were used in the results analysis process, as shown in Fig. 2.

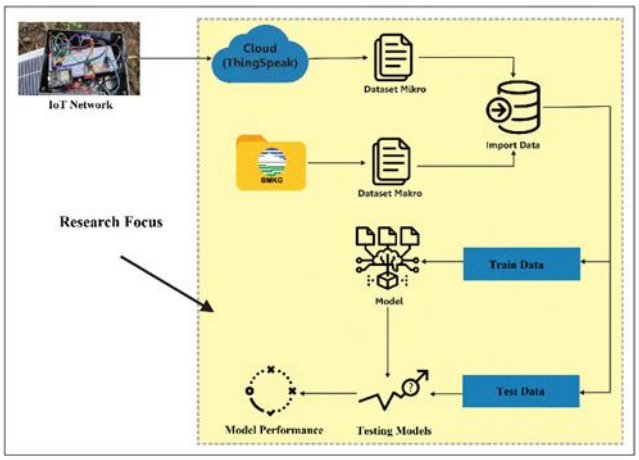


Fig. 2. Research focus.

A. Implementation of GRU-LSTM Hybrid Model

The GRU-LSTM hybrid model implementation process stages are shown in Fig. 3.

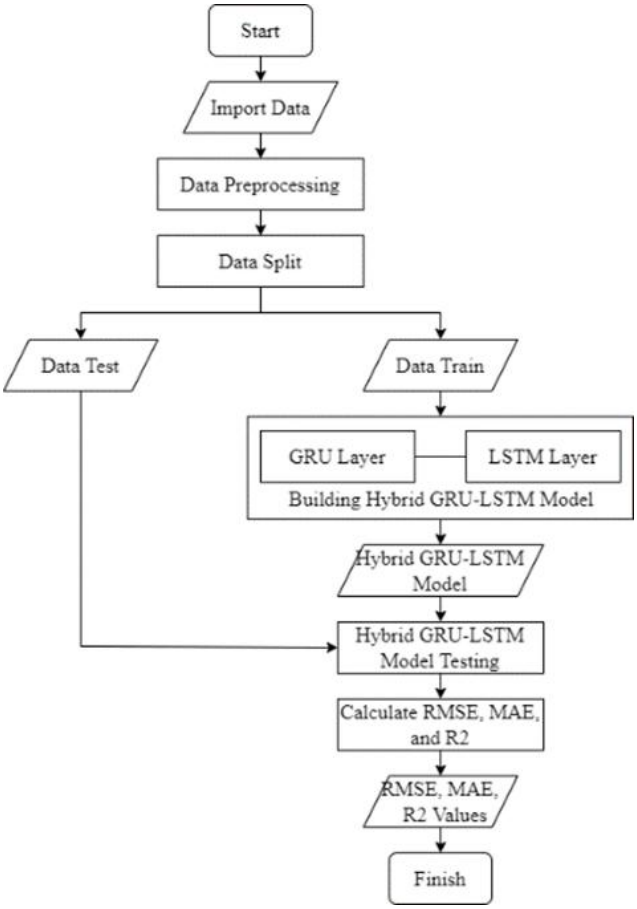


Fig. 3. Workflow model.

B. Data Preprocessing

The imported data was used as a data series during the data preprocessing stage, first by testing various input parameters and then using the Exponential (Weighted) Moving Average to fill in NaN values or empty values before being transformed using the Fast Fourier Transform.

1) Exponential (Weighted) Moving Average (EWMA): EWMA is a statistical method for smoothing datasets by decreasing data weight over time, and it is sensitive to process shifts [19]. The standard EWMA calculation is as follows:

$$S_t = \lambda e_t + (1-\lambda) S_{t-1} \quad (1)$$

Where S_t is the EWMA result, t is the time stamp, e_t is the average output, and λ is a constant ranging from 0 to 1 that controls the impact of historical data.

2) Fast Fourier Transform (FFT): Time series data is changed using FFT. FFT enhances the Discrete Fourier Transform (DFT) algorithm by leveraging periodicity and symmetry, significantly reducing the number of operations. FFT has the advantages of simplicity and speed. As a result, FFT is used to represent errors in the frequency domain to facilitate error identification [20], [21]. FFT can be calculated with [22]:

$$X_t(\mathrm{k}) = \sum_{t=1}^{\bar{N}} x_{ti} e^{\frac{-j2\pi}{\bar{N}} k(t-1)} \quad (2)$$

Where $X_i = [\, x_{1i}\, x_{2i} \cdots x_{\bar{N}i}]^T$, i $= 1, 2, \cdots,$ y, and $[\cdot]^T$ represents transpose from vector $[\cdot]$, $k = 0, 1, 2, \cdots, \bar{N} - 1$.

C. Data Sharing

For benchmarking, the datasets must be divided into training and test data. The model is trained using training data and evaluated using test data. This study conducted tests to assess various types of data-sharing ratios.

D. Model Establishing

The GRU-LSTM hybrid model was built with the GRU and LSTM layers from library keras.

1) Gated Recurrent Unit (GRU): GRU is a promising algorithm within the Recurrent Neural Network (RNN) [18]. GRU and LSTM cells operate similarly, but GRU cells use a hidden state that combines the forget gate and the input gate to form an update gate. GRU also combines hidden and cell states into a single state [22]. As a result, GRU is known as a simplified variant of LSTM [15]. The GRU equation is as follows [23]:

$$z_t = \sigma(W_z \cdot x_t + U_z \cdot h_{(t-1)} + b_z) \quad (3)$$

$$r_t = \sigma(W_r \cdot x_t + U_r \cdot h_{(t-1)} + b_r) \quad (4)$$

$$\tilde{h}_t = \tanh(W_h \cdot x_t + (r_t \circ U_h \cdot h_{(t-1)}) + b_h) \quad (5)$$

$$h_t = z_t \circ h_{(t-1)} + (1 - z_t) \circ \tilde{h}_t \quad (6)$$

Where z_t = update gate, σ = activation of the sigmoid function, W_z = weight of update gate, U_z = weight of hidden state, x_t = input value (input vector x in timestep t), $h_{(t-1)}$ = value of prior hidden state cell, b_z = bias of update gate, r_t = reset gate, W_r = weight of reset gate, U_r = weight of hidden state, b_r = bias of reset gate, $\tilde{h}_t$ = Output candidate from cell state vector, W_h = weight of cell state vector, U_h = weight of hidden state, b_h = bias of cell state vector, and h_t = cell state vector.

2) Long-Short Term Memory (LSTM): LSTM is a type of Recurrent Neural Network (RNN) that, when compared to conventional RNNs, allows the network to maintain long-term dependence between data at a specific time from many timesteps because LSTM uses special hidden blocks that remember input data for a long time [24], [25]. A typical LSTM unit consists of a memory cell, forget gate, input gate, and output gate, with the forget gate's purpose being to forget information in the cell state selectively, the input gate deciding what new information is stored in the cell state, and the output gate deciding what value to remove. The cell remembers the value over arbitrary time intervals, and three gates control the data flow into and out of the cell [26]. The LSTM application procedure is as follows [26]:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (7)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (8)$$

$$\tilde{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (9)$$

$$c_t = (f_t * c_{t-1} + i_t * \tilde{C}_t) \quad (10)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (11)$$

$$h_t = o_t * \tanh(c_t) \quad (12)$$

Where ft = forget gate, σ = sigmoid activation function, Wf = weight of forget gate, $ht-1$ = value of prior hidden state cell, xt = input value (input vector x in timestep t), bf = bias of forget gate, it = input gate, Wi = weight of input gate, bi = bias of input gate, $\tilde{C}t$ = candidate gate, Tanh = tanh activation function, Wc = weight of candidate gate, bc = bias of candidate gate, ct = cell gate, it = input gate, $\tilde{C}t$ = candidate gate, ft= forget gate, $ct-1$ = value of prior cell state, ot = output gate, Wo = weight of output gate, bo = bias of output gate, ht = hidden state, ct = cell gate. The LSTM architecture is shown in Fig. 4 [15]:

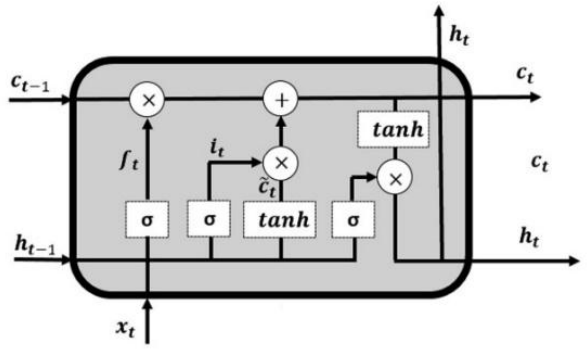


Fig. 4. LSTM architecture.

E. Model Testing

Model testing was carried out using test data to put the previously built GRU-LSTM hybrid model to the test.

F. Calculation of RMSE, MAE and R-Squared

The model test results was used to calculate performance using:

1) Root Mean Squared Error (RMSE): The root of the error value between the predicted and actual values is calculated to test the accuracy of prediction results [23]. The equation for calculating the RMSE value is as follow [27]:

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\tilde{y}_i - y_i)^2}{n}} \quad (13)$$

2) Mean Absolute Error (MAE): MAE measures the absolute difference between predicted and actual data, regardless of whether the difference is positive or negative [28]. The equation for calculating the MAE value is as follow [29]:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\tilde{y}_i - y_i| \quad (14)$$

3) R-Squared (R^2): R2 or called by coefficient of determination is a statistical measure used to investigate the relationship between actual and predicted data results. R2 can range between $-\infty$ up to 1; the closer the R^2 value is to 1, the better the model fits the dataset [18], [27]. The equation for calculating the value of R^2 is as follows [28]:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad (15)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\tilde{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{Y} - y_i)^2} \quad (16)$$

Where n = total of the data, i = (1, 2, 3, 4, 5, …, l), l is the entire data set, $\bar{Y}$ = the mean value of the actual data, y_i = actual value, dan $\tilde{y}_i$ = predicted value.

G. Testing Research Indicators Using Macro Data

In this test scheme, tests was performed to determine the performance of the Hybrid GRU-LSTM model by varying the number of input parameters, data sharing ratios, epochs, batch sizes, neuron sizes, and the arrangement of the model's architectural layers using daily, weekly, and monthly macro air temperature data.

H. Testing the Macro GRU-LSTM Hybrid Model Using Micro Data

In this test scheme, the best model for each use of the data obtained in the previous test scheme was used to predict air temperature on a microscale interval of 3 hours and daily.

I. Comparing the GRU-LSTM Hybrid Model Using GRU and LSTM

In this test scheme, the model performance results of the GRU-LSTM Hybrid Model, the GRU model, and the LSTM Model were compared.

J. Analysis Results

The Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2) were used to calculate the results of model testing. The lower the RMSE and MAE values obtained, the better the performance of the tested model, and the closer to one the R2 value, the better the resulting model. All models' training times were compared to determine which model had the fastest performance.

IV. Results

The GRU-LSTM Hybrid model was created on the Google Collab platform using the Python language and various Python libraries, such as time, numpy, pandas, sklearn, TensorFlow, and Keras. Fig. 5 shows the architecture of the GRU-LSTM Hybrid model [17].

The GRU model was created using 2 Layer GRU and 2 Layer Dense neurons with a size of 240. ReLU activation was used in the first GRU layer, ELU activation in the second GRU layer, and ReLU activation in the two Dense layers. The LSTM model was created using two layers of LSTM and two layers of Dense with 240 neuron sizes. The first LSTM layer was activated with ReLU, the second GRU layer with ELU, and the two Dense layers with ReLU. Using the add() function from keras.layers, the GRU and LSTM models were combined into a parallel GRU-LSTM model. The model now includes 3 Layer Dense with ELU activation and 240 neuron size, as well as 1 Layer Dense with Linear activation and 240 neuron size. Finally, as the model's output, a Dense Layer with one neuron size was added. The Adam Optimizer was used to build the model, with a learning rate of 0.00015, an L2 Regularizer of 10-4, 20 epochs, and a batch size of 64. This GRU-LSTM hybrid model was used to train each test scheme.

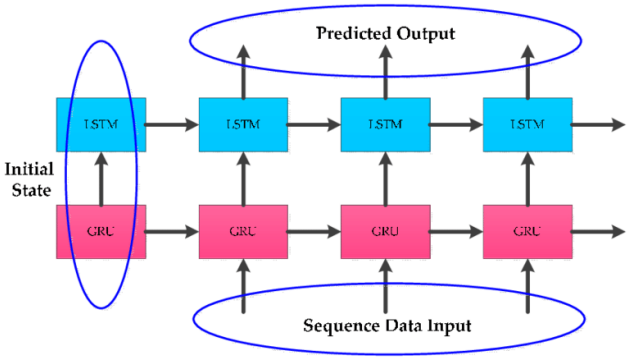


Fig. 5. Architecture of GRU-LSTM hybrid model.

A. Testing Data Preprocessing Techniques

This test was carried out by comparing the use of Min-Max Scaling with the Fast Fourier Transform to 7 NaN Handling techniques. Table III shows the results of testing each data preprocessing scheme using Min-Max Scaling:

TABLE III. COMPARISON OF NAN TECHNIQUES USING MIN-MAX SCALING

Missing Value Techniques	RMSE	MAE	R^2	Duration (s)
Mean	0.6844244536	0.4679480707	-0.1507817381	11.535
Median	0.6739228724	0.4521220334	-0.1240336428	12.755
Last Observation Carried Forward	0.7017514004	0.5292969394	-0.0704251303	11.038
Exponential (Weighted) Moving Average	0.6816600892	0.4810926004	-0.2355169779	11.272
Interpolation Linear	0.7437746843	0.5975289210	0.0155255931	12.202
Interpolation Spline	0.7421904651	0.5940747193	0.0197149326	10.993
Interpolation Time	0.7461391790	0.5987105870	0.0092562629	12.290

Based on this test, it is clear that the various preprocessing stages impact the model's performance. Four of the seven techniques tested with Min-Max Scaling had a negative R^2 value, indicating that the model did not adequately fit the data. This study obtained the best preprocessing stages using the Exponential (Weighted) Moving Average for NaN handling and the Fast Fourier Transform for data transformation, as shown in Table IV.

TABLE IV. COMPARISON OF NAN HANDLING TECHNIQUES USING THE FAST-FOURIER TRANSFORM

Missing Value Techniques	RMSE	MAE	R^2	Duration (s)
Mean	0.1290396570	0.1057177850	0.9590938765	11.822
Median	0.1270965006	0.1043401824	0.9600215361	11.272
Last Observation Carried Forward	0.1391963981	0.1151423968	0.9578841359	12.311
Exponential (Weighted) Moving Average	0.1161321395	0.0949022568	0.9641393763	12.104
Interpolation Linear	0.1638662900	0.1452299168	0.9522139770	11.063
Interpolation Spline	0.1664941613	0.1479380850	0.9506690292	14.679
Interpolation Time	0.1645815491	0.1457774991	0.9517959046	11.428

B. Testing the Total of Data

Table V shows the test results obtained from testing the amount of data using daily macro air temperature data.

TABLE V. RESULTS OF DAILY MACRO DATA TESTING BASED ON TOTAL DATA

Total of the Data	RMSE	MAE	R^2	Duration (s)
50	0.1851642148	0.1498968825	0.9340657953	9.079
100	0.1523455190	0.1254774662	0.9623533541	10.857
200	0.1265524442	0.1079908816	0.9674943384	11.948
300	0.1174715265	0.0959396310	0.9633074235	13.293
500	0.1700496630	0.1323210940	0.9593256173	18.942
1000	0.1802263938	0.1456180500	0.9560157537	22.164
2000	0.1680635190	0.1345518316	0.9591325611	29.965
4000	0.1667981339	0.1313218270	0.9629232147	50.314
8000	0.1735634802	0.1304684863	0.9647927358	86.563

When testing the amount of data with daily average macro air temperature data, 300 data points produced the lowest RMSE and MAE values, 200 data points produced the highest R2 values, and 50 data points produced the shortest model training duration. The test scheme applied 300 data points to obtain the best model performance on this research indicator. The test results obtained from testing the amount of data using weekly macro air temperature data are shown in Table VI.

TABLE VI. RESULTS OF WEEKLY MACRO DATA TESTING BASED ON TOTAL DATA

Total of the Data	RMSE	MAE	R^2	Duration (s)
50	0.0623367201	0.0511048617	0.9455028859	10.697
100	0.1371391608	0.1088339692	0.8295841737	13.122

Total of the Data	RMSE	MAE	R^2	Duration (s)
200	0.1554007142	0.1264640826	0.9198697611	10.850
300	0.1634344265	0.1335772436	0.9185153251	13.031
500	0.1330690852	0.1018681449	0.9483418575	14.694
1000	0.1250917579	0.0985749315	0.9648619295	26.889

When testing the amount of data with weekly average macro air temperature data, the lowest RMSE and MAE values were obtained using 50 data, the highest R^2 values were obtained using 1000 data, and the fastest model training duration was obtained using 50 data. The test scheme applied 50 data points to get the best model performance on this research indicator. Table VII shows the results obtained in testing the amount of data using monthly macro air temperature data.

TABLE VII. RESULTS OF MONTHLY MACRO DATA TESTING BASED ON TOTAL DATA

Total of the Data	RMSE	MAE	R^2	Duration (s)
50	0.1712211232	0.1550369370	0.7706810377	10.477
100	0.1649493670	0.1474490374	0.8194394220	9.683
200	0.1452670353	0.1231163900	0.9068988647	11.418
300	0.1420492412	0.1233844163	0.9335440337	13.327

Testing with monthly average macro air temperature data showed that 300 data points yielded the lowest RMSE and highest R^2, 200 data points had the lowest MAE, and 50 data points resulted in the shortest training time. The test scheme applied 300 data points to achieve the best model performance, as shown in Fig. 6.

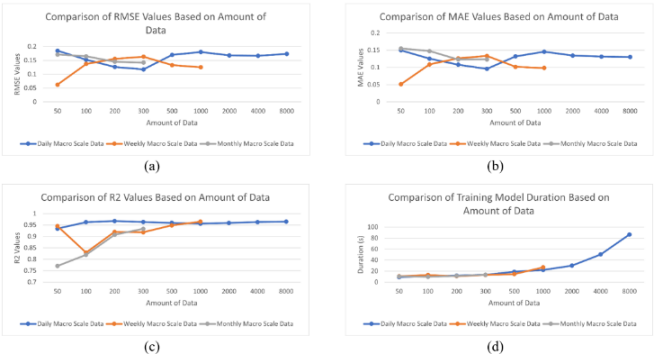


Fig. 6. Analysis of model performance results based on total data (a) RMSE Value, (b) MAE Value, (c) R2 Value, (d) Model training duration.

Increasing the data amount did not continually improve performance; data variations influenced results. For instance, with daily macro temperature data, 300 data points had a lower error rate than 200 or 500, whereas with weekly data, 300 data points had the highest error rate.

C. Parameter Input Testing

Table VIII shows the test results obtained in the input parameter test using daily macro air temperature data.

TABLE VIII. RESULTS OF DAILY MACRO DATA TESTING BASED ON PARAMETER INPUT

Parameter Input	RMSE	MAE	R^2	Duration (s)
1 day	0.4947327263	0.3840819742	0.3491899320	7.326
2 days	0.2478204263	0.1922136994	0.8366996247	9.103
3 days	0.1838271449	0.1453921775	0.9101470651	6.824
4 days	0.1541403522	0.1312984414	0.9368249379	6.975
5 days	0.1393977268	0.1110161938	0.9483316768	7.318
6 days	0.1163984780	0.0952625401	0.9639747016	6.925
7 days	0.1002650901	0.0804510812	0.9732691678	9.645
8 days	0.0902920465	0.0726037929	0.9783223612	7.895
9 days	0.0782327439	0.0637689138	0.9837261575	6.967
10 days	0.0744849001	0.0633215273	0.9852480489	7.635
11 days	0.0731881517	0.0594071249	0.9857572273	7.047
12 days	0.0765548540	0.0585217093	0.9844167352	7.538

The lowest RMSE value was in the use of 11 daily average data to predict the next day's average, the lowest MAE value was in the use of 12 daily average data to predict the next day's average, the highest R^2 value was in the use of 11 daily average data to predict the next day's average, and the fastest model training duration was in the use of 3 daily average data to predict the next day's average. To achieve the best model performance on this research indicator, the test scheme applied 11 Input Parameters that use 11 daily average data to predict the next day's average. The test results obtained in the input parameter test using weekly macro air temperature data is shown in Table IX.

TABLE IX. RESULTS OF WEEKLY MACRO DATA TESTING BASED ON PARAMETER INPUT

Parameter Input	RMSE	MAE	R^2	Time (s)
1 week	0.3119153610	0.2365223692	0.7032010388	6.849
2 weeks	0.1542141168	0.1180309918	0.9274500509	7.787
3 weeks	0.1909112509	0.1624875251	0.8888135417	7.846
4 weeks	0.1851261596	0.1540472429	0.8954499045	7.708
5 weeks	0.1701543518	0.1398497052	0.9116767640	9.665
6 weeks	0.1594423340	0.1269611427	0.9224474403	8.832
7 weeks	0.1582014176	0.1261986688	0.9236499033	7.102
8 weeks	0.1462549959	0.1217502412	0.9347455294	9.227
9 weeks	0.1272247907	0.1041676132	0.9506221313	11.002
10 weeks	0.1145790877	0.0905048298	0.9599502721	8.880
11 weeks	0.1155339656	0.0965910716	0.9592799586	9.063
12 weeks	0.0976961789	0.0799839595	0.9708831757	10.617

When the input parameters were tested using weekly average macro air temperature data, the lowest RMSE and MAE values

were obtained using 12 weekly average data to predict the following week's average, the highest R^2 value was obtained using 12 weekly average data to predict the following week's average, and the shortest model training duration was obtained by using one weekly average data to predict the following week's average. To achieve the best model performance on this research indicator, the test scheme applied 12 Input Parameters that used 12 weekly average data to predict the following week's average. The test results obtained in the input parameter test using monthly macro air temperature data are shown in Table X.

TABLE X. RESULTS OF MONTHLY MACRO DATA TESTING BASED ON INPUT DATA

Parameter Input	RMSE	MAE	R^2	Duration (s)
1 month	0.2520620441	0.2141007492	0.7907473106	7.227
2 months	0.1328624270	0.1141378974	0.9418619443	7.976
3 months	0.1890981064	0.1546116300	0.8822311320	10.817
4 months	0.1929430235	0.1650484307	0.8773932727	9.214
5 months	0.1738967881	0.1457596050	0.9004046042	9.102
6 months	0.1413953839	0.1238171764	0.9341544236	8.743
7 months	0.1078759913	0.0922371984	0.9616729194	11.342
8 months	0.0801424696	0.0654195946	0.9788465317	13.374
9 months	0.0738925268	0.0626033554	0.9820172057	16.962
10 months	0.0588448555	0.0470539990	0.9885955816	13.186
11 months	0.0712976266	0.0640248735	0.9832580405	12.678
12 months	0.0568075286	0.0465924708	0.9893715990	14.567

When the input parameters were tested using monthly average macro air temperature data, the lowest RMSE and MAE values were obtained using 12 monthly average data to predict the next month's average, the highest R^2 value was obtained using 12 monthly average data to predict the next month's average, and the shortest model training duration was obtained using one monthly average data to predict the next month's average. To achieve the best model performance on this research indicator, the test scheme applied 12 Input Parameters that used 12 monthly average data to predict the next month's average. The analysis of the effect of the number of input parameters on the performance of the Hybrid GRU-LSTM model is shown in Fig. 7.

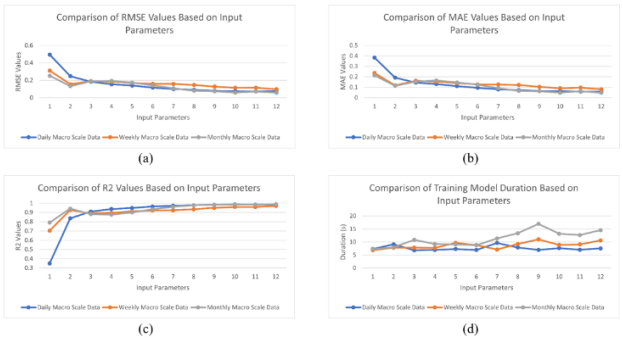


Fig. 7. Analysis of model performance results based on input parameters (a) RMSE Value, (b) MAE Value, (c) R^2 Value, (d) Model training duration.

Increasing the number of input parameters affected the model's performance. Each data use almost always resulted in a non-significant decrease in error rate from 3 input parameters to 12 input parameters. Increasing the number of input parameters from one to two significantly improved model performance, so using a GRU-LSTM hybrid model with one input parameter was not recommended.

D. Testing of Data Sharing Ratio

Table XI shows the test results obtained when testing the data sharing ratio using daily macro air temperature data.

TABLE XI. RESULTS OF DAILY MACRO DATA TESTING BASED ON DATA SHARING RATIO

Data Sharing Ratio	RMSE	MAE	R^2	Duration (s)
50 : 50	0.1746709404	0.1380718406	0.9607895299	10.457
60 : 40	0.1704924270	0.1366307721	0.9618051593	13.888
70 : 30	0.1854317031	0.1515861214	0.9537578721	13.381
80 : 20	0.1651222128	0.1313805477	0.9569454454	11.056
90 : 10	0.1164872255	0.0956705397	0.9639197460	12.978

Testing the data-sharing ratio using daily average macro air temperature data revealed that a 90:10 ratio provided the best model performance, with the lowest RMSE and MAE values and the highest R^2 value. In contrast, a 50:50 ratio significantly reduced the training time but at the cost of lower model accuracy. These findings indicate that a more extensive training dataset (90%) enhances model learning, leading to better predictive performance. The test scheme applied a 90:10 data-sharing ratio to optimize performance based on these results. Further analysis of the impact of data-sharing ratios using weekly average macro air temperature data is presented in Table XII.

TABLE XII. RESULTS OF WEEKLY MACRO DATA TESTING BASED ON DATA SHARING RATIO

Data Sharing Ratio	RMSE	MAE	R^2	Duration (s)
50 : 50	0.1310978587	0.1054439719	0.9654619062	11.361
60 : 40	0.1430146436	0.1128988945	0.9627798949	11.728
70 : 30	0.1393393249	0.1107925050	0.9557609879	12.986
80 : 20	0.1304721654	0.1024591358	0.9551738898	12.774
90 : 10	0.1573456424	0.1253423013	0.9244736860	15.262

When testing the data sharing ratio with weekly average macro air temperature data, the 80:20 ratio had the lowest RMSE and MAE values, the 50:50 ratio had the highest R^2 value, and the 50:50 ratio had the shortest model training duration. Table XIII shows the results obtained when testing the data-sharing ratio using monthly macro air temperature data.

TABLE XIII. RESULTS OF MONTHLY MACRO DATA TESTING BASED ON DATA SHARING RATIO

Data Sharing Ratio	RMSE	MAE	R^2	Duration (s)
50 : 50	0.1763222753	0.1431057810	0.8977571251	11.401
60 : 40	0.1730447912	0.1381097616	0.8966686071	11.724
70 : 30	0.1717020795	0.1433728574	0.9059654873	15.100
80 : 20	0.1353351706	0.1142427518	0.9230716267	13.004
90 : 10	0.1533416137	0.1331341740	0.9225580644	10.967

When testing the data sharing ratio with monthly average macro air temperature data, the 80:20 ratio produced the lowest RMSE and MAE values, the 80:20 ratio produced the highest R^2, and the 90:10 ratio produced the shortest model training duration. To achieve the best model performance on this research indicator, the test scheme applied an 80:20 Data Sharing Ratio. Fig. 8 shows the analysis of the effect of the data-sharing ratio on the performance of the Hybrid GRU-LSTM model.

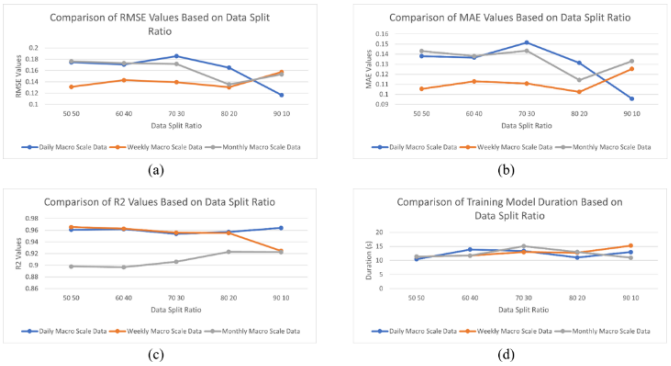


Fig. 8. Analysis of model performance results based on data sharing ratio (a) RMSE Value, (b) MAE Value, (c) R^2 Value, (d) Model training duration.

Changes in the data share ratio affected model performance, but data also affected it. Based on this test, the best data-sharing ratio for each data set was not found.

E. Epoch Testing

The epoch test results obtained using daily macro air temperature data are shown in Table XIV.

TABLE XIV. RESULTS OF MONTHLY MACRO DATA TESTING BASED ON EPOCH

Epochs	RMSE	MAE	R^2	Duration (s)
10	0.1412090605	0.1098501957	0.9469801969	10.081
20	0.1178582018	0.0964495148	0.9630654677	12.043
30	0.1159206586	0.0952576068	0.9642698645	17.625
40	0.1158310883	0.0951873015	0.9643250595	18.672
50	0.1162878559	0.0955324685	0.9640431441	18.745
60	0.1159376773	0.0948069858	0.9642593724	21.236
70	0.1173887879	0.0965781303	0.9633590926	22.936
80	0.1162496231	0.0956332868	0.9640667838	27.657
90	0.1163170192	0.0956855566	0.9640251069	27.304
100	0.1161146882	0.0951103535	0.9641501532	30.338

On this research indicator, the test scheme with 40 Epoch yielded the best model performance. The epoch test results obtained using weekly macro air temperature data is shown in Table XV:

TABLE XV. RESULTS OF WEEKLY MACRO DATA TESTING BASED ON EPOCH

Epochs	RMSE	MAE	R^2	Duration (s)
10	0.1625175407	0.1327365961	0.9194270372	11.416
20	0.1600507480	0.1281993821	0.9218544474	12.325
30	0.1612495635	0.1296542749	0.9206794082	17.127
40	0.1615903931	0.1300313808	0.9203437375	15.127
50	0.1609623787	0.1293785623	0.9209616959	28.846
60	0.1603825770	0.1283413292	0.9215300773	27.610
70	0.1610848606	0.1292554419	0.9208413641	28.765
80	0.1584860572	0.1260031276	0.9233749144	28.937
90	0.1589206752	0.1266394984	0.9229540786	24.613
100	0.1607210281	0.1285202697	0.9211985418	50.008

The test scheme with 80 Epoch yielded the best model performance on this research indicator. The epoch test results obtained using monthly macro air temperature data is shown in Table XVI.

TABLE XVI. RESULTS OF MONTHLY MACRO DATA TESTING BASED ON EPOCH

Epochs	RMSE	MAE	R^2	Duration (s)
10	0.1768219120	0.1485940880	0.8970258287	14.240
20	0.1443285145	0.1265315739	0.9313942646	11.598
30	0.1401524357	0.1229330155	0.9353069778	17.838
40	0.1396633099	0.1222866899	0.9357577414	17.241
50	0.1388041207	0.1213296601	0.9365457289	28.092
60	0.1409545112	0.1233228180	0.9345643983	20.015
70	0.1397881914	0.1223504527	0.9356428042	28.075
80	0.1404582682	0.1230280480	0.9350243311	28.311
90	0.1391538159	0.1214379346	0.9362256001	27.242
100	0.1401636231	0.1227904198	0.9352966494	47.246

The test scheme with 50 epochs yielded the best model performance on this research indicator. Fig. 9 analyzes the effect

of the number of epochs on the performance of the Hybrid GRU-LSTM model.

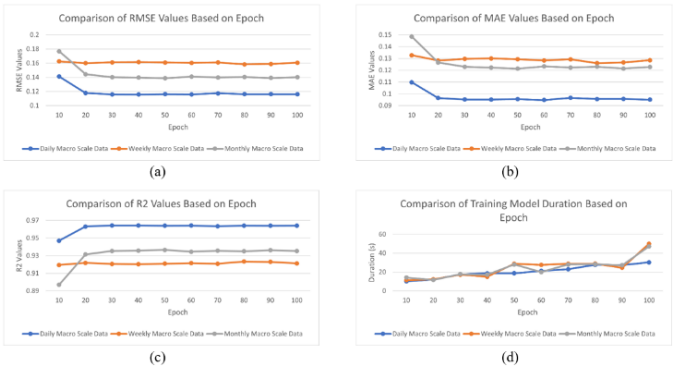


Fig. 9. Analysis of model performance results based on the number of Epoch (a) RMSE Value, (b) MAE Value, (c) R^2 Value, (d) Model training duration.

The number of epochs did not significantly affect the model's performance; only when the number of epochs was increased from 10 to 20, the error rate decreased by 0.02 - 0.04, and the MAE value decreased by 0.03 - 0.02. Other tests tended to have the same error rate, with the RMSE value being 0.12 for daily macro data, 0.16 for weekly macro data, and 0.14 for monthly macro data.

F. Batch Size Testing

Table XVII presents the results of the batch size test using daily macro air temperature data, comparing model performance across different batch sizes. The table shows how batch size affects error values (RMSE, MAE, and R^2) and training stability. Smaller batch sizes may improve generalization but require more iterations, while larger batch sizes can speed up training but may lead to overfitting. This analysis helps determine the optimal batch size for the Hybrid GRU-LSTM model.

TABLE XVII. RESULTS OF DAILY DATA TESTING BASED ON BATCH SIZE

Batch Size	RMSE	MAE	R^2	Duration (s)
32	0.1165713735	0.0960287397	0.9638675999	16.838
64	0.1161702623	0.0956351955	0.9641158285	12.916
128	0.1182457040	0.0959229463	0.9628221966	10.385

The test scheme with 64 Batch Sizes yielded the best model performance on this research indicator. The results of the batch size test using weekly macro air temperature data is shown in Table XVIII:

TABLE XVIII. RESULTS OF WEEKLY MACRO DATA TESTING BASED ON BATCH SIZE

Batch Size	RMSE	MAE	R^2	Duration (s)
32	0.1599404943	0.1278730529	0.9219620741	15.048
64	0.1579052215	0.1257131268	0.9239355319	12.542
128	0.1635250788	0.1345633975	0.9184249057	11.566

The test scheme with 64 Batch Sizes yielded the best model performance on this research indicator. The results of the batch size test using monthly macro air temperature data are shown in Table XIX:

TABLE XIX. RESULTS OF MONTHLY DATA TESTING BASED ON BATCH SIZE

Batch Size	RMSE	MAE	R^2	Duration (s)
32	0.1405752504	0.1233503645	0.9349160546	12.019
64	0.1423798040	0.1245906945	0.9332343744	13.026
128	0.1836717030	0.1569849137	0.8888931982	11.357

The test scheme with 32 batch sizes yielded the best model performance for this research indicator. Fig. 10 shows the analysis of the effect of batch size on the performance of the Hybrid GRU-LSTM model.

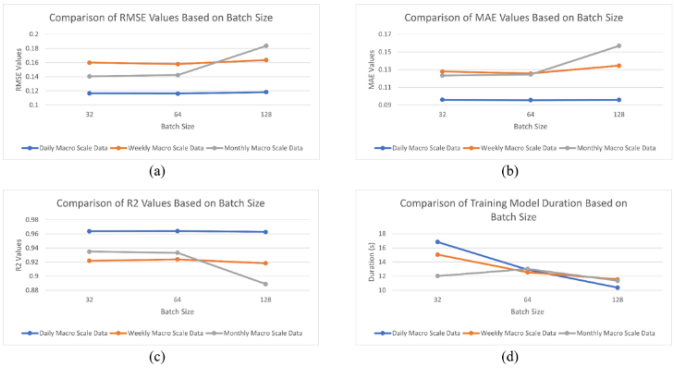


Fig. 10. Analysis of model performance results based on total batch size (a) RMSE Value, (b) MAE Value, (c) R^2 Value, (d) Model training duration.

Increasing the number of batch sizes did not affect the model's performance. Except for using 128 batch sizes, the model performance was consistent across tests, with the RMSE value at 0.12 for daily macro data, 0.16 for weekly macro data, and 0.14 for monthly macro data.

G. Neuron Size Testing

The neuron size test results obtained using daily macro air temperature data are shown in Table XX.

TABLE XX. RESULTS OF DAILY MACRO DATA TESTING BASED ON NEURON SIZE

Neuron Size	RMSE	MAE	R^2	Duration (s)
10	0.1474003564	0.1178712074	0.9422289780	8.242
20	0.1516744342	0.1224722308	0.9388301028	11.471
30	0.1472699419	0.1163316259	0.9423311602	9.093
60	0.1228602713	0.0997256246	0.9598638304	10.605
120	0.1157847143	0.0940748270	0.9643536193	10.974
240	0.1165024016	0.0957993578	0.9639103442	12.659

Based on this research indicator, the test scheme with a neuron size of 120 produced the best model performance. The

results of the neuron size test using weekly macro air temperature data are presented in Table XXI.

TABLE XXI. RESULTS OF WEEKLY MACRO DATA TESTING BASED ON NEURON SIZE

Neuron Size	RMSE	MAE	R^2	Duration (s)
10	0.1682801592	0.1389587059	0.9136117495	9.663
20	0.1677301907	0.1378873890	0.9141754902	10.070
30	0.1620301639	0.1322839767	0.9199095760	8.496
60	0.1605895360	0.1284312306	0.9213274301	9.727
120	0.1627621243	0.1322071073	0.9191843353	10.361
240	0.1603681788	0.1286783949	0.9215441658	11.524

The test scheme with 240 Neuron Size produced the best model performance on this research indicator. The neuron size test results obtained using monthly macro air temperature data are shown in Table XXII.

TABLE XXII. RESULTS OF MONTHLY MACRO DATA TESTING BASED ON NEURON SIZE

Neuron Size	RMSE	MAE	R^2	Duration (s)
10	0.1822280375	0.1554668319	0.8906329402	9.285
20	0.1878851786	0.1603905481	0.8837370911	10.220
30	0.1896031054	0.1620292374	0.8816012730	10.887
60	0.1441418736	0.1270096016	0.9315715872	9.028
120	0.1481143123	0.1263619519	0.9277479467	11.299
240	0.1413978662	0.1237826017	0.9341521117	12.373

The test scheme with 240 neuron sizes produced the best model performance for this research indicator. Fig. 11 examines the effect of neuron size on the performance of the Hybrid GRU-LSTM model. Increasing the number of neuron sizes affected the model's performance. Increasing the number of neuron sizes from 30 to 60 significantly improved model performance, and the error rate decreased by around 0.03 - 0.05 in tests using daily macro data and monthly macro data.

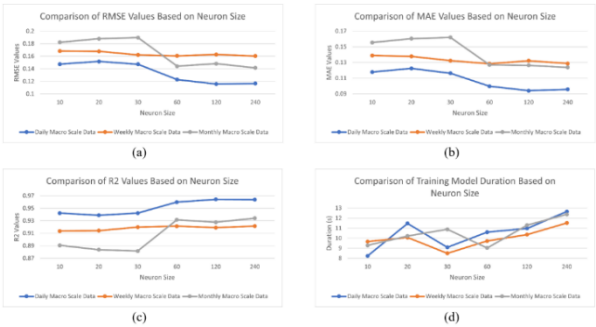


Fig. 11. Analysis of model performance based on number of neuron sizes (a) RMSE value, (b) MAE value, (c) R^2 value, (d) Training duration.

H. Layer Testing

Layer testing was conducted by modifying the model's layer arrangement. The scheme used to test the layer is shown in Table XXIII.

TABLE XXIII. LAYER TESTING SCHEME

Scheme	Description
Scheme 1	In the GRU-LSTM parallel model, each has 1 GRU/LSTM Layer and 1 Dense Layer
Scheme 2	In the GRU-LSTM parallel model, each has 2 GRU/LSTM Layer and 1 Dense Layer
Scheme 3	In the GRU-LSTM parallel model, each has 3 GRU/LSTM Layer and 1 Dense Layer
Scheme 4	In the parallel GRU-LSTM model, each has 1 GRU/LSTM Layer and 2 Dense Layers
Scheme 5	In the parallel GRU-LSTM model, each has 2 GRU/LSTM Layer and 2 Dense Layers
Scheme 6	In the parallel GRU-LSTM model, each has 3 GRU/LSTM Layer and 2 Dense Layers
Scheme 7	In the parallel GRU-LSTM model, each has 1 GRU/LSTM Layer and 3 Dense Layers
Scheme 8	In the parallel GRU-LSTM model, each has 2 GRU/LSTM Layer and 3 Dense Layers
Scheme 9	In the parallel GRU-LSTM model, each has 3 GRU/LSTM Layer and 3 Dense Layers
Scheme 10	In the parallel GRU-LSTM model, each has 2 GRU/LSTM Layers, 1 Dropout Layer, and 2 Dense Layers
Scheme 11	In the parallel GRU-LSTM model, each has 1 Layer GRU/LSTM, 1 Layer Dropout, 1 Layer GRU/LSTM, 1 Layer Dropout, and 2 Layers Dense
Scheme 12	Using 1 Layer Dense after implementing the GRU-LSTM parallel model
Scheme 13	Using 2 Layer Dense after implementing the GRU-LSTM parallel model
Scheme 14	Using 3 Layer Dense after implementing the GRU-LSTM parallel model
Scheme 15	Using 4 Layer Dense after implementing the GRU-LSTM parallel model

The layer test results obtained using daily macro air temperature data are shown in Table XXIV:

TABLE XXIV. RESULTS OF DAILY MACRO DATA TESTING BASED ON LAYER

Layers	RMSE	MAE	R^2	Duration (s)
Scheme 1	0.1207798570	0.0960987393	0.9612115875	9.724
Scheme 2	0.1170422441	0.0950615363	0.9635751088	13.741
Scheme 3	0.1164223853	0.0956228881	0.9639599015	16.659
Scheme 4	0.1193676096	0.0970861849	0.9621133700	11.356
Scheme 5	0.1171691442	0.0944435854	0.9634960804	12.952
Scheme 6	0.1166691458	0.0958401778	0.9638069636	21.968
Scheme 7	0.1168505315	0.0937090292	0.9636943374	12.021
Scheme 8	0.1156994266	0.0948360588	0.9644061147	13.936

Layers	RMSE	MAE	R^2	Duration (s)
Scheme 9	0.1160412787	0.0953250178	0.9641954685	23.327
Scheme 10	0.1205027634	0.0988276975	0.9613893604	14.647
Scheme 11	0.1491179453	0.1198267348	0.9408747752	13.310
Scheme 12	0.1152844223	0.0943347440	0.9646610014	13.840
Scheme 13	0.1160242160	0.0946575514	0.9642059972	12.546
Scheme 14	0.1163461744	0.0952354286	0.9640070702	13.246
Scheme 15	0.1166766606	0.0955005634	0.9638023009	15.742

After implementing the GRU-LSTM parallel model, test scheme 12 achieved the best model performance on this research indicator. The layer test results obtained using weekly macro air temperature data are shown in Table XXV:

TABLE XXV. RESULTS OF WEEKLY MACRO DATA TESTING BASED ON LAYER

Layers	RMSE	MAE	R^2	Duration (s)
Scheme 1	0.1585801594	0.1294481339	0.9232838940	8.031
Scheme 2	0.1592887200	0.1307524235	0.9225968037	12.092
Scheme 3	0.1570656604	0.1245989907	0.9247422310	20.229
Scheme 4	0.1617638061	0.1308678964	0.9201726773	9.351
Scheme 5	0.1632904734	0.1330755769	0.9186588054	11.745
Scheme 6	0.1612453088	0.1299866103	0.9206835940	17.198
Scheme 7	0.1621873110	0.1322764476	0.9197541471	9.056
Scheme 8	0.1642158546	0.1350723155	0.9177342578	15.201
Scheme 9	0.1632028674	0.1321671039	0.9187460617	21.451
Scheme 10	0.1612673131	0.1325025413	0.9206619447	13.687
Scheme 11	0.1683629685	0.1372377593	0.9135267067	19.908
Scheme 12	0.1573133076	0.1271241494	0.9245047244	12.781
Scheme 13	0.1586990147	0.1287871174	0.9231688540	13.876
Scheme 14	0.1563978327	0.1247707114	0.9253808476	11.979
Scheme 15	0.1619698941	0.1309035883	0.9199691468	11.850

The best model performance on this research indicator was obtained using Test Scheme 14 after implementing the parallel GRU-LSTM model. The layer test results obtained using monthly macro air temperature data are shown in Table XXVI.

TABLE XXVI. RESULTS OF MONTHLY MACRO DATA TESTING BASED ON LAYER

Layers	RMSE	MAE	R^2	Duration (s)
Scheme 1	0.1624325311	0.1378718658	0.9131035240	7.353
Scheme 2	0.1377480782	0.1186656237	0.9375075951	11.547
Scheme 3	0.1401070555	0.1226267907	0.9353488651	19.437
Scheme 4	0.1708900003	0.1450752292	0.9038189659	8.928
Scheme 5	0.1455686171	0.1262893909	0.9302102483	12.174
Scheme 6	0.1410352850	0.1234592980	0.9344893812	16.611
Scheme 7	0.1519643950	0.1326691130	0.9239428879	11.777
Scheme 8	0.1394894457	0.1219289526	0.9359175898	13.114
Scheme 9	0.1407462996	0.1233650677	0.9347575725	18.069
Scheme 10	0.1684723507	0.1447563020	0.9065211390	14.408
Scheme 11	0.1872637291	0.1604907882	0.8845049222	12.991
Scheme 12	0.1590729268	0.1363708035	0.9166609234	11.487
Scheme 13	0.1464591550	0.1272421046	0.9293537377	14.149
Scheme 14	0.1395472120	0.1213907857	0.9358645023	13.124
Scheme 15	0.1439520390	0.1267799049	0.9317517087	14.256

Test Scheme 2 achieved the best model performance on this research indicator using a parallel GRU-LSTM model with 2 Layers of GRU/LSTM and one dense layer. Fig. 12 shows the analysis of the effect of layers on the performance of the Hybrid GRU-LSTM model.

Fig. 12. Analysis of model performance results based on layer testing scheme (a) RMSE Value, (b) MAE Value, (c) R^2 Value, (d) Model Training Duration.

Layer changes had no discernible effect on model performance. Increasing or decreasing the number of layers did not affect the model error rate; the model error rate only increased significantly when the Dropout Layer was added in Schemes 10 and 11. As a result, using the GRU-LSTM hybrid model with a dropout layer in testing with 300 data points was not recommended.

I. Testing the GRU-LSTM Hybrid Model for Macro Data for Micro Data

This test applied a model trained on macro air temperature data to test micro air temperature data. The model used is as follows:

1) Model 1: This GRU-LSTM hybrid model was built with Daily Macro Data, incorporating 11 daily averages to predict the average of the following days.

2) Model 2: The GRU-LSTM LSTM hybrid model was built using Weekly Macro Data and 50 weekly average data points.

3) Model 3: This GRU-LSTM hybrid model was built using monthly macro data by combining 12 monthly averages

to predict the next month's average.

Table XXVII shows the test results obtained in the GRU-LSTM Hybrid Model t2st, which was built using macro data to predict micro data every 3 hours:

TABLE XXVII. RESULTS OF MACRO DATA MODEL USING MICRO DATA PER 3 HOURS

GRU-LSTM Hybrid Models	RMSE	MAE	R^2
Model 1	1.0422397749	0.9398446809	0.9536844965
Model 2	1.8543152148	1.5887022426	0.8533918118
Model 3	0.8673054280	0.7361485392	0.9679273214

A test that used Model 3 to predict micro air temperature every 3 hours had the best model performance in this test. The temperature changes during the day were significant, reaching 3 - 14 °C, so micro air temperature data with 3-hour intervals had several outlier data. These significant temperature changes affected the model's performance. As a result, using models trained on daily, weekly, and monthly macro data still resulted in a relatively high error rate. The test results obtained in the GRU-LSTM Hybrid Model test, which was built using macro data to predict daily microdata, are shown in Table XXVIII:

TABLE XXVIII. TESTING RESULTS OF THE MACRO DATA MODEL USING DAILY MICRO DATA

GRU-LSTM Hybrid Models	RMSE	MAE	R^2
Model 1	0.2307456696	0.1950775580	0.9812115284
Model 2	0.3606277068	0.2694757832	0.9541074158
Model 3	0.2270860682	0.1908012237	0.9818027687

The best model performance in this test was in a test that used Model 3 to predict micro air temperature every 3 hours. The graph of model performance results is shown in Fig. 13.

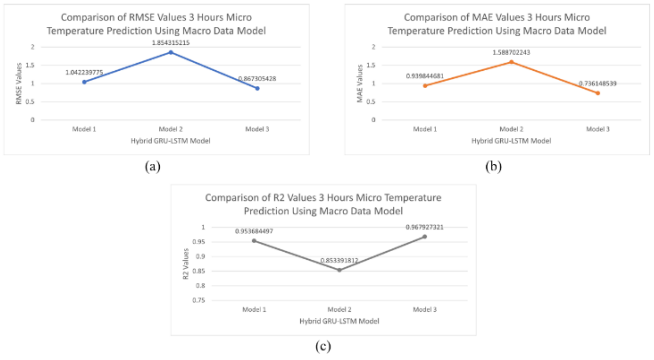


Fig. 13. Analysis of macro data model performance results for prediction of microtemperature per 3 hours (a) RMSE value, (b) MAE value, (c) R^2 value.

The model's performance was tested in predicting daily micro air temperature using a model trained using daily, weekly, and monthly macro data. This model showed a lower error rate than testing using data with 3-hour intervals.

J. GRU, LSTM and Hybrid GRU-LSTM Model Testing

This evaluation compares the performance of the GRU, LSTM, and Hybrid GRU-LSTM models in handling sequential data. Fig. 14 presents the results for each test scheme, highlighting their accuracy and efficiency. This provides insights into the strengths and suitability of each model for forecasting tasks.

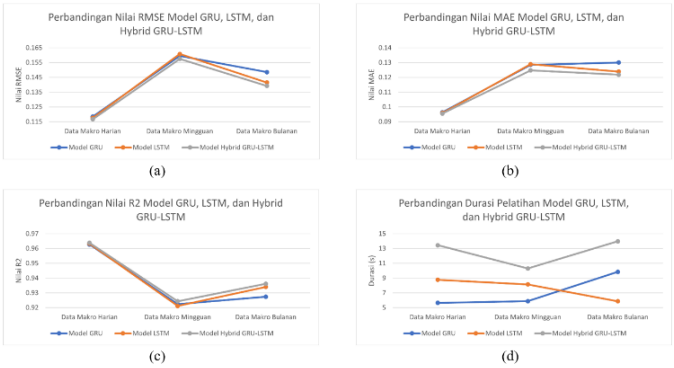


Fig. 14. Analysis of performance model results of GRU Model, LSTM Model, and GRU-LSTM Hybrid Model (a) RMSE Value, (b) MAE Value, (c) R^2 Value, (d) Model Training Duration.

Table XXIX compares the Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), and Hybrid GRU-LSTM models. This table highlights their predictive accuracy, efficiency, and overall performance in handling sequential data.

In every use of the data tested, the GRU-LSTM Hybrid Model outperformed the GRU-only and LSTM-only Models. Fig. 15 shows a performance comparison analysis of the GRU, LSTM, and Hybrid GRU-LSTM Modes. The prediction results of the GRU-only model, the LSTM-only model, and the GRU-LSTM Hybrid model on daily, weekly, and monthly macro data did not differ significantly. The model performance results showed significant results. If larger quantities were predicted, this has not been investigated further.

TABLE XXIX. COMPARISON RESULTS OF GRU, LSTM, AND GRU-LSTM HYBRID MODELS

Data Using	Model Using	RMSE	MAE	R^2	Duration (s)
Daily Macro Data	GRU model	0.1184584205	0.0964402422	0.9626883153	5.650
	LSTM models	0.1176304218	0.0959534831	0.9632080936	8.774
	GRU-LSTM Hybrid models	0.1166597038	0.0955331357	0.9638128215	13.437
Weekly Macro Data	GRU model	0.1594879534	0.1285021928	0.9224030555	5.875
	LSTM models	0.1607783534	0.1289006750	0.9211423186	8.146
	GRU-LSTM Hybrid models	0.1575065981	0.1247597571	0.9243190886	10.293

Data Using	Model Using	RMSE	MAE	R^2	Duration (s)
Monthly Macro Data	GRU model	0.1485245039	0.1300583339	0.9273471991	9.837
	LSTM models	0.1415222437	0.1239307644	0.9340362175	5.858
	GRU-LSTM Hybrid models	0.1392682847	0.1217903526	0.9361206344	13.975

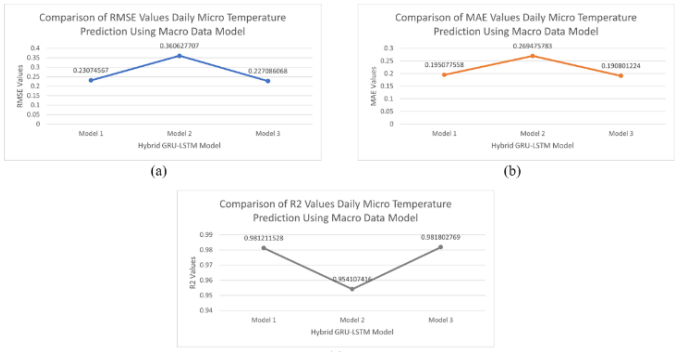


Fig. 15. Analysis of performance model results of GRU Model, LSTM Model, and GRU-LSTM Hybrid Model (a) RMSE Value, (b) MAE Value, (c) R^2 Value, (d) Model training duration.

V. DISCUSSION

The choice of data preprocessing techniques significantly influenced the performance of the GRU-LSTM hybrid model. Using Min-Max Scaling with various NaN handling methods resulted in suboptimal performance, with four out of seven techniques showing negative R^2 values, indicating a poor model fit. However, applying the Fast Fourier Transform (FFT) improved model accuracy across all NaN handling techniques. The Exponential (Weighted) Moving Average technique produced the best results with the lowest RMSE and MAE values and the highest R^2 value. These results indicate that FFT effectively transforms data so that the model can better recognize patterns.

Increasing the amount of training data did not continuously improve model performance. In daily macro temperature data, 300 data points produced the lowest RMSE and MAE values, while 200 data points yielded the highest R^2 value. However, the error rate also increased when the data size increased to 500 or more. A similar pattern was found in the weekly and monthly data tests, where the optimal amount of data varied. These results suggest that excessive data can lead to information overload or noise accumulation, which disrupts model accuracy. This finding is significant for real-world applications where computational resources are limited. This study provides a practical guideline for selecting training data without unnecessary computational costs by identifying the optimal data size for different timescales.

The number of input parameters played a crucial role in improving prediction accuracy. For daily macro temperature predictions, the model achieved the best performance using 11 days of historical data, while for weekly and monthly data, 12 weeks and 12 months of historical data provided the most optimal results. Increasing input parameters beyond a certain threshold did not significantly enhance model performance. This indicates that only a certain amount of historical data provides relevant information for the model to predict future air temperatures. In practical applications, this insight helps optimize data storage and processing requirements by preventing excessive use of historical data that does not contribute to better predictions.

The ratio of training to testing data affected the model's effectiveness. A 90:10 split produced the best daily macro temperature data accuracy, while an 80:20 split yielded optimal results in weekly and monthly datasets. This difference suggests that the optimal ratio depends on the scale of the data used. An extreme split, such as 50:50, speeds up training time but reduces the model's ability to generalize new data. Conversely, a 90:10 ratio risks overfitting because the training data dominates too much. These findings are helpful for practitioners who need to balance training time and model generalization, particularly in operational forecasting systems where real-time updates are essential.

The number of epochs used in training did not significantly affect model accuracy. The most notable improvement occurred between 10 and 20 epochs, where RMSE and MAE values decreased significantly. However, after exceeding 20 epochs, performance gains became insignificant, indicating a saturation point in model training. The best results were obtained with 40 epochs for daily data, 80 epochs for weekly data, and 50 epochs for monthly data. These differences suggest that the model requires a varying number of epochs depending on the scale of the data used. Training time can be optimized for practical deployment by selecting the appropriate number of epochs, reducing computational overhead without sacrificing prediction accuracy.

Batch size had minimal impact on model accuracy. A batch size of 64 yielded the best daily and weekly data performance, while a batch size of 32 was more optimal for monthly data. However, increasing the batch size to 128 slightly decreased model performance. These results indicate that while batch size optimization can affect training stability, changes in batch size do not significantly impact prediction accuracy. This suggests practitioners can choose moderate batch sizes to balance memory usage and model performance. This makes the model scalable for different computing environments, from high-performance servers to edge devices.

Increasing the number of neurons positively affected model accuracy, mainly when the number of neurons increased from 30 to 60, where RMSE values significantly decreased. The best results were obtained with 120 neurons for daily data and 240 neurons for weekly and monthly data. This suggests that more neurons allow the model to capture more complex patterns, but an excessive number of neurons risks overfitting. In practical applications, understanding the optimal number of neurons helps in designing efficient models that maximize accuracy without unnecessary computational complexity, making the model more scalable for large-scale implementation.

Changes in the number of GRU/LSTM and Dense layers did not significantly impact prediction accuracy. The best

configuration for daily data was obtained using one Dense Layer after GRU-LSTM, weekly data using three Dense Layers after GRU-LSTM, and monthly data using two GRU/LSTM Layers and one Dense Layer. Adding Dropout Layers worsened model performance, indicating that dropout is unsuitable for this air temperature prediction scenario. These insights provide practical recommendations for model architecture design, allowing practitioners to avoid unnecessary layers that do not contribute to accuracy improvements, thereby reducing training time and computational costs.

Applying a model trained on macro temperature data to predict micro temperature data every three hours resulted in a high error rate. The large temperature fluctuations within this timeframe made it difficult for the model to recognize temperature change patterns. Among the three models tested, the model trained with monthly data performed the best. This indicates that longer-term macro data provides more stable patterns for micro-temperature predictions. However, the still relatively high error rate suggests that further adjustments are needed for the model to predict micro temperatures accurately. This finding is particularly relevant for real-world applications where localized micro-temperature forecasts are required, such as in precision agriculture or climate-sensitive industries. Future work should explore additional feature engineering techniques or hybrid models incorporating real-time weather variables.

The GRU-LSTM hybrid model consistently performed better than the standalone GRU or LSTM models in all test scenarios. Although the differences in RMSE, MAE, and R^2 values between the three models were insignificant, the hybrid model still showed slightly higher accuracy. However, this advantage came at the cost of longer training times. These results indicate that using only the GRU or LSTM model is sufficient for applications requiring fast predictions. In contrast, the GRU-LSTM hybrid model is recommended for more precise forecasting needs. From a scalability perspective, this suggests that the choice of model should depend on computational resources and accuracy requirements. The standalone GRU or LSTM model is more suitable for real-time applications with limited processing power. In contrast, the GRU-LSTM hybrid model is ideal for high-accuracy forecasting in research and large-scale monitoring systems.

VI. Conclusion

The objective of this study was to see how the number of input parameters, number of data sharing ratios, number of epochs, number of batch sizes, number of neuron sizes, and layer arrangement of the model architecture affected the performance results of the GRU-LSTM hybrid model. The best model performance was obtained from all tests using Macro Air Temperature Data when testing with monthly Macro Air Temperature Data and a 12 Input Parameter scheme. The obtained RMSE is 0.056807, the MAE is 0.046592, and the R^2 is 0.989371. Because the error rate derived from the RMSE and MAE values is relatively low for prediction, the GRU-LSTM Hybrid Model is appropriate for predicting macro air temperature. The data preprocessing stages, the number of input parameters used, and the presence or absence of a Dropout Layer in the model architecture are the indicators that have the most significant influence on model performance. Furthermore, this study investigates whether a model trained on macro data can be used to predict microdata. The GRU-LSTM Hybrid Model, which was built using monthly macro data and 12 Parameter Inputs, produced the best results in predicting Micro Air Temperatures every 3 Hours and Daily Micro Air Temperatures. In microdata tests at daily intervals, the best RMSE, MAE, and R^2 values were 0.227086, 0.190801, and 0.981802, respectively. Because the error rate obtained in testing using daily microdata is relatively low, it can be concluded that micro air temperature predictions can be performed using the GRU-LSTM Hybrid Model, which was trained using monthly macro temperature data. Because of their high accuracy, the results of the 12 input parameters can be used to build a time series air temperature prediction system. The number of inputs indicates the impact on the model's performance results. The Hybrid GRU-LSTM model with 12 inputs can be used to design a temperature prediction application in a wetland environment.

Acknowledgment

This research was funded by the DRTPM Higher Education Excellence Research scheme with Agreement/Contract Number 026/E5/PG.02.00.PL/2023 and Internal PNBP Funds of Universitas Lambung Mangkurat with contract Number 615.82/UN8.2/PL/2023

References

[1] A. Ibrahem Ahmed Osman, A. Najah Ahmed, M. F. Chow, Y. Feng Huang, and A. El-Shafie, "Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia," Ain Shams Eng. J., vol. 12, no. 2, 2021, doi: 10.1016/j.asej.2020.11.011.

[2] Z. Yahya Dewangga and S. Koesuma, "Development of forest fire early warning system based on the wireless sensor network in Lawu Mountain.," J. Phys. Conf. Ser., vol. 1153, no. 1, 2019, doi: 10.1088/1742-6596/1153/1/012025.

[3] A. Ashok, H. P. Rani, and K. V. Jayakumar, "Monitoring of dynamic wetland changes using NDVI and NDWI based landsat imagery," Remote Sens. Appl. Soc. Environ., vol. 23, no. May, p. 100547, 2021, doi: 10.1016/j.rsase.2021.100547.

[4] S. J. Lite, K. J. Bagstad, and J. C. Stromberg, "Riparian plant species richness along lateral and longitudinal gradients of water stress and flood disturbance, San Pedro River, Arizona, USA," J. Arid Environ., vol. 63, no. 4, pp. 785–813, 2005, doi: 10.1016/j.jaridenv.2005.03.026.

[5] R. Tawalbeh, F. Alasali, Z. Ghanem, M. Alghazzawi, and A. Abu-raideh, "Innovative Characterization and Comparative Analysis of Water Level Sensors for Enhanced Early Detection and Warning of Floods," 2023.

[6] N. Novitasari, Y. Sari, Y. F. Arifin, N. F. Mustamin, and E. Maulidiya, "Use Of UAV Images For Peatland Cover Classification Using The Convolutional Neural Network Method," J. Southwest Jiaotong Univ., vol. 6, no. 3, pp. 1–13, 2023.

[7] A. F. Zulkarnain, Y. Sari, and R. Rakhmadani, "Monitoring System for Early Detection of Fire in Wetlands based Internet of Things (IoT) using Fuzzy Methods," IOP Conf. Ser. Mater. Sci. Eng., vol. 1115, no. 1, p. 012007, 2021, doi: 10.1088/1757-899x/1115/1/012007.

[8] A. R. Baskara, Y. Sari, A. A. Anugerah, E. S. Wijaya, and R. A. Pramunendar, "Fire Detection In Wetland Using YOLOv4 And Deep Learning Architecture," in 2022 Seventh International Conference on Informatics and Computing (ICIC), 2022, pp. 1–6. doi: 10.1109/ICIC56845.2022.10006963.

[9] B. Guha and G. Bandyopadhyay, "Gold Price Forecasting Using ARIMA Model," J. Adv. Manag. Sci., vol. 4, no. 2, pp. 117–121, 2016, doi: 10.12720/joams.4.2.117-121.

[10] L. R. Jácome-Galarza, M. A. Realpe-Robalino, J. Paillacho-Corredores, and J. L. Benavides-Maldonado, "Time Series in Sensor Data Using State-of-the-Art Deep Learning Approaches: A Systematic Literature Review,"

in Smart Innovation, Systems and Technologies, 2022, vol. 252. doi: 10.1007/978-981-16-4126-8_45.

[11] H. Kim et al., "Exsolution of Ru Nanoparticles on BaCe0.9Y0.1O3-δ Modifying Geometry and Electronic Structure of Ru for Ammonia Synthesis Reaction Under Mild Conditions," Small, vol. 19, no. 6, 2023, doi: 10.1002/smll.202205424.

[12] M. Alkaff, H. Khatimi, W. Puspita, and Y. Sari, "Modelling and predicting wetland rice production using support vector regression," Telkomnika (Telecommunication Comput. Electron. Control., vol. 17, no. 2, pp. 819–825, 2019, doi: 10.12928/TELKOMNIKA.V17I2.10145.

[13] T. Bernatin, G. Sundari, S. A. Nisha, and M. S. Godwin Premi, "Optimized inter prediction for h.264 video codec," Int. J. Electron. Telecommun., vol. 67, no. 2, 2021, doi: 10.24425/ijet.2021.135981.

[14] V. S. Suruthhi, V. Smita, J. Rolant Gini, and K. I. Ramachandran, "Detection and localization of audio event for home surveillance using CRNN," Int. J. Electron. Telecommun., vol. 67, no. 4, 2021, doi: 10.24425/ijet.2021.139771.

[15] K. E. ArunKumar, D. V. Kalaga, C. M. S. Kumar, M. Kawaji, and T. M. Brenza, "Forecasting of COVID-19 using deep layer Recurrent Neural Networks (RNNs) with Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) cells," Chaos, Solitons and Fractals, vol. 146, 2021, doi: 10.1016/j.chaos.2021.110861.

[16] W. Wu, W. Liao, J. Miao, and G. Du, "Using gated recurrent unit network to forecast short-term load considering impact of electricity price," in Energy Procedia, 2019, vol. 158. doi: 10.1016/j.egypro.2019.01.950.

[17] E. Haque, S. Tabassum, and E. Hossain, "A Comparative Analysis of Deep Neural Networks for Hourly Temperature Forecasting," IEEE Access, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3131533.

[18] M. S. Islam and E. Hossain, "Foreign exchange currency rate prediction using a GRU-LSTM hybrid network," Soft Comput. Lett., vol. 3, 2021, doi: 10.1016/j.socl.2020.100009.

[19] Z. Kong, B. Tang, L. Deng, W. Liu, and Y. Han, "Condition monitoring of wind turbines based on spatio-temporal fusion of SCADA data by convolutional neural networks and gated recurrent units," Renew. Energy, vol. 146, 2020, doi: 10.1016/j.renene.2019.07.033.

[20] K. Khalil, O. Eldash, A. Kumar, and M. Bayoumi, "Machine Learning-Based Approach for Hardware Faults Prediction," IEEE Trans. Circuits Syst. I Regul. Pap., vol. 67, no. 11, 2020, doi: 10.1109/TCSI.2020.3010743.

[21] S. Mishra, C. Bordin, K. Taharaguchi, and I. Palu, "Comparison of deep learning models for multivariate prediction of time series wind power generation and temperature," Energy Reports, vol. 6, 2020, doi: 10.1016/j.egyr.2019.11.009.

[22] Y. Fu, Z. Gao, Y. Liu, A. Zhang, and X. Yin, "Actuator and sensor fault classification for wind turbine systems based on fast fourier transform and uncorrelated multi-linear principal component analysis techniques," Processes, vol. 8, no. 9, 2020, doi: 10.3390/pr8091066.

[23] X. Liu, Z. Lin, and Z. Feng, "Short-term offshore wind speed forecast by seasonal ARIMA - A comparison against GRU and LSTM," Energy, vol. 227, 2021, doi: 10.1016/j.energy.2021.120492.

[24] D. Tukymbekov, A. Saymbetov, M. Nurgaliyev, N. Kuttybay, G. Dosymbetova, and Y. Svanbayev, "Intelligent autonomous street lighting system based on weather forecast using LSTM," Energy, vol. 231, 2021, doi: 10.1016/j.energy.2021.120902.

[25] H. D. Nguyen, K. P. Tran, S. Thomassey, and M. Hamad, "Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management," Int. J. Inf. Manage., vol. 57, 2021, doi: 10.1016/j.ijinfomgt.2020.102282.

[26] J. Luo, Z. Zhang, Y. Fu, and F. Rao, "Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms," Results Phys., vol. 27, 2021, doi: 10.1016/j.rinp.2021.104462.

[27] N. AlDahoul et al., "Suspended sediment load prediction using long short-term memory neural network," Sci. Rep., vol. 11, no. 1, 2021, doi: 10.1038/s41598-021-87415-4.

[28] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," PeerJ Comput. Sci., vol. 7, 2021, doi: 10.7717/PEERJ-CS.623.

[29] X. Zhang, Q. Zhang, G. Zhang, Z. Nie, Z. Gui, and H. Que, "A novel hybrid data-driven model for daily land surface temperature forecasting using long short-term memory neural network based on ensemble empirical mode decomposition," Int. J. Environ. Res. Public Health, vol. 15, no. 5, 2018, doi: 10.3390/ijerph15051032.

Lightweight Parabola Chaotic Keyed Hash Using SRAM-PUF for IoT Authentication

Nattagit Jiteurtragool[1], Jirayu Samkunta[2], Patinya Ketthong[3]
Department of Computer and Information Sciences-Faculty of Applied Science,
King Mongkut's University of Technology, North Bangkok, Bangkok, Thailand[1]
Graduate School of Science and Technology, Gunma University, Kiryu, Japan[2]
Intelligent Electronics System Laboratory (IES), Thai-Nichi Institute of Technology, Bangkok, Thailand[3]

***Abstract*—This paper introduces a lightweight and efficient keyed hash function tailored for resource-constrained Internet of Things (IoT) environments, leveraging the chaotic properties of the Parabola Chaotic Map. By combining the inherent unpredictability of chaotic systems with a streamlined cryptographic design, the proposed hash function ensures robust security and low computational overhead. The function is further strengthened by integrating it with a Physical Unclonable Function (PUF) based on SRAM initial values, which serves as a secure and tamper-resistant source of device-specific keys. Experimental validation on an ESP32 microcontroller demonstrates the function's high sensitivity to input variations, exceptional statistical randomness, and resistance to cryptographic attacks, including collisions and differential analysis. With a mean bit-change probability nearing the ideal 50% and 100% reliability in key generation under varying conditions, the system addresses critical IoT security challenges such as cloning, replay attacks, and tampering. This work contributes a novel solution that combines chaos theory and hardware-based security to advance secure, efficient, and scalable authentication mechanisms for IoT applications.**

***Keywords*—SRAM PUF; PUF key generation; chaotic keyed hash; device authentication; discrete-time chaotic**

I. Introduction

The Internet of Things (IoT) has revolutionized the way we interact with technology, connecting billions of devices globally to create a vast network of interconnected systems. IoT extends to various applications, including smart homes, healthcare, industrial automation, and transportation systems [1-2]. For instance, IoT-enabled medical devices can monitor patients in real-time, while industrial sensors optimize manufacturing processes. The ability of IoT to enhance operational efficiency, reduce costs, and enable innovative services underscores its importance in modern technological advancements. However, the rapid proliferation of IoT devices also brings significant challenges, particularly in managing the security of these interconnected networks [3-4].

Device authentication is a fundamental aspect of IoT security, ensuring that only legitimate devices can communicate within the network. Conventional methods for authentication often rely on cryptographic keys stored in memory, which are vulnerable to physical attacks, cloning, and software-based exploits. These issues highlight the need for more robust and tamper-resistant mechanisms for IoT security [5-6].

However, IoT devices face unique constraints compared to traditional IT systems. These constraints include limited computational resources, such as low processing power, minimal memory, and restricted energy availability [7]. These limitations necessitate lightweight and efficient security mechanisms that do not overwhelm the device's capabilities. Moreover, IoT devices are often deployed in physically exposed environments, making them more susceptible to tampering, theft, and side-channel attacks. Unlike traditional IT systems, which are often housed in secure, controlled environments, IoT devices may operate in untrusted or hostile settings, increasing the potential attack surface [8]. As a result, IoT security mechanisms must address these heightened risks while remaining compatible with resource-constrained hardware.

In this paper, we propose a novel approach that integrates SRAM-based PUFs with a keyed hash function based on the Parabola Chaotic Map. This combination provides a lightweight and secure mechanism for device authentication in resource-constrained IoT environments. By leveraging the inherent randomness of SRAM PUFs and the entropy-enhancing properties of the chaotic hash function, our method ensures robust security while maintaining computational efficiency over a resource-constrained environment of IoT.

The rest of this paper is organized as follows: Section II provides an in-depth discussion of the background concepts, including an overview of IoT security challenges, Physical Unclonable Functions (PUFs), and chaotic hash functions. Section III introduces the proposed system, elaborating on the integration of SRAM-based PUFs with the Parabola Chaotic Map-based keyed hash function and detailing its implementation on the ESP32 microcontroller. Section IV presents the results and analysis of the system's performance. Finally, Section V concludes the paper by summarizing key findings and proposing directions for future research.

II. Background

A. IoT Security Challenges

Security is a critical concern in IoT due to the sensitive data these devices handle and their deployment in diverse environments. Fig. 1 shows a classification of IoT security threats along with the security approaches for each IoT layer [9-10]. As IoT architectures are commonly divided into three layers [11]: The perception layer, which includes sensors and actuators, involves data collection through directly interfaces

with the physical environment, faces risks such as interception of data transmitted in and out of devices, physical tampering or node capture attacks. The network layer, which is responsible for transmitting data between devices, gateways, and cloud platforms, may have vulnerability to data interception during the transmission, or being attack by overloading the network traffic to make it unavailable. Lastly, the application layer, which able to provide the user interfacing control and processes the collected data, can be compromised by software vulnerabilities and unauthorized access. Unauthorized access to IoT devices can lead to data breaches, operational disruptions, and even physical damage in critical systems.

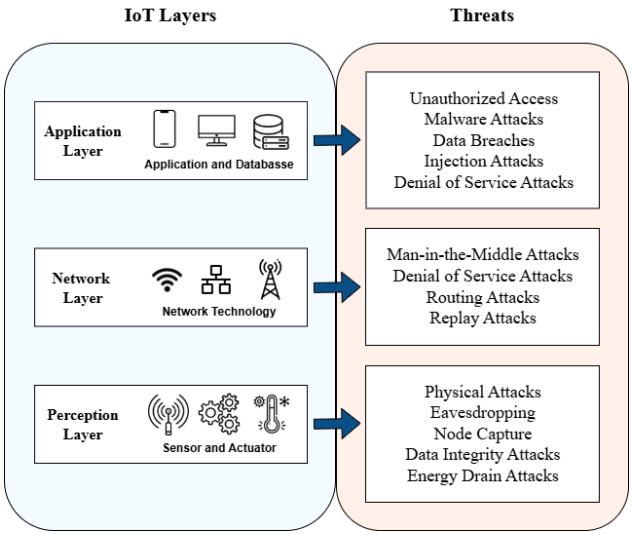


Fig. 1. IoT layers, and security threats.

B. Physical Unclonable Functions (PUFs)

Physical Unclonable Functions (PUFs) have emerged as a promising solution to IoT security challenges [12-13]. A PUF leverages the intrinsic physical variations in hardware components to generate unique and unclonable identifiers [14]. Among the various types of PUFs [15], SRAM-based PUFs are particularly attractive due to their simplicity and widespread availability in existing hardware [16-17]. When powered on, SRAM cells exhibit random initialization values influenced by manufacturing variations, making them an ideal candidate for generating device-specific keys. The stability of these values can be enhanced through error correction, enabling consistent and reliable use for cryptographic purposes, including as keys for keyed hash functions.

PUFs leverage manufacturing variations to generate unique and unclonable responses from hardware components. SRAM PUFs use the power-up state of uninitialized SRAM cells as a source of randomness. These initial values are device-specific and can be stabilized using error correction mechanisms for consistent key generation. PUFs eliminate the need to store cryptographic keys, thereby reducing risks associated with key theft or compromise. SRAM-based PUFs, specifically, are advantageous because of their availability in existing devices and their ability to function without modifications to the hardware design.

C. Hash Function

Hash functions have been implemented in the contexts of cybersecurity and information security applications [18-20], such as ensuring data integrity, authentication, digital signatures [21], protocol encryption [22], number generation [23], password security [24], or blockchain applications [25]. Hash functions are generally categorized into two types: unkeyed and keyed. Unkeyed hash functions rely solely on the input message to generate the hash value. On the contrary, Keyed hash functions, which incorporate a secret key into the hashing process, can provide enhanced security by protecting against tampering and brute-force attacks. Unkeyed hash functions such as MD5 and SHA-1, which are widely used in security applications and protocols, have primarily relied on logical operations and multi-round iterations.

Aside from hash functions, chaotic systems have gained significant traction in the field of cryptography over the last decade [26-27]. Chaotic systems are defined by their deterministic nature combined with an inherent unpredictability. These systems exhibit extreme sensitivity to initial conditions, long-term unpredictability, and complex yet non-random behavior. Such characteristics make them ideal for cryptographic applications [28], where unpredictability and high entropy are essential for ensuring secure operations. Among the tools derived from chaotic systems, chaotic maps stand out as mathematical functions capable of generating sequences with these properties. Examples include the Logistic Map, Tent Map, and Parabola Map, each of which provides unique advantages in terms of randomness and computational efficiency. The ability to generate pseudo-random values from simple mathematical operations makes chaotic maps particularly valuable for lightweight and efficient cryptographic systems, especially in resource-constrained environments like IoT.

Hashing efficiency is contingent upon inherent ciphers which necessarily require extensive computation processes. Combining input sensitivity with the entropy-generating properties of chaotic systems resulted in keyed hash functions such as the Chaotic Map-based approach [29-31], which offers a lightweight and secure solution for IoT environments. These functions ensure robust performance even in resource-constrained devices, making them a suitable choice for integrating with other security approaches such as hardware-based mechanisms like PUFs [32-33]. Moreover, the chaotic map-based approach provides an inherent randomness that enhances cryptographic strength, making it resilient against cryptanalysis. This approach also enables adaptability in various IoT applications by allowing parameter adjustments, ensuring that it can balance security needs with computational overhead effectively.

III. Proposed System

A. SRAM-PUF Key Generation on ESP32

The ESP32 is a low-cost, low-power, and highly versatile microcontroller system-on-chip (SoC) designed by Espressif Systems [34]. It is widely used in Internet of Things (IoT) applications due to its powerful features, extensive connectivity options, and efficient performance. The ESP32 microcontroller features a versatile internal memory architecture that includes

multiple Static Random Access Memory (SRAM) regions. The primary SRAM, divided into three memory blocks—SRAM0, SRAM1, and SRAM2—totals 520 KB. These memory blocks are shared between instruction and data storage, enabling efficient use of memory resources. SRAM0 and SRAM1 are typically used for both instruction and data storage, while SRAM2 is primarily reserved for general-purpose data storage. The instruction storage is accessible through the instruction memory bus (IRAM), allowing executable code to run efficiently. Conversely, the data storage, accessed through the data memory bus (DRAM), is non-executable and dedicated to runtime data. Fig. 2 illustrates the internal SRAM memory architecture of the ESP32.

In this paper, the ESP32 microcontroller is used as a test platform. The device's on-chip SRAM is leveraged for PUF generation, while its computational capabilities handle the chaotic hash function. Memory block from SRAM2 is chosen for the PUF generation. This region is preferred because it remains uninitialized during the boot process, preserving its unique power-up state. These states, influenced by inherent manufacturing differences, exhibit sufficient entropy to serve as a reliable source for PUF generation. By reserving this address range exclusively for PUF purposes, the implementation ensures that other system operations do not interfere with the PUF response. This allocation also simplifies access and management of the PUF-specific memory blocks while maintaining the system's overall efficiency and stability.

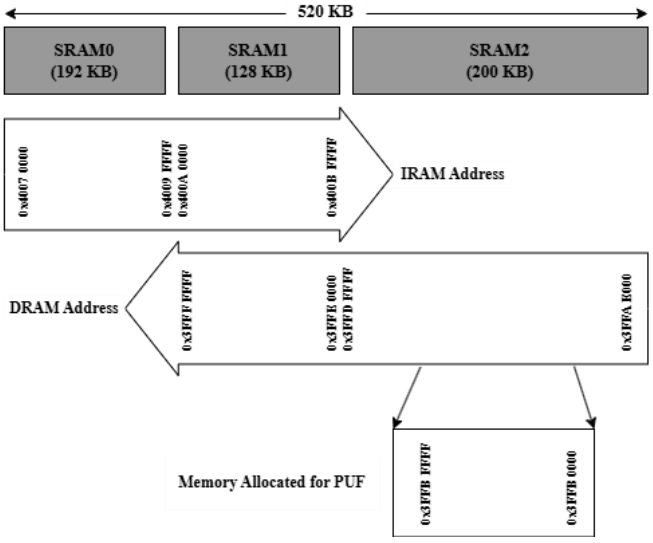


Fig. 2. Internal SRAM memory allocation of ESP32 chip.

The ESP32 microcontroller's low power consumption and computational efficiency make it an ideal platform for implementing this SRAM-based PUF mechanism. Its versatile memory architecture supports the storage and processing of the raw and stabilized PUF responses, while its dual-core processor efficiently handles the computations required for Error Correction to stabilize the output.

The key workflow of the implementation of SRAM-PUF Key Generation on ESP32 involves extracting device-specific responses from SRAM and stabilizing these responses using a Majority Vote Key instead of Error Correction Codes (ECC). This process ensures unique, tamper-resistant authentication for IoT devices, addressing the challenges of environmental variability and hardware noise. By leveraging both techniques, the system achieves a high degree of stability and reliability while maintaining computational efficiency suitable for resource-constrained IoT applications.

During device power-up, uninitialized SRAM cells generate random values due to manufacturing variations. These values are read and processed to form a raw binary response. However, environmental factors such as temperature fluctuations, voltage variations, and hardware aging can introduce inconsistencies in the raw responses across multiple power cycles. To address these challenges, the Majority Vote Key approach is applied.

With Majority Vote Key, the binary representations of the SRAM data undergo majority voting over a defined number of iterations, enhancing the robustness of the generated key against noise and instability. This process effectively averages out noise and random bit flips, ensuring that transient errors caused by environmental factors do not impact the final response. The output of this stage is a binary response that is significantly more stable and reliable than any single raw response. The stabilized output then serves as the key for the chaotic hash function.

B. Parabola Chaotic Keyed Hash Function

The Generalized Parabola Chaotic Map, developed in our previous work [35], serves as a robust randomness source for the proposed keyed hash function. This map builds upon well-known chaotic systems like the tent map, logistic map, and Gauss map, offering enhanced entropy generation and sensitivity to initial conditions. The mathematical definition of the Parabola Chaotic Map is as follows:

$$x_{n+1} = \mp A f_{\mathrm{NL}}(B x_n) \pm C \tag{1}$$

where the parameters *A*, *B* and *C* are real constants, and the $f_{NL}(x)$ is a parabola function. These parameters contribute to the map's chaotic behavior, ensuring its suitability for cryptographic applications.

The Parabola Chaotic Map-based keyed hash function is specifically designed for resource-constrained environments, making it an ideal choice for IoT applications. This lightweight cryptographic mechanism capitalizes on the pseudo-random behavior and high entropy characteristics of the chaotic map, which enhances its resistance to attacks such as collision, preimage, and differential cryptanalysis. Furthermore, the efficient computational requirements ensure compatibility with IoT devices that often have limited processing power and energy resources.

The hash function operates through the following steps, illustrated in Fig. 3:

1) Input handling: The input data and the stabilized keys from the SRAM-PUF are prepared. If the input length is not a multiple of 8, it is padded with zeros to meet the length requirement.

2) Initialization: A chaotic sequence is initialized, with the number of iterations determined by the length of the padded input.

3) Chaos mapping loop: The chaotic sequence is iteratively calculated, where each new value is derived from the previous value and a portion of the input bits.

4) Continuation of chaos mapping: Additional values are generated as needed to complete the hash output.

5) Output generation: The final chaotic sequence is processed to extract the relevant portion, forming the hash output.

The streamlined structure of the hash function ensures high efficiency while maintaining robust cryptographic properties. The nonlinear dynamics of the Parabola Chaotic Map enable the generation of highly unpredictable outputs, providing a secure foundation for cryptographic applications in IoT environments.

C. IoT Device Authentication Mechanism

The combination of the parabola chaotic map-based keyed hash function with SRAM-PUF technology is aimed to create a secure and lightweight authentication mechanism. The SRAM-PUF, implemented on an ESP32 microcontroller, generates a stabilized output key derived from the intrinsic physical variations of the SRAM memory. This stabilized key serves as the secret input to the keyed hash function, which, in turn, generates unique and unpredictable authentication tokens. The system leverages these tokens to implement a secure authentication mechanism.

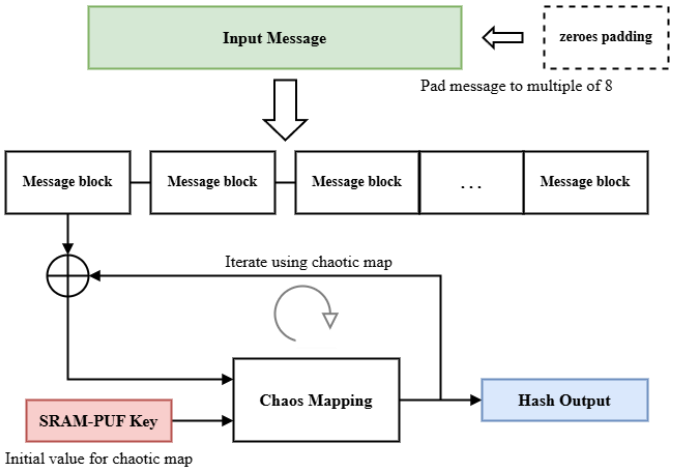


Fig. 3. SRAM-PUF based chaotic keyed hash.

As challenge-response pair (CRP) was chosen as an authentication mechanism, the authenticator sends a challenge to the device during an authentication session. The device combines this challenge with its stabilized SRAM-PUF key and processes through the chaotic keyed hash function. The result is an authentication token that is unique to the device and the specific challenge. This response is then sent back to the authenticator for authentication. Since the SRAM-PUF-derived key is unclonable and never directly transmitted, the system ensures that even if an attacker intercepts the challenge or response, they cannot reconstruct the key or generate valid tokens. The authentication mechanism is shown in Fig. 4.

The verifier, having the original challenge and a stored copy of the expected key or hash behavior, computes the expected token using the same keyed hash function. If the computed token matches the received token, the device is authenticated successfully. This process ensures the following:

1) Device uniqueness: The unclonable nature of the SRAM-PUF key ensures that each device produces a unique response, making it nearly impossible to replicate the authentication behavior of another device.

2) Resistance to replay attacks: Since the response is dynamically generated for each challenge, replaying a previously captured response will fail to authenticate the device.

3) Lightweight operation: The combination of SRAM-PUF and chaotic keyed hash function provides a computationally efficient authentication mechanism suitable for IoT devices with limited resources.

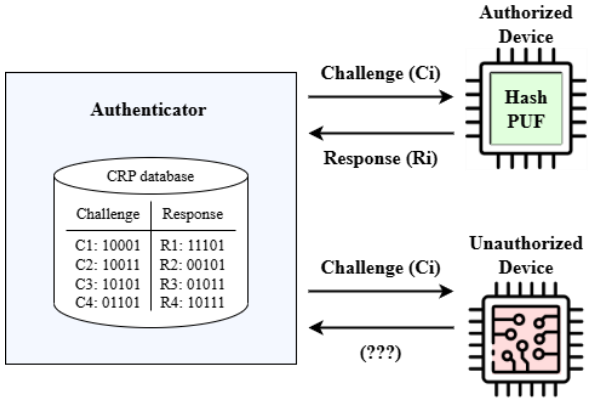


Fig. 4. The authentication mechanism.

By integrating the parabola chaotic map-based keyed hash function with SRAM-PUF, the proposed system achieves a high degree of security and efficiency, ensuring robust authentication for IoT devices in diverse and potentially hostile environments.

IV. Results and Analysis

This section evaluates the performance of the Parabola Chaotic Map-based keyed hash function and the SRAM-PUF mechanism implemented on the ESP32 microcontroller on security, efficiency, and robustness in IoT authentication.

A. Sensitivity and Reliability Analysis

The reliability of the SRAM-PUF is critical factor in determining the robustness of the proposed keyed hash function. In this study, experimental test was conducted using two ESP32 boards with the same model (ESP32WROOM32).

Reliability measures the consistency of the PUF response to the same challenge across multiple instances and varying environmental conditions, such as temperature fluctuations and power supply variations. The proposed system leverages the Majority Vote Key technique instead of traditional error correction codes, ensuring consistent key generation without significant computational overhead.

The evaluation of reliability involved extensive testing with two ESP32 boards. To assess reliability, a set of predefined challenges was repeatedly presented to the PUF, and the responses were collected over 100 iterations per challenge. The

collected responses were then analyzed to identify any inconsistencies.

Remarkably, the PUF demonstrated a reliability of 100%, consistently producing identical responses to the same challenges across all tests. This perfect reliability validates the robustness of the Majority Vote Key mechanism, which effectively mitigated the effects of noise and variability. Over prolonged operational periods, the mechanism maintained its performance, demonstrating resilience against potential aging effects in the SRAM hardware. This result validates the unparalleled reliability of the proposed SRAM-PUF mechanism, reinforcing its suitability for an integration with the parabola chaotic map-based keyed hash function.

Consequently, to demonstrate exceptional sensitivity to minor changes in the input message of the proposed keyed hash function, a series of experiments were conducted such as character replacements, additions, or deletions.

Case1: The Original message: “This is a simple absolute chaos based keyed hash function."

Case2: Replace the first character of the original message ‘‘T’’ by ‘‘t’’.

Case3: Add a space to the end of the original message.

Case4: Delete the full stop symbol at the end of the message.

The corresponding 128-bit hash values for each input message are presented in hexadecimal format as follows.

Case1: FB47 2C55 6481 D81D 379B 13F4 7616 733C

Case2: 80FB 471D 72B4 71C6 A735 7FA0 7868 CF77

Case3: 81E9 1D90 FB39 0FB4 5028 1F38 00D1 55AF

Case4: 90E9 1D71 D72B 5639 6F4B 75F9 195B 2CD1

The graphical display of binary sequences is shown in Fig. 5.

The results clearly demonstrate the proposed hash function exceptional sensitivity to such alterations and its suitability for tamper-resistant IoT authentication systems and dependable performance across diverse scenarios.

B. Collision analysis

The collision test consists in computing the difference between the hash values obtained from the original message and from a modified version of this message. The difference is computed based on the following formula:

$$d = \sum_{i=1}^{N} |dec(m_i) - dec(m_i')| \qquad (2)$$

where m_i and m_i' is the ith ASCII character of the original and the new hash value respectively, while the *dec*(.) converts an ASCII character to its decimal value.

Table I, show the minimum, maximum and mean values of the absolute difference of original and new hash values, where simulation repeat N = 10,000 time.

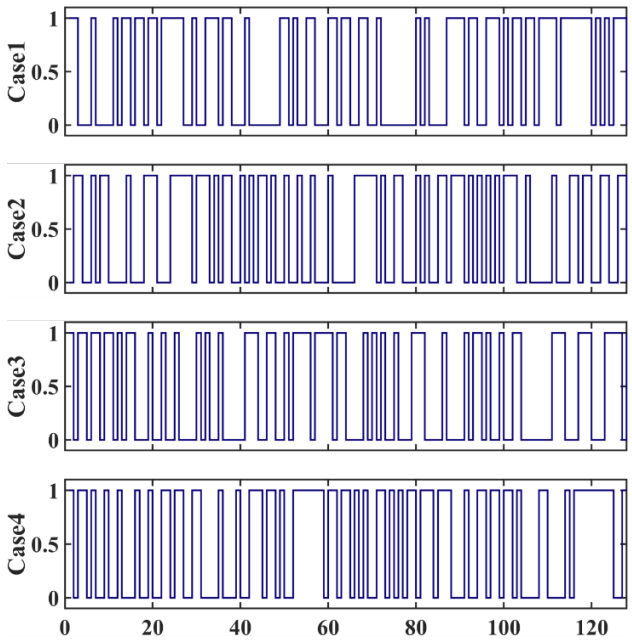


Fig. 5. The 128-bit hash values under different conditions.

TABLE I. COLLISION ANALYSIS OF HASH ALGORITHMS

Hash Algorithm	Min	Max	Mean
MD5(128bit)	590	2074	1304
SHA-1(160bit)	795	2730	1603
Proposed scheme (128bit)	549	2590	1387
Proposed scheme (160bit)	602	2702	1647
Proposed scheme (256bit)	1509	4005	2710

C. Statistical Analysis

Confusion and diffusion are critical metrics for evaluating the performance of hashing algorithms. To conduct a statistical analysis of these properties, the following experiments were performed:

- A random message of length L = n*50 was generated, where n represents the desired hash value length.
- A single bit of the original message was randomly selected and flipped, and a new n-bit hash value was calculated.
- The two n-bit hash values were compared to determine the number of changed bits.
- This simulation was repeated N times for different values of N (256, 512, 1024, 2048, and 10,000) and varying hash value lengths n (128, 256, and 512).

The following statistics were collected from the results obtained (see Tables II, III and IV):

- Minimum number of changed bits (B_{min}):
- Maximum changed bit (B_{max}):
- Mean changed bit $\overline{(B)}$:
- Mean changed probability (P):

- Standard deviation of the changed bit (ΔB):
- Standard deviation (ΔP):

TABLE II. STATISTICAL RESULT FOR A 128-BIT HASH VALUE

128-bit	*N* times				
	256	512	1024	2048	10000
B_{min}	48	41	38	21	8
B_{max}	80	83	83	96	86
$\overline{B}$	64.40	63.93	63.73	63.76	63.83
P (%)	50.31	49.95	49.79	49.82	49.87
ΔB	5.33	6.05	6.11	6.27	6.27
ΔP (%)	4.16	4.73	4.77	4.9	4.89

TABLE III. STATISTICAL RESULT FOR A 256-BIT HASH VALUE

256-bit	*N* times				
	256	512	1024	2048	10000
B_{min}	95	95	86	82	80
B_{max}	151	153	154	154	158
$\overline{B}$	127.93	127.80	127.35	127.40	127.41
P (%)	49.97	49.92	49.75	49.77	49.77
ΔB	8.95	8.99	9.28	8.93	9.16
ΔP (%)	3.50	3.51	3.63	3.49	3.58

TABLE IV. STATISTICAL RESULT FOR A 512-BIT HASH VALUE

512-bit	*N* times				
	256	512	1024	2048	10000
B_{min}	195	195	195	195	175
B_{max}	288	288	288	296	296
$\overline{B}$	255.17	254.87	254.71	254.82	254.80
P (%)	49.84	49.78	49.75	49.77	49.76
ΔB	12.64	12.52	12.61	12.86	12.94
ΔP (%)	2.47	2.45	2.46	2.51	2.53

The distribution of the number of changed bits can be analyzed using both a plot of the distribution of changed bits and a histogram as shown in Fig. 6 (a) and (b), respectively. This illustrates the distributions for n = 256 and N = 10000. The results indicate that the number of changed bits is consistently close to n/2 or 50%, and the histograms exhibit a normal distribution with a mean of n/2. This corresponds to a mean changed bit number and a mean changed probability that are very close to the ideal values of n/2 bits and 50%, respectively.

To achieve high security and prevent major attacks, the parabola chaotic keyed hash function based on SRAM-PUF key demonstrate high sensitivity to input changes, with minor variations in the input producing significant differences in the hash output. Statistical analyses confirm that the reliability of the proposed system again a single bit changes in the challenge would result in nearly 50% change in responses, which is an ideal state.

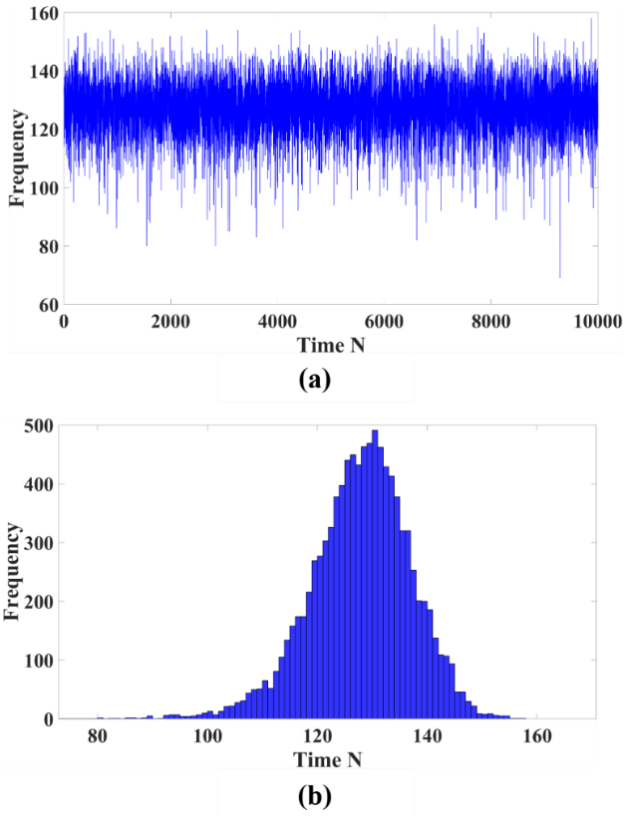


Fig. 6. Plot distribution of the number of bits changed (*a*) and Histogram distribution of the number of changed bits (*b*) for n = 256 and N = 10000.

V. CONCLUSION

This paper introduces a novel authentication mechanism that combines the hardware-level uniqueness of SRAM-based Physical Unclonable Functions (PUFs) with the cryptographic robustness of a Parabola Chaotic Map-based keyed hash function. This integration is designed to meet the stringent requirements of security, reliability, and computational efficiency for Internet of Things (IoT) applications operating in resource-constrained environments.

The proposed approach leverages the inherent physical variations in SRAM to generate unclonable keys, ensuring device-specific uniqueness without the need for external key storage. The chaotic hash function, built on the properties of the Parabola Chaotic Map, enhances security through its high sensitivity to input changes and entropy-generating capabilities, offering strong resistance to cryptographic attacks such as collision, preimage, and differential analysis.

Experimental evaluations confirm the system's effectiveness. The Majority Vote Key technique, employed to stabilize SRAM responses, demonstrates 100% reliability. Statistical analyses further validate the hash function's robustness, achieving near-ideal diffusion and confusion properties with a mean bit-change probability close to 50% in response to input variations. These results underscore the system's capability to deliver robust security and consistent performance.

This work contributes to the field of IoT security by providing a lightweight, tamper-resistant solution that mitigates risks associated with cloning, replay attacks, and physical tampering. The simplicity of the proposed design ensures compatibility with the limited resources of IoT devices, while its high reliability and security make it a promising candidate for broader adoption.

Future research will explore further optimizations of the system's computational and energy efficiency. Additionally, extending the application of this approach to other cryptographic contexts and implementing hardware prototypes for large-scale testing will provide further validation and refinement of the proposed solution.

REFERENCES

[1] Minerva, Roberto, Abyi Biru, and Domenico Rotondi, "Towards a definition of the Internet of Things (IoT). *IEEE Internet Initiative*, vol. 1, pp. 1-86, 2015

[2] M. Mohammadi, M. Aledhari, A. Al-Fuqaha, Internet of things: a survey on enabling technologies, protocols and applications. *IEEE Commun. Surveys Tuts*. vol. 17, pp. 2347–2376, 2015

[3] Shafiq, M., Gu, Z., Cheikhrouhou, O., Alhakami, W., & Hamam, H. (2022). The Rise of "Internet of Things": Review and Open Research Issues Related to Detection and Prevention of IoT-Based Security Attacks. *Wireless Communications and Mobile Computing*, *2022*(1), 8669348.

[4] Abosata, N., Al-Rubaye, S., Inalhan, G., & Emmanouilidis, C. (2021). Internet of things for system integrity: A comprehensive survey on security, attacks and countermeasures for industrial applications. *Sensors*, *21*(11), 3654.

[5] S. Dargaoui, et al., "Internet of Things Authentication Protocols: Comparative Study," *Computers, Materials & Continua*, vol. 79, no. 1, 2024.

[6] M. Trnka, et al., "Systematic review of authentication and authorization advancements for the Internet of Things," *Sensors*, vol. 22, no. 4, p. 1361, 2022.

[7] F. Pereira, et al., "Challenges in resource-constrained IoT devices: Energy and communication as critical success factors for future IoT deployment," *Sensors*, vol. 20, no. 22, p. 6420, 2020.

[8] T. Sasi, et al., "A comprehensive survey on IoT attacks: Taxonomy, detection mechanisms and challenges," *Journal of Information and Intelligence*, vol. 2, no. 6, pp. 455-513, 2024.

[9] S. Shin and T. Kwon, "A privacy-preserving authentication, authorization, and key agreement scheme for wireless sensor networks in 5G-integrated Internet of Things," IEEE Access, vol. 8, pp. 67555–67571, 2020.

[10] N. Abosata, S. Al-Rubaye, G. Inalhan, and C. Emmanouilidis, "Internet of things for system integrity: A comprehensive survey on security, attacks and countermeasures for industrial applications," *Sensors*, vol. 21, no. 11, p. 3654, 2021.

[11] Institute of Electrical and Electronics Engineers, "IEEE Standard for an Architectural Framework for the Internet of Things (IoT)," IEEE Std 2413-2019, pp. 1-269, 2020.

[12] C. Herder, M. -D. Yu, F. Koushanfar and S. Devadas, "Physical Unclonable Functions and Applications: A Tutorial," in *Proceedings of the IEEE*, vol. 102, no. 8, pp. 1126-1141, Aug. 2014

[13] A. Shamsoshoara, et al., "A survey on physical unclonable function (PUF)-based security solutions for Internet of Things," *Computer Networks*, vol. 183, p. 107593, 2020.

[14] U. Rührmair et al., "PUFs: Myth, fact or busted? A security evaluation of physically unclonable functions (PUFs) cast in silicon," *IEEE Trans. Dependable Secure Comput*., vol. 10, no. 3, pp. 193–206, 2013

[15] F. Farha, H. Ning, K. Ali, L. Chen, and C. D. Nugent, "SRAM-PUF based entities authentication scheme for resource-constrained IoT devices," *IEEE Internet of Things Journal*, vol. 0, pp. 1-10, 2020.

[16] G. E. Suh and S. Devadas, "Physical unclonable functions for device authentication and secret key generation," in *Proc. 44th Annu. Design Autom. Conf*., pp. 9-14, 2007.

[17] Böhm, Christoph, Maximilian Hofer, and Wolfgang Pribyl. "A microcontroller sram-puf." in *2011 5th International Conference on Network and System Security*, pp. 269-273. 2011.

[18] ISO/IEC 10118-3:2004, "Information technology - Security techniques - Hash-functions - Part 3: Dedicated hash-functions," 2004.

[19] A. Menezes, P. Van Oorschot, and S. Vanstone, "Handbook of Applied Cryptography," CRC Press, 1996.

[20] D. R. Stinson, "Cryptography: Theory and Practice," CRC Press, 2005.

[21] J. Buchmann, E. Dahmen, and M. Szydlo, "Hash-based digital signature schemes," in *Post-quantum cryptography*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 35-93, 2009.

[22] Y. Yang, et al., "Research on the hash function structures and its application," *Wireless Personal Communications*, vol. 94, pp. 2969-2985, 2017.

[23] Van der Leest, Vincent, Erik Van der Sluis, Geert-Jan Schrijen, Pim Tuyls, and Helena Handschuh. "Efficient implementation of true random number generator based on SRAM PUFs." In *Cryptography and Security: From Theory to Applications: Essays Dedicated to Jean-Jacques Quisquater on the Occasion of His 65th Birthday*, pp. 300-318, 2012.

[24] Q. Guo, et al., "PUFPass: A password management mechanism based on software/hardware codesign," *Integration*, vol. 64, pp. 173-183, 2019.

[25] P. Velmurugadass, et al., "Enhancing Blockchain security in cloud computing with IoT environment using ECIES and cryptography hash algorithm," *Materials Today: Proceedings*, vol. 37, pp. 2653-2659, 2021.

[26] S. N. Elaydi, Discrete Chaos: With Applications in Science and Engineering. Chapman and Hall/CRC, 2007.

[27] S. Boccaletti, C. Grebogi, Y.-C. Lai, H. Mancini, and D. Maza, "The control of chaos: theory and applications," *Physics Reports*, vol. 329, no. 3, pp. 103-197, 2000.

[28] N. Jiteurtragool, T. Masayoshi, and W. San-Um, "Robustification of a one-dimensional generic sigmoidal chaotic map with application of true random bit generation," *Entropy*, vol. 20, no. 2, p. 136, 2018.

[29] P. Ayubi, S. Setayeshi, and A.M. Rahmani, "Chaotic complex hashing: A simple chaotic keyed hash function based on complex quadratic map," *Chaos, Solitons & Fractals*, vol. 173, pp. 113647, 2023.

[30] N. Jiteurtragool, et al., ''A topologically simple keyed hash function based on circular chaotic sinusoidal map network,'' in *15th International Conference on Advanced Communications Technology* (ICACT), pp. 1089–1094, 2013.

[31] H Liu, et al., "Keyed hash function using hyper chaotic system with time-varying parameters perturbation," *IEEE Access*, vol. 7, pp.37211-37219, 2019.

[32] A. Mostafa, S. J. Lee, and Y. K. Peker, "Physical unclonable function and hashing are all you need to mutually authenticate IoT devices," *Sensors*, vol. 20, no. 16, p. 4361, 2020.

[33] A. Braeken, "PUF based authentication protocol for IoT," *Symmetry*, vol. 10, no. 8, p. 352, 2018.

[34] Espressif Systems, "ESP32 Technical Reference Manual," 2023.

[35] N. Jiteurtragool, "Generalized parabola chaotic map for pseudorandom random number generator," in *26th International Conference on Advanced Communications Technology (ICACT)*, pp. 53-56, , 2024.

A Systematic Review of Metaheuristic Algorithms in Human Activity Recognition: Applications, Trends, and Challenges

John Deutero Kisoi, Norfadzlan Yusup, Syahrul Nizam Junaini
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Kota Samarahan, Malaysia

***Abstract*—Metaheuristic algorithms have emerged as promising techniques for optimizing human activity recognition (HAR) systems. This systematic review examines the application of these algorithms in HAR by analyzing relevant literature published between 2019 and 2024. A comprehensive search across multiple databases yielded 27 studies that met the inclusion criteria. The analysis revealed that Genetic Algorithms (GA) exhibit classification accuracy rates ranging from 88.25% to 96.00% in activity recognition and up to 90.63% in localization tasks. Notably, Oppositional and Chaos Particle Swarm Optimization (OCPSO) combined with MI-1DCNN significantly improves detection accuracy, demonstrating a 2.82% improvement over standard PSO with Support Vector Machine (SVM) as classifier approaches. Our analysis highlights a growing trend toward hybrid metaheuristic approaches that enhance feature selection and classifier optimization. However, challenges related to computational cost and scalability persist, underscoring key areas for future research. These findings emphasize the potential of metaheuristic algorithms to significantly advance HAR. Future studies should explore the development of more computationally efficient hybrid models and the integration of metaheuristic optimization with deep learning architectures to enhance system robustness and adaptability.**

***Keywords*—Metaheuristic algorithm; human activity recognition; systematic review; application; trend; challenge; literature**

I. INTRODUCTION

Human Activity Recognition (HAR) is a crucial field that extends its influence across diverse domains, including healthcare, sports, and security, by employing its ability to classify human body movements and gestures based on sensor data [1], [2]. The incorporation of HAR within these domains has initiated significant transformative possibilities, reshaping patient care, enhancing athletic training, and strengthening security protocols [3]. Nevertheless, the effectiveness of HAR relies heavily on their ability to process complex, multidimensional data accurately. To address this complexity, metaheuristic algorithms have proven to be invaluable resources, drawing inspiration from natural phenomena such as evolution and swarm behaviors [4], [5]. Metaheuristic algorithms are powerful optimization strategies designed to tackle complex problems that traditional methods struggle to solve. They allow for the exploration of extensive search spaces to identify the best possible solutions to various issues [6]. The development of these algorithms has taken place over several decades and draws inspiration from natural systems, leading to a range of innovative optimization techniques. We believe that metaheuristic algorithms will continue to be instrumental in driving advancements in new technologies and applications, proving to be invaluable tools for addressing intricate optimization challenges across different fields.

In recent years, HAR technology has revolutionized healthcare monitoring through sophisticated patient movement analysis and personalized rehabilitation programs, while enabling early detection of health issues through non-invasive monitoring methods that provide real-time treatment efficacy feedback. This innovative technology extends its capabilities into sports applications, where it facilitates precise analysis of athletic performance through comprehensive movement and strain monitoring, and further demonstrates its versatility in security systems by identifying unauthorized access and unusual behavioral patterns. The integration of HAR systems across these diverse domains exemplifies its fundamental role in advancing human-centric technological solutions that enhance monitoring, analysis, and decision-making processes in critical sectors.

Despite the advancements in HAR technology, challenges persist, particularly when it comes to processing large and unstructured datasets. This is where metaheuristic algorithms come into play. They offer a promising solution by optimizing the recognition process through efficient exploration of various solution spaces. However, integrating these algorithms with HAR presents unique challenges, as they must be capable of identifying meaningful patterns without falling into the trap of overfitting [1].

Metaheuristic algorithms tackle several specific challenges within HAR. One significant issue is the high dimensionality of sensor data, which can overwhelm traditional algorithms. By ensuring robust feature selection, these algorithms enhance classification accuracy while reducing computational burdens. They also adapt to variations in human activities and device usage, adding complexity to data interpretation. Furthermore, advancements in sensor technologies such as depth sensors and wearable devices allow metaheuristic algorithms to leverage richer data for more nuanced activity inferences [4].

The broader implications of HAR technologies combined with metaheuristic algorithms are profound. Improved patient monitoring and tailored rehabilitation protocols can lead to better health outcomes and lower healthcare costs. In sports, optimized training and injury prevention strategies can prolong athletes' careers while enhancing their performance. In terms of

security, advanced surveillance capabilities can bolster public safety and protect critical infrastructure [7].

In our systematic review, we explored the technical complexities of employing metaheuristic algorithms in HAR by assessing their strengths and limitations in optimizing these systems. By analyzing recent advancements and considering future research directions, we aim to provide a comprehensive understanding of how metaheuristic algorithms can be applied within HAR. This exploration not only highlights the current state of the field but also serves as a foundation for future innovations that could unlock even more sophisticated HAR capabilities. As a response to curiosity towards the capability of metaheuristic algorithms in HAR, the following research questions (RQ) were developed as part of this work:

RQ1: How do metaheuristic algorithms enhance feature selection and improve the performance of machine learning models in human activity recognition compared to traditional feature selection methods?

RQ2: What are the most effective adaptations and enhancements for metaheuristic algorithms that have proven to be most effective for human activity recognition?

RQ3: What are the computational challenges faced by metaheuristic algorithms in human activity recognition?

RQ4: What are the emerging trends and significant research gaps in the application of metaheuristic algorithms for human activity recognition?

This research advances the field of metaheuristic algorithm in HAR with the following key contributions and implications:

1) Identify enhanced feature selection and optimization: This study demonstrates how metaheuristic algorithms improve feature selection and classification accuracy in HAR, reducing redundancy and computational costs while maintaining high recognition performance.

2) Insight into the advancement of hybrid metaheuristic approaches: By reviewing recent hybrid metaheuristic techniques, this research highlights their role in overcoming the limitations of individual algorithms, leading to more robust and efficient HAR systems.

3) Theoretical and practical insights: This study offers both theoretical contributions and practical considerations, helping researchers and practitioners navigate key challenges in optimizing HAR systems.

4) Bridging research and innovation: By analyzing emerging trends, this work serves as a foundation for future advancements, encouraging further exploration of novel strategies in HAR optimization.

The rest of this paper is organized as follows: Section II provides a review of related works. Section III presents the review method used for this research where it highlights the use of PRISMA approach. Section IV presents the results and discussions of this research to answer the research questions that have been raised. Finally, Section V presents the conclusion of the entire research work.

II. Related Work

The integration of metaheuristic algorithms into HAR systems has garnered significant scholarly attention, driven by the need to optimize feature selection, classification accuracy, and computational efficiency in complex sensor-driven environments. Prior studies have explored diverse applications of metaheuristics, though gaps persist in systematic evaluations of algorithmic adaptations and scalability challenges.

Helmi et al. [8] conducted a foundational analysis of nine metaheuristics, including Marine Predators Algorithm (MPA) for HAR and fall detection, demonstrating their efficacy in binary classification tasks. While their work established the viability of swarm intelligence for sensor data optimization, it focused narrowly on fall detection scenarios, leaving broader HAR applications underexplored. Similarly, Al-Wesabi et al. [9] employed Chaos Game Optimization to tune BiLSTM hyperparameters, achieving 93.9% accuracy on the UCI-HAD dataset, yet their methodology neglected feature selection dynamics critical for real-time deployment.

Recent advancements in hybrid metaheuristics have reshaped the field. Zhang et al. [10] introduced Oppositional and Chaos Particle Swarm Optimization (OCPSO), which elevated MI-1DCNN classification precision to 97.92%, outperforming conventional PSO-SVM models by 2.82%. Parallel developments by Tian et al. [11] utilized Improved Binary Glowworm Swarm Optimization to achieve 98.25% F-scores in ensemble learning frameworks, though their analysis omitted computational cost comparisons across algorithm classes.

Prior systematic analyses in metaheuristic research have emphasized breadth of algorithmic coverage over domain-specific methodological evaluation. Foundational works like Alorf's [7] meta-analysis provided comprehensive taxonomies of optimization techniques but offered limited assessment of their practical implementation efficacy in HAR contexts. Subsequent domain-focused reviews, such as Raj et al.'s [3] examination of healthcare applications, demonstrated rigorous vertical analysis while overlooking horizontal scalability across activity recognition domains. The field has seen notable technical innovations like Challa et al.'s [12] Rao-3 algorithm for BiLSTM optimization, which achieved benchmark performance across multiple datasets but left unexplored synergies with emerging deep learning architectures. This pattern reveals a persistent dichotomy in the literature between expansive algorithmic surveys and narrowly focused application studies, creating critical knowledge gaps in cross-domain performance evaluation and architectural hybridization potential, especially in the HAR domain.

III. Review Method

This review was conducted according to best practices in scoping reviews and is reported according to the PRISMA scoping review reporting guidelines [13].

A. Eligibility Criteria

This systematic review, conducted in Jan 2024, explored the application of metaheuristic algorithms in HAR, focusing on recent advancements from 2019 to 2024. The review included studies from the Scopus, IEEE Xplore, and Web of Science

databases, specifically targeting journal articles and conference proceedings with keywords related to HAR, such as "recognition," "estimation," "classification," and "detection," as well as metaheuristic optimization algorithms through terms like "application" and "implementation." Non-essential materials, including book chapters, reviews, and articles in press, were excluded to maintain academic rigor. The timeline was restricted to the last five years to ensure the review reflected current knowledge, avoiding outdated methodologies. Only final, peer-reviewed publications were considered, and non-peer-reviewed sources, like book series and trade journals, were excluded. The review focused solely on English-language studies for consistency and convenience, ensuring it accurately represents the current landscape of metaheuristic algorithm applications in HAR. The inclusion and exclusion criteria are summarized in Table I.

TABLE I. THE INCLUSION AND EXCLUSION CRITERIA

Criteria	Inclusion	Exclusion
Timeline	2019-2024	>2024
Document type	Journal article, Conference paper	Other than mentioned in the inclusion criteria]
Publication stage	Final	Article in press
Exact keywords	Human activity recognition, activity detection, motion recognition, behaviour recognition, recognition, estimation, classification, detection, metaheuristic, optimization, application, implementation	[Other than mentioned in the inclusion criteria]
Source type	Journal, Conference proceeding	Book series, book, trade journal
Language	English	[Other than English]

B. Information Sources

Three major academic databases were used as information sources: Scopus, IEEE Xplore, and Web of Science. These databases focused on studies of metaheuristic algorithms and their applications in HAR.

C. Search

This study systematically reviews recent literature on metaheuristic algorithms for HAR. We conducted a comprehensive search using Scopus, IEEE Xplore, and Web of Science, focusing on peer-reviewed studies published between 2019 and 2024. The search strategy employed precise keywords related to metaheuristics and HAR to identify relevant studies.

Table II provides the detailed search strategies employed across three major academic databases: Scopus, IEEE Xplore, and Web of Science. The advanced query strings were carefully crafted to capture a comprehensive range of studies focusing on metaheuristic algorithms and their applications in HAR.

In Scopus, the search query included terms related to various metaheuristic algorithms like "evolutionary algorithm," "swarm intelligence," and "particle swarm optimization," combined with keywords associated with HAR, such as "recognition," "classification," and "detection.".

TABLE II. ADVANCED SEARCH QUERY

Database	Advanced search query string
Scopus	TITLE-ABS-KEY ((metaheuristic* OR "evolutionary algorithm*" OR "swarm intelligence" OR "genetic algorithm*" OR "particle swarm optimization") AND (algorithm* OR optimization* OR method*) AND (application* OR implementation* OR use) AND ("human activity recognition" OR har OR "activity detection" OR "motion recognition" OR "behavior recognition") AND (recognition* OR estimation* OR classification* OR detection*)) AND PUBYEAR > 2018 AND PUBYEAR < 2025 AND (LIMIT-TO (SRCTYPE , "j") OR LIMIT-TO (SRCTYPE , "p")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (PUBSTAGE , "final")) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp"))
IEEE Xplore	("Document Title":metaheuristic* OR "Document Title":"evolutionary algorithms" OR "Document Title":"swarm intelligence" OR "Document Title":"genetic algorithms" OR "Document Title":"particle swarm optimization") AND ("All Metadata":algorithm* OR "All Metadata":optimization* OR "All Metadata":method*) AND ("All Metadata":application* OR "All Metadata":implementation* OR "All Metadata":use) AND ("All Metadata":"human activity recognition" OR "All Metadata":"activity detection" OR "All Metadata":"motion recognition" OR "All Metadata":"behavior recognition") AND ("All Metadata":recognition* OR "All Metadata":classification* OR "All Metadata":detection*)
Web Of Science	TS=((metaheuristic* OR "evolutionary algorithm*" OR "swarm intelligence" OR "genetic algorithm*" OR "particle swarm optimization") AND (algorithm* OR optimization* OR method*) AND (application* OR implementation* OR use) AND ("human activity recognition" OR HAR OR "activity detection" OR "motion recognition" OR "behavior recognition") AND (recognition* OR estimation* OR classification* OR detection*)) and 2024 and 2023 or 2022 or 2021 or 2020 or 2019 (Publication Years) and Article (Document Types) and English (Languages)

In Scopus, the search query included terms related to various metaheuristic algorithms like "evolutionary algorithm," "swarm intelligence," and "particle swarm optimization," combined with keywords associated with HAR, such as "recognition," "classification," and "detection.".

The query string used for IEEE Xplore follows a similar format, specifying document types such as conference papers and journal articles while filtering by publication stage to exclude in-progress works. To ensure consistent and comparable results across all databases, the same keywords related to metaheuristics and HAR were applied throughout.

In Web of Science, the search strategy also included combinations of relevant keywords and was refined further by applying filters for document type, publication years, and language. By aligning the search terms and filters across these databases, this review aimed to ensure a consistent and thorough collection of relevant literature. These query strings play a crucial role in ensuring the systematic review captures the most relevant and high-quality studies within the specified timeframe, providing a robust foundation for the subsequent analysis.

D. Data Extraction Process

Fig. 1 shows a PRISMA diagram for the systematic process we undertook to identify and select the most relevant studies for our review on the application of metaheuristic algorithms in HAR. Our initial search across Scopus, IEEE Xplore, and Web of Science databases produced 159 articles.

To refine our dataset, we used Mendeley Reference Manager to eliminate duplicates, resulting in 122 unique articles. We then conducted an initial screening based on the titles and abstracts. During this phase, we excluded 78 articles for various reasons: some had unavailable full texts (3 articles), others focused primarily on vision-based approaches (8 articles), and a significant number did not closely align with the specific focus of our review (67 articles).

Following this, 44 articles underwent a rigorous full-text review. This in-depth assessment resulted in the exclusion of 17 articles primarily due to methodological shortcomings or insufficient relevance to our review's objectives. Ultimately, 27 studies met the inclusion criteria and were qualitatively synthesized to gain insights into the application of metaheuristic algorithms in HAR. These 27 selected papers were then meticulously examined and assessed to extract relevant information aligned with each research question, guided by the rationale outlined in Table III.

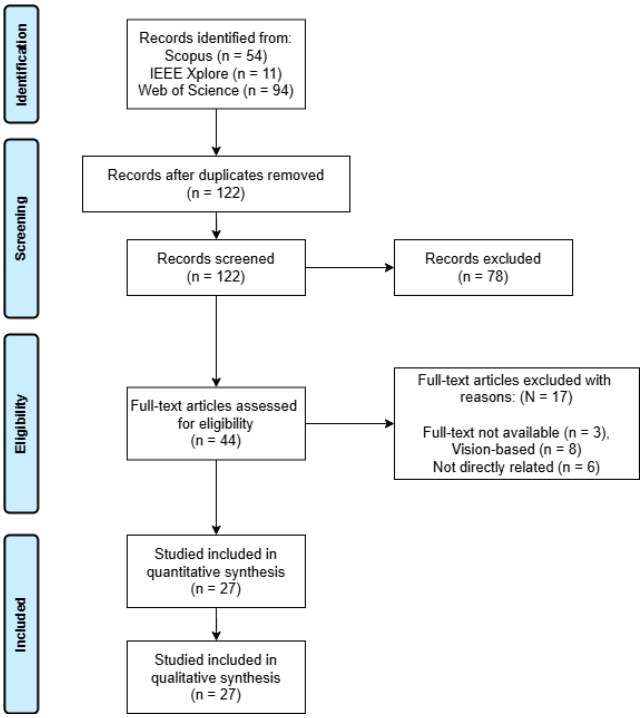


Fig. 1. PRISMA diagram.

TABLE III. RESEARCH QUESTIONS AND THEIR CORRESPONDING RATIONALES

Number	Research Questions	Rationale
RQ1	How do metaheuristic algorithms enhance feature selection and improve the performance of machine learning models in human activity recognition compared to traditional feature selection methods?	This question explores how advanced optimization techniques address persistent challenges in feature selection and model performance by reducing redundancy and improving predictive accuracy across various applications.
RQ2	What are the most effective adaptations and enhancements for metaheuristic algorithms that have proven to be most effective for human activity recognition?	This question investigates diverse adaptations and enhancements applied to metaheuristic algorithms, showcasing how these modifications enhance their effectiveness, adaptability, and efficiency for complex optimization problems.
RQ3	What are the computational challenges faced by metaheuristic algorithms in human activity recognition?	This question identifies significant computational challenges arising in the implementation of metaheuristic algorithms and offers insights into strategies for addressing these barriers effectively.
RQ4	What are the emerging trends and significant research gaps in the application of metaheuristic algorithms for human activity recognition?	This question uncovers transformative trends reshaping the field and highlights critical research gaps to provide a roadmap for future innovations and deepen understanding of these algorithms' potential.

IV. RESULTS AND DISCUSSIONS

In this section, we investigate all final selected articles (27 articles). The data is discussed to address the four mentioned research questions.

A. Improvement of Machine Learning Performance by Metaheuristic Algorithms over Traditional Methods Feature Selection (RQ1)

HAR rely heavily on the quality and relevance of the features extracted from sensor data. Traditional feature selection methods, while offering a certain level of effectiveness, can struggle with the complexities inherent in HAR tasks. Metaheuristic algorithms have emerged as powerful tools, offering significant advantages over traditional approaches. Fig. 2 shows the general workflow of how metaheuristic algorithm is being used to do feature selection. Feature selection happens after feature extraction, and metaheuristic algorithm will be applied during feature selection phase, though there are many approaches in applying metaheuristic algorithms during this phase.

One of the key strengths of metaheuristic algorithms is their ability to efficiently navigate large and complex search spaces. Unlike traditional feature selection methods, which may become stuck in local optima, metaheuristics employ a broader search strategy inspired by natural phenomena. Algorithms such as Genetic Algorithms (GA) and Grey Wolf Optimizers (GWO) mimic processes such as evolution and predator-prey interactions, respectively, to explore diverse regions within the feature space [14], [15]. This global search capability allows them to identify feature subsets that traditional methods might miss, potentially leading to superior classification performance in HAR applications.

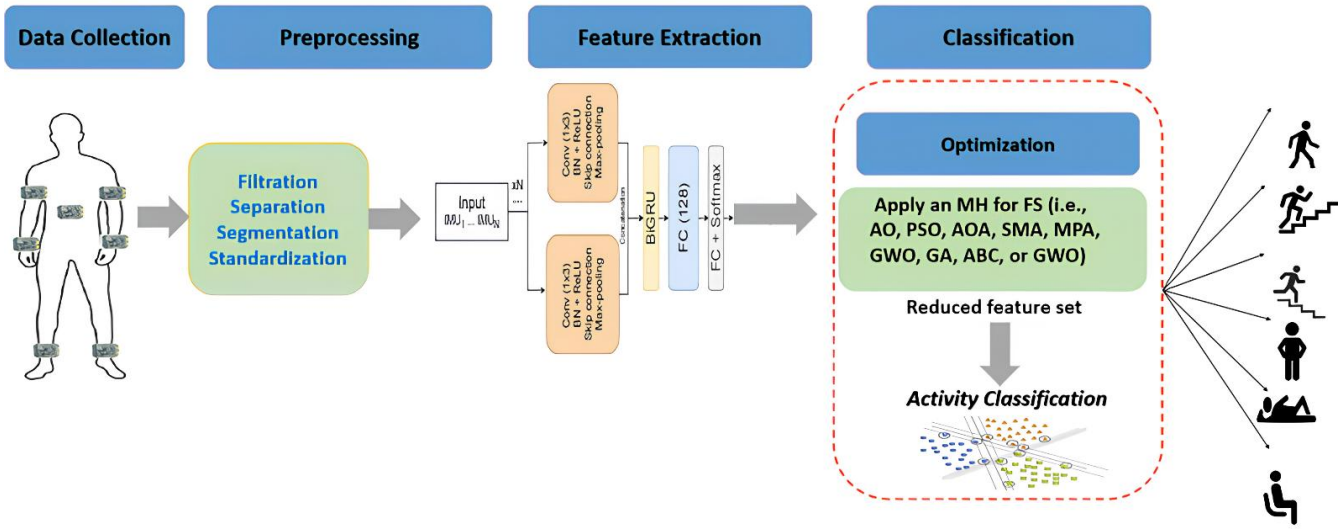


Fig. 2. Process workflow of metaheuristic feature selection for HAR [8].

Furthermore, metaheuristic algorithms play a key role in optimizing selected features by minimizing redundancy and identifying a small yet highly relevant set of features. This not only improves classification accuracy but also reduces the computational costs associated with managing large feature sets. These advantages make metaheuristic algorithms particularly valuable in critical medical applications, where both efficiency and accuracy are essential for reliable diagnoses. In contrast, traditional methods often rely on manual feature engineering or simpler optimization techniques, which may struggle to achieve the same level of efficiency or effectively capture the complex patterns found in sensor data used for HAR [16].

The fusion of heuristic algorithms and deep-learning approaches further boosts the benefits of feature selection in HAR. Deep-learning models demonstrate proficiency in unveiling intricate associations within data; however, they frequently require substantial quantities of high-quality features for optimal efficiency. By integrating metaheuristic algorithms into a deep learning framework, researchers can leverage their feature selection capabilities to pinpoint the most relevant features for a given task. This enhancement not only improves the training procedure but also results in models that exhibit greater resilience and generalizability [16]. For example, Alam et al. [17] has introduced NeuroHAR in 2024, which integrated Multilayer Perceptron (MLP) with Real-valued Genetic Algorithm (RGA). MLP performs the deep learning task of understanding and classifying complex human activity patterns, while RGA optimizes the hyperparameters by iterating through combinations to find the optimal model configuration. This synergy allows NeuroHAR to execute fewer models while exploring comprehensive hyperparameter ranges, making it computationally efficient while maintaining high prediction accuracy. Furthermore, the inherent adaptability of numerous metaheuristic algorithms enables them to manage the varied and high-dimensional characteristics of the sensor data commonly encountered in HAR applications. This adaptability makes them strong tools for researchers and developers to advance in HAR.

Metaheuristic algorithms have shown superior performance compared to traditional feature selection methods in effectively and adaptively handling high-dimensional data in HAR tasks. These approaches excel at navigating complex search spaces, enabling the discovery of optimal or near-optimal feature subsets that improve recognition performance. While earlier studies primarily relied on manual feature extraction and selection, the adoption of metaheuristic algorithms in HAR represents significant progress. By automating and refining this process, these algorithms address challenges related to feature interpretability and dimensionality, marking a substantial advancement in the field.

B. Effective Adaptation and Enhancement of Metaheuristic Algorithms for HAR (RQ2)

Researchers have investigated how metaheuristic algorithms can be tailored for systems related to HAR by analyzing a variety of approaches aimed at enhancing their effectiveness. The application of GAs has played a key role in the field of feature selection and classifier optimization, with significant achievements noted in Reweighted GAs, which have displayed remarkable accuracy in detecting daily activities. Furthermore, studies have shown their effectiveness in optimizing Support Vector Machines (SVMs) for HAR tasks [18], [19], [20], [21]. PSO has also displayed potential, especially in conjunction with SVMs (referred to as PSO-SVM), leading to enhancements in both detection accuracy and optimization efficiency [22]. Additional adaptations, including Quantum-behaved PSO (QPSO) and conventional PSO have shown advantages in refining kernel extreme learning machines (KELMs) and base extreme learning machines (ELMs), respectively, within the context of HAR [23].

Moreover, researchers have explored hybrid approaches that combine metaheuristic algorithms to address HAR challenges effectively. Hybrid techniques such as Hybrid Artificial Bee Colony and PSO (hABCPSO) and Oppositional and Chaos PSO (OCPSO) have demonstrated superior performance by leveraging the strengths of different algorithms to enhance feature selection, classifier parameter optimization, and overall recognition accuracy [24]. In addition to GAs and PSO, other metaheuristic algorithms, such as Chaos Game Optimization

(CGO), Binary Cuckoo Search (BCS), Rao-3 Optimization, Improved Binary Glowworm Swarm Optimization (IBGSO), and Gradient-based Grey Wolf Optimizer (GBOGWO), and Improved Cat Swarm Optimization (ICSO) have been adapted to fine-tune hyperparameters, explore search spaces efficiently, and improve the recognition performance of HAR [25], [26], [27].

Table V shows that metaheuristic algorithms have demonstrated superior and more efficient adaptation than conventional methodologies. Some enhanced metaheuristic algorithms even showed their superiority over the base-type metaheuristic algorithms. In contrast to traditional techniques that depend on manual feature engineering, metaheuristic algorithms automate the feature selection process, consequently enhancing the robustness and adaptability of HAR. Conversely, hybrid metaheuristic approaches, which combine the capabilities of algorithms such as GA and PSO, exhibit enhanced performance in terms of accuracy and computational efficiency when compared with conventional machine-learning models that typically require extensive data preprocessing and feature selection. Although traditional feature selection methods are simple and easy to understand, but metaheuristic algorithms provide a more powerful and flexible solution for complex HAR optimization, though they come with computational challenges.

C. Computational Challenges and Influence on Metaheuristic Algorithms for HAR (RQ3)

This research question discussed the key computational challenges encountered by metaheuristic algorithms in HAR and explored how these challenges influence their design and implementation. The computational challenges identified in the studied papers can be classified into two categories: Algorithmic Complexity, and Data Complexity as shown in Table IV.

TABLE IV. ALGORITHMIC AND DATA COMPLEXITY

Category	Computational Challenges
Algorithmic Complexity	Randomness, Premature Convergence, Complexity, Balancing Exploration and Exploitation, Slow Learning, High Computation Time
Data Complexity	Insufficient, Irrelevant, or Redundant Features, High Dimensionality, Computation Cost

Challenges related to algorithmic complexity include randomness, premature convergence, complexity, balancing exploration and exploitation, slow learning, and high computation time. Randomness is inherent in many metaheuristic algorithms and can lead to inconsistent results. Premature convergence occurs when the algorithm gets stuck in a local optimum and is unable to find the global optimum. Slow learning alludes to the prolonged duration required for the algorithm to effectively grasp and derive insights from the dataset, thus hindering the overall efficiency of the learning process. The high computational time stems from the fact that metaheuristic algorithms must evaluate a large number of possible solutions. On the other hand, challenges related to data complexity include insufficient, irrelevant, or redundant features, high dimensionality, and computation cost. High dimensionality points to the fact that the dataset comprises a wide range of features, thereby presenting a challenge for the algorithm to detect underlying patterns within the data. Moreover, the presence of insufficient, irrelevant, or redundant features within the dataset could also significantly hinder the efficiency of the algorithm, resulting in less-than-optimal outcomes. Table VI summarizes the computational challenges faced by the papers studied.

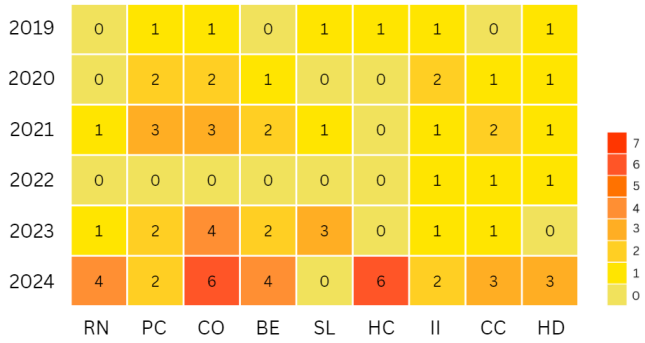


Fig. 3. Heatmap of computational challenges from 2019 to 2024.

From Fig. 3, it is proven that the computational challenges faced by metaheuristic algorithms in HAR have evolved over the years from 2019 to 2024. The most frequently encountered challenge appears to be balancing exploration and exploitation, particularly peaking in 2024 with four instances. This indicates a growing concern within the research community regarding the need for algorithms to effectively navigate the trade-off between exploration and exploitation of new solutions. This balance is crucial for optimizing algorithm performance without getting trapped in local optima or failing to converge on a global solution. The consistent presence of this challenge across the years underscores its significance in the field of HAR.

An increasing focus on complexity and high computation time has also emerged as a notable trend. This review has shown the complexity challenge grew steadily from 2019 to 2024, peaking in 2024 and maintaining a significant presence in subsequent years. Complexity was the most prominent challenge identified, accounting for 59.26% (16 out of 27) of the studies reviewed. This indicates that as metaheuristic algorithms grow more sophisticated, managing their complexity becomes increasingly critical. Similarly, high computation time saw a sudden spike in 2024, highlighting the significant computational burden of implementing advanced algorithms. This trend underscores the need for more efficient computational strategies or improved hardware to handle the intensive processing demands of metaheuristic algorithms in HAR applications.

Interestingly, some challenges such as randomness and premature convergence have relatively lower but varying instances across the years, with a noticeable spike in randomness in 2024. This variability might be due to the differing nature of the studies each year and the specific focus of the metaheuristic algorithms applied. The persistent challenge of insufficient, irrelevant, or redundant features shows that feature selection remains a critical area needing attention, impacting the accuracy and effectiveness of HAR.

TABLE V. METAHEURISTIC ALGORITHMS IN HUMAN ACTIVITY RECOGNITION

Metaheuristic Algorithm		Effective adaptation	Classifier	Performance
Base-type	**Enhanced-type**			
Particle Swarm Optimization (PSO)	Data not available (NA) [28], [29]	Optimal sensor combination through randomness and convergence.	SVM, RF, CNN, DNNLSTM, DeepCNN	SVM has the highest accuracy (95.13%)
	Oppositional and Chaos PSO (OCPSO) [10]	Helps in feature selection and improving recognition accuracy.	MI-1D-CNN	Results: • Precision: 97.92% • Recall: 97.85% • Accuracy: 97.81% • F1-score: 97.87%
	Quantum Behaved PSO (QPSO) [23], [30]	Enhances and optimizes kernel extreme learning machine (KELM) for HAR, resulting reduced misrecognized samples.	QPSO-KELM	QPSO-KELM shows highest accuracy at 91.3% for LDA and 96.2% for KDA [23] Accuracy of 96.4% [30]
	Hybrid Artificial Bee Colony and PSO (hABCPSO) [24]	Enhances local and global search capabilities for optimization.	Stacked AutoEncoder (SAE)	Has the most best performance values (17 out of 30 runs) over its competitors (ABC, DE, PSO, GA)
	PSO-Support Vector Machine (PSO-SVM) [22], [31]	PSO-SVM enhances detection accuracy, reliability, and optimization efficiency for SVM parameters.	PSO-SVM	Overall accuracy of 94.0% [22] Accuracy of 92.30%, F-Measure of 92.63% [31]
	Hierarchical PSO (H-PSO) [32]	Optimizes architecture-level parameters and layer-level hyperparameters simultaneously, enhancing the search for optimal configurations in CNNs.	1D-CNN	Accuracy on dataset: • UCI-HAR: 99.72% • PAMAP2: 96.03% • Daphnet Gait: 98.52% • Opportunity: 99.82%
	Adaptive Binary PSO (ABPSO) [33]	ABPSO utilizes a self-adaptive operator pool to enhance the feature selection process.	SVM & KNN	Accuracy: ReliefF: 95.62% (with 293 features selected) mRMR: 95.80% (with 201 features selected)
Biogeography Based Optimization (BBO)	Reweighted GA (rGA) [19]	Helps achieve high recognition accuracy of daily activities.	Reweighted Genetic Algorithm (rGA)	Accuracy on dataset): • CMU-MMAC: 88% • WISDM: 88.75% • IMSB: 83.33%
Bee Swarm Optimization (BSO)	BSO with deep Q-network (BAROQUE) [14]	BAROQUE lbalances exploitation and exploration for feature searching. It provides self-organization and self-adaptation capabilities for optimization.	KNN	Performance (accuracy on dataset): • UCI-HAR:98.41%
Cuckoo Search (CS)	NA [34]	Effectively contributed om high predictive accuracy due to its global convergence and efficient search space exploration.	NA	The accuracy value is 93.77% .
	CS with Recursive Feature Elimination (CSRFE) [21]	Optimizes the feature selection process, significantly reducing the number of features while maintaining or improving classification accuracy, minimizing temporal complexity in HAR systems.	SVM, RF, LR	Accuracy by classifier: RF: 96.76% SVM: 96.23% LR: 96.16%
Rao-3 Optimization	NA [12]	Optimization technique for ideal hyperparameters value to improve recognition performance.	BiLSTM	Accuracy on dataset: • PAMAP2: 94.91% • UCI-HAR: 97.11% • MHEALTH: 99.25%
Glowworm Swarm Optimization (GSO)	Improved Binary GSO (IBGSO) [11]	Enhances learning by selects a superior subset for ensemble pruning to find optimal subensemble models.	IBGSO	• Precision: 98.25% • Recall: 98.17% • Accuracy: 98.25% • F-score: 97.94%
Grey Wolf Optimizers (GWO)	Gradient-based Optimization & GWO (GBOGWO) [16]	Improves performance by balancing exploration and exploitation stages.	SVM	Mean accuracy: 98.87%
Chaos Game Optimization (CGO)	NA [9]	Fine-tunes BiLSTM hyperparameters for enhanced performance.	BiLSTM	UCI-HAR dataset: • Precision: 77% - 80.1% • Recall: 75.9% - 79.3% • Accuracy 92.0% - 93.2% • F-score: 76.0% - 79.4% UCI-HAD dataset:

				• Precision: 81.4% • Recall: 81.1% • Accuracy: 93.9% • F-score: 81.0%
Genetic Algorithm (GA)	NA [15], [18], [20], [26], [35], [36]	Used for selecting important features to improve classification performance while keeping the model small. It helps increase accuracy by filtering out noise and is also effective in optimizing fuzzy logic systems.	Deep Neural Decision Forest [18] KNN, SVM, RF [15] NA [20] SVM & RF [36] DCNN-LSTM [26] KNN [35]	[18] Accuracy: • ExtraSensory: 88.25% • Sussex-Huawei: 96.00% [15] F-measure: • KNN: 98.2% • SVM: 98.2% • RF: 97.6% [20] Accuracy: 99.1% [26] Result: • F1 Score: 98.89% • Average Recall: 99.01% • Average Precision: 98.90% • Total Accuracy: 99.92% [35] Accuracy using MHEALTH dataset is the highest.
	GA with a centroid-based clustering approach [36]	Helps in managing data efficiently while retaining high classification accuracy.	SVC, RF, KNN	Accuracy on dataset: • UCI-HAR: 93.45% • WISDM: 72.8%
	Non-dominated Sorting GA II (NSGA-II) [37]	Using a multi-objective optimization approach that simultaneously evolves LSTMs for classification accuracy.	LSTM	Classification accuracy on SMARTPHONE dataset: • With NGSA-II: 99.03% • Without NGSA-II: 97.69%
	Real-valued GA (RGA) with Multilayer Perceptron (MLP) [17]	Dynamic optimization of network architectures and hyperparameters, which allows for better handling of task complexities compared to traditional methods.	MLP	Accuracy based on model: NeuroHAR: 89.91% Grid Search:84.04% Notes: NeuroHAR is the proposed model (RGA with MLP).
Cat Swarm Optimization (CSO)	Improved CSO (ICSO) [27]	Improves CNN parameter tuning for HAR tasks.	CNN	ICSO-CNN achieved an accuracy of 99.79%, outperforming other methods such as CNN with Long Short-Term Memory (CNN-LSTM), PSO based CNN (PSO-CNN), and CNN-BiLSTM.

TABLE VI. COMPUTATIONAL CHALLENGES

ID	Paper	Computational Challenges								
		RN	PC	CO	BE	SL	HC	II	CC	HD
ID1	[28]	✓	✓							
ID2	[29]		✓	✓	✓					
ID3	[10]		✓			✓				
ID4	[30]		✓							
ID5	[23]					✓	✓	✓		
ID6	[24]								✓	
ID7	[22]		✓						✓	
ID8	[31]			✓						✓
ID9	[19]		✓	✓	✓					
ID10	[18]			✓	✓	✓				
ID11	[15]			✓	✓					
ID12	[20]			✓				✓		✓
ID13	[36]			✓						
ID14	[26]	✓								
ID15	[35]							✓	✓	
ID16	[14]			✓		✓				
ID17	[25]			✓						
ID18	[34]		✓					✓		
ID19	[12]			✓		✓				
ID20	[11]							✓	✓	✓
ID21	[16]		✓		✓			✓	✓	✓
ID22	[33]	✓		✓	✓		✓	✓	✓	✓
ID23	[21]	✓		✓	✓		✓	✓	✓	✓
ID24	[37]		✓	✓	✓		✓			
ID25	[27]	✓		✓			✓			✓
ID26	[17]			✓			✓		✓	
ID27	[32]	✓	✓	✓	✓		✓			

Abbreviations: RN = Randomness, PC = Premature Convergence, CO = Complexity, BE = Balancing Exploration and Exploitation, SL = Slow Learning, HC = High Computation Time, II = Insufficient, Irrelevant, or Redundant Features, CC = Computation Cost, HD = High Dimensionality.

In HAR, metaheuristic algorithms face various computational challenges that shape their design and

implementation. Table VII summarizes these challenges across 27 reviewed papers, highlighting issues such as complexity and the need for balancing exploration and exploitation. Hybrid models that combine multiple algorithms are essential to address these challenges by leveraging their individual strengths. For example, combining GA with PSO achieves a balance between exploration and exploitation, while hybrid approaches like ABC with PSO tackle complexity and premature convergence by using ABC for local search and PSO for global optimization.

High-dimensional data and redundant or irrelevant features significantly affect classification accuracy, with 55.56% (15 out of 27) of studies identifying this as a critical challenge. To address this, efficient feature selection techniques, such as using QPSO to optimize feature sets, have improved KELMs by retaining only the most discriminative features. Additionally, computational overhead, including high processing times and costs, poses scalability challenges for HAR systems. One solution has been implementing BBO within scalable architectures, dynamically optimizing resource allocation to manage system load and reduce computational demands.

The efficiency of feature selection and extraction is another critical area influenced by high dimensionality and slow learning processes, impacting 59.26% (16 out of 27) of the studies. IBGSO, for instance, enhances the learning process by identifying high-value features and reducing redundancy, thereby increasing overall system efficiency. As HAR systems increasingly demand real-time processing capabilities, algorithmic designs must focus on minimizing computational burdens without compromising accuracy, ensuring these systems are both effective and scalable for practical applications.

Additionally, intensive numerical calculations demand high computational resources; despite not being much, 4 out of the 27 papers studied have impacted the feasibility and scalability of designs. The iterative nature of the optimization process necessitates frequent updates and fitness value calculations, contributing to computational intensity. Addressing optimal training to prevent underfitting or overfitting entails computational hurdles, prompting the design of early stopping mechanisms. The transition from handcrafted feature extraction to deep learning techniques, driven by the limitations of traditional machine learning, introduces further computational demands [35]. The real-time processing requirements inherent in HAR pose significant computational challenges, guiding the selection of sensors and algorithms. Also, from Fig. 3 we can see that from 2019 to 2024, most of the challenges faced were algorithmic complexity compared to data complexity. In six years since 2019, the cumulative computational challenges faced by the papers studied were 76, and algorithmic complexity contributed 53 (69.74%) from the total, while data complexity contributed 23 (30.26%) in total, though both categories showed an increasing pattern over the years and peaked in 2024.

Metaheuristic algorithms offer a flexible strategy for optimizing intricate issues in HAR, unlike conventional techniques that might necessitate explicit mathematical formulations. Traditional machine learning approaches typically rely on predefined models and assumptions, limiting their ability to adapt to the dynamic nature of human activities. In contrast, metaheuristic algorithms seek solutions based on heuristic principles, enabling more resilient HAR solutions. The development of metaheuristic algorithms is guided by the need to balance exploration and exploitation to efficiently navigate the search space of HAR problems, a challenge less prominent in traditional optimization methods. Computational obstacles such as dimensionality and local optima are more effectively tackled by metaheuristics through population-based search strategies, a capability that traditional methods may struggle to achieve.

TABLE VII. INFLUENCE OF COMPUTATIONAL CHALLENGES TO DESIGN

ID	Paper	Influence on Design			
		Hybrid model	Low Classification Accuracy	System scalability	Feature selection and extraction efficiency
ID1	[28]				✓
ID2	[29]		✓		
ID3	[10]	✓	✓		✓
ID4	[30]	✓	✓		
ID5	[23]	✓	✓		✓
ID6	[24]	✓		✓	
ID7	[22]	✓	✓		
ID8	[31]	✓			✓
ID9	[19]	✓			✓
ID10	[18]		✓	✓	
ID11	[15]		✓		✓
ID12	[20]				✓
ID13	[36]		✓		
ID14	[26]		✓		✓
ID15	[35]		✓		✓
ID16	[14]				✓
ID17	[25]		✓		✓
ID18	[34]	✓			
ID19	[12]				✓
ID20	[11]	✓	✓		✓
ID21	[16]	✓	✓		✓
ID22	[33]	✓	✓	✓	✓
ID23	[21]	✓			✓
ID24	[37]	✓			
ID25	[27]	✓			
ID26	[17]	✓	✓	✓	
ID27	[32]	✓			

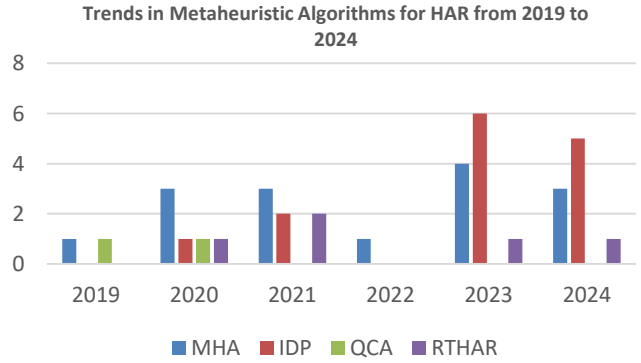
D. Trends and Research Gaps of Metaheuristic Algorithm for HAR (RQ4)

This research question explores the trends and research gaps in the use of metaheuristic algorithms for HAR between 2019 and 2024. A summary of the findings is provided in Table VIII. One notable trend is the rise of hybrid metaheuristic algorithms. These hybrid approaches, such as combining GAs with PSO, capitalize on the strengths of multiple algorithms to strike a balance between exploration and exploitation, tackle complexity, and improve convergence rates. For instance,

hybrid models that integrate ABC and PSO algorithms have demonstrated potential in enhancing both local and global search capabilities. Another emerging trend is the integration of deep learning techniques with metaheuristic algorithms. Deep learning models, particularly CNNs and Recurrent Neural Networks (RNNs), are being employed to automate feature extraction and boost classification accuracy. In addition, the application of quantum computing in this domain is gaining traction, with QPSO showing promise for optimizing processes in high-dimensional spaces. The growing demand for real-time processing in HAR has also boosted the development of more efficient and scalable algorithms. For instance, techniques like BBO are being applied to dynamically optimize resource allocation, ensuring smooth and effective operations in real-time environments.

TABLE VIII. SUMMARY OF EMERGING TRENDS IN METAHEURISTIC ALGORITHMS FOR HAR

Trend	Description	Examples
Hybrid Metaheuristic Algorithms	Combining different algorithms to balance exploration and exploitation and improve convergence rates.	GA + PSO, ABC + PSO
Integration with Deep Learning	Using deep learning models to automate feature extraction and improve classification accuracy.	CNNs, RNNs combined with metaheuristics
Quantum Computing Applications	Enhancing optimization processes in high-dimensional spaces.	QPSO
Real-Time Processing in HAR Systems	Developing scalable algorithms for dynamic optimization in real-time environments.	BBO for resource allocation



Abbreviations: MHA = Hybrid Metaheuristic Algorithms, IDP = Integration with Deep Learning, QCP = Quantum Computing Applications, RTHAR = Real-Time Processing in HAR Systems

Fig. 4. Trends in Metaheuristic algorithms for HAR from 2019 to 2024.

Fig. 4 shows key trends in metaheuristic algorithms for HAR from 2019 to 2024, focusing on four major areas: Hybrid Metaheuristic Algorithms (MHA), Integration with Deep Learning (IDP), Quantum Computing Applications (QCA), and Real-Time Processing (RTHAR). Research on MHA has remained steady, peaking with four publications in 2023 before slightly dropping to three in 2024. This reflects the continued importance of hybrid approaches in tackling complex HAR challenges. IDP has emerged as the dominant focus, with publications surging to six in 2023 and maintaining strong momentum with five in 2024. This trend highlights the growing synergy between deep learning and HAR systems, showcasing its potential to enhance automation and accuracy. QCA, while still in the experimental stage, has seen consistent yet minimal activity, with one publication annually through 2020. This indicates ongoing exploration of quantum computing's potential in HAR. RTHAR, which focuses on real-time processing techniques, saw moderate activity with a peak in 2021 before stabilizing at one publication per year through 2024. This trend suggests a gradual shift toward standardization in real-time solutions for HAR. The sustained high levels of research in MHA and IDP through 2024 demonstrate the community's strong focus on advancing hybrid models and integrating deep learning, while interest in experimental and specialized methods like QCA and RTHAR continues to provide avenues for innovation.

Significant research gaps persist in HAR, as shown in Table IX despite the trends. A key challenge lies in the scalability and efficiency of metaheuristic algorithms when handling large datasets typical in HAR applications [30]. Many of these algorithms become computationally expensive with increasing data sizes, limiting their applicability in big data scenarios. Additionally, there is a lack of comprehensive studies comparing the effectiveness of various metaheuristic algorithms across HAR contexts. While individual algorithms have been explored, systematic comparisons across scenarios, such as sensor-based activity recognition in smart homes versus smartphone-based recognition during exercise, are needed to identify the most suitable options for specific applications. Combining deep learning with metaheuristic algorithms for hyperparameter optimization has improved HAR systems but understanding how these models interpret selected features remains a major gap. This lack of insight hampers our ability to evaluate the importance of features in accurately recognizing activities. Most studies focus on improving efficiency and accuracy, often overlooking computational costs and real-time feasibility, particularly in wearable sensor-based HAR systems. Addressing these challenges is crucial for developing practical, effective solutions that enhance seamless human activity recognition.

TABLE IX. SUMMARY OF RESEARCH GAPS IN METAHEURISTIC ALGORITHMS FOR HAR

Research Gap	Description
Scalability and Efficiency	Metaheuristic algorithms often struggle with computational costs as data sizes increase in HAR applications.
Lack of Comprehensive Comparisons	Insufficient studies comparing different metaheuristic algorithms across various HAR contexts.
Interpretability of Metaheuristic-Optimized Models	Limited research on understanding the significance of features selected by metaheuristic algorithms in HAR.
Real-Time Feasibility	Need for lightweight algorithms that can operate efficiently in real-time, especially in wearable sensor-based HAR.
Adaptability to Dynamic Environments	Lack of dynamic adaptive algorithms that can adjust to changing human activity patterns.
Integration with IoT and Edge Computing	Underexplored integration of metaheuristics with IoT devices and edge computing for practical HAR applications.

V. CONCLUSION

This systematic review underscores the transformative role of metaheuristic algorithms in advancing HAR, synthesizing insights across four research questions. The analysis revealed that metaheuristic algorithms such as GWO significantly enhance feature selection and classification accuracy (RQ1), achieving up to 96.00% recognition rates by efficiently navigating high-dimensional data and reducing redundancy. Hybrid and enhanced adaptations like OCPSO and Quantum-inspired PSO emerged as pivotal solutions (RQ2), boosting detection accuracy by 2.82% and optimizing models such as QPSO-KELM to 96.2% precision. However, computational challenges persist (RQ3), in which 69.74% are algorithmic complexity, and 30.26% are data complexity. Emerging trends (RQ4) highlight a shift toward hybrid metaheuristic algorithms, integration with deep learning, and quantum-inspired techniques which had peaked in 2024, yet gaps in scalability, real-time feasibility, and interpretability demand urgent attention. Collectively, metaheuristic algorithms demonstrate immense potential to revolutionize HAR not limited to healthcare, sports, and security, but future innovations must prioritize lightweight, adaptive frameworks to bridge practical implementation gaps and unlock their full societal impact.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Ministry of Higher Education Malaysia for funding the research via Fundamental Research Grant Scheme (FRGS), under grant number FRGS/1/2022/ICT02/UNIMAS/03/1. Additionally, we acknowledge the assistance provided by ChatGPT and Perplexity in the writing of this paper.

REFERENCES

[1] P. P. Ariza-Colpas *et al.*, "Human Activity Recognition Data Analysis: History, Evolutions, and New Trends," *Sensors*, vol. 22, no. 9, 2022, doi: 10.3390/s22093401.

[2] V. Dentamaro, V. Gattulli, D. Impedovo, and F. Manca, "Human activity recognition with smartphone-integrated sensors: A survey," *Expert Syst Appl*, vol. 246, p. 123143, 2024, doi: https://doi.org/10.1016/j.eswa.2024.123143.

[3] A. K. Ravi Raj, "An improved human activity recognition technique based on convolutional neural network," *National Library of Medicine*, vol. 13, no. 1, p. 22581, 2023, doi: 10.1038/s41598-023-49739-1.

[4] M. H. Arshad, M. Bilal, and A. Gani, "Human Activity Recognition: Review, Taxonomy and Open Challenges," Sep. 01, 2022, *MDPI*. doi: 10.3390/s22176463.

[5] A. M. Helmi, M. A. A. Al-qaness, A. Dahou, and M. Abd Elaziz, "Human activity recognition using marine predators algorithm with deep learning," *Future Generation Computer Systems*, vol. 142, pp. 340–350, 2023, doi: 10.1016/j.future.2023.01.006.

[6] S. Raja Sekaran, P. Y. Han, and O. S. Yin, "Smartphone-based human activity recognition using lightweight multiheaded temporal convolutional network," *Expert Syst Appl*, vol. 227, p. 120132, 2023, doi: https://doi.org/10.1016/j.eswa.2023.120132.

[7] A. Alorf, "A survey of recently developed metaheuristics and their comparative analysis," *Eng Appl Artif Intell*, vol. 117, p. 105622, Jan. 2023, doi: 10.1016/j.engappai.2022.105622.

[8] M. Al-qaness, A. Helmi, A. Dahou, and M. Elsayed Abd Elaziz, "The Applications of Metaheuristics for Human Activity Recognition and Fall Detection Using Wearable Sensors: A Comprehensive Analysis," *Biosensors (Basel)*, vol. 12, p. 821, Jan. 2022, doi: 10.3390/bios12100821.

[9] F. N. Al-Wesabi *et al.*, "Design of Optimal Deep Learning Based Human Activity Recognition on Sensor Enabled Internet of Things Environment," *IEEE Access*, vol. 9, pp. 143988–143996, 2021, doi: 10.1109/ACCESS.2021.3112973.

[10] Y. Zhang, X. Yao, Q. Fei, and Z. Chen, "Smartphone sensors-based human activity recognition using feature selection and deep decision fusion," *IET Cyber-Physical Systems: Theory & Applications*, vol. 8, no. 2, pp. 76–90, Jan. 2023, doi: 10.1049/cps2.12045.

[11] Y. Tian, J. Zhang, Q. Chen, and Z. Liu, "A Novel Selective Ensemble Learning Method for Smartphone Sensor-Based Human Activity Recognition Based on Hybrid Diversity Enhancement and Improved Binary Glowworm Swarm Optimization," *IEEE Access*, vol. 10, pp. 125027–125041, 2022, doi: 10.1109/ACCESS.2022.3225652.

[12] S. K. Challa, A. Kumar, V. B. Semwal, and N. Dua, "An optimized deep learning model for human activity recognition using inertial measurement units," *Expert Syst*, vol. 40, no. 10, p. e13457, 2023, doi: 10.1111/exsy.13457.

[13] M. J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, pp. e112–e112, 2021, doi: 10.1136/bmj.n71.

[14] C. Fan and F. Gao, "Enhanced Human Activity Recognition Using Wearable Sensors via a Hybrid Feature Selection Method," *Sensors*, vol. 21, no. 19, p. 6434, Sep. 2021, doi: 10.3390/s21196434.

[15] J. Chen, Y. Sun, and S. Sun, "Improving Human Activity Recognition Performance by Data Fusion and Feature Engineering," *Sensors*, vol. 21, no. 3, p. 692, Jan. 2021, doi: 10.3390/s21030692.

[16] A. M. Helmi, M. A. A. Al-qaness, A. Dahou, R. Damaševičius, T. Krilavičius , and M. A. Elaziz, "A Novel Hybrid Gradient-Based Optimizer and Grey Wolf Optimizer Feature Selection Method for Human Activity Recognition Using Smartphone Sensors," *Entropy*, vol. 23, no. 8, p. 1065, Aug. 2021, doi: 10.3390/e23081065.

[17] F. Alam, P. Plawiak, A. Almaghthawi, M. R. C. Qazani, S. Mohanty, and A. Roohallah Alizadehsani, "NeuroHAR: A Neuroevolutionary Method for Human Activity Recognition (HAR) for Health Monitoring," *IEEE Access*, vol. 12, pp. 112232–112248, 2024, doi: 10.1109/ACCESS.2024.3441108.

[18] A. Alazeb *et al.*, "Intelligent Localization and Deep Human Activity Recognition through IoT Devices," *Sensors*, vol. 23, no. 17, p. 7363, 2023, doi: 10.3390/s23177363.

[19] M. Batool, A. Jalal, and K. Kim, "Telemonitoring of Daily Activity Using Accelerometer and Gyroscope in Smart Home Environments," *Journal of Electrical Engineering & Technology*, vol. 15, no. 6, pp. 2801–2809, Jan. 2020, doi: 10.1007/s42835-020-00554-y.

[20] Z. Huang, Q. Niu, and S. Xiao, "Human Behavior Recognition Based on Motion Data Analysis," *Intern J Pattern Recognit Artif Intell*, vol. 34, no. 09, p. 2056005, Jan. 2020, doi: 10.1142/S0218001420560054.

[21] R. Saifi, A. Achroufene, H. Attoumi, and L. Souici, "A Hybrid Feature Selection Method for Human Activity Recognition," in *PAIS 2024 - Proceedings: 6th International Conference on Pattern Analysis and Intelligent Systems*, 2024. doi: 10.1109/PAIS62114.2024.10541202.

[22] H. Wang and L. Liu, "Characterization of human motion by the use of an accelerometer-based detection system," *Instrum Sci Technol*, vol. 49, no. 1, pp. 55–64, Jan. 2021, doi: 10.1080/10739149.2020.1779083.

[23] [23] Y. Tian, J. Zhang, L. Chen, Y. Geng, and X. Wang, "Single Wearable Accelerometer-Based Human Activity Recognition via Kernel Discriminant Analysis and QPSO-KELM Classifier," *IEEE Access*, vol. 7, pp. 109216–109227, 2019, doi: 10.1109/ACCESS.2019.2933852.

[24] T. Ozcan and A. Basturk, "Human action recognition with deep learning and structural optimization using a hybrid heuristic algorithm," *Cluster Comput*, vol. 23, no. 4, pp. 2847–2860, Jan. 2020, doi: 10.1007/s10586-020-03050-0.

[25] F. N. Al-Wesabi *et al.*, "Design of Optimal Deep Learning Based Human Activity Recognition on Sensor Enabled Internet of Things Environment," *IEEE Access*, vol. 9, pp. 143988–143996, 2021, doi: 10.1109/ACCESS.2021.3112973.

[26] S. Jameer and H. Syed, "A DCNN-LSTM based human activity recognition by mobile and wearable sensor networks," *Alexandria Engineering Journal*, vol. 80, pp. 542–552, 2023, doi: 10.1016/j.aej.2023.09.013.

[27] Y. Chanti, A. H. Shnain, R. Banoth, R. V. S. S. B. Rupavath, and C. Sushama, "Human Activity Recognition Using Improved Cat Swarm Optimization Algorithm and Convolutional Neural Network," in *2nd IEEE International Conference on Networks, Multimedia and Information Technology, NMITCON 2024*, 2024. doi: 10.1109/NMITCON62075.2024.10699228.

[28] C. Xia and Y. Sugiura, "Wearable Accelerometer Layout Optimization for Activity Recognition Based on Swarm Intelligence and User Preference," *IEEE Access*, vol. 9, pp. 166906–166919, Jan. 2021, doi: 10.1109/ACCESS.2021.3134262.

[29] R. T. Al_Hassani and D. C. Atilla, "Human Activity Detection Using Smart Wearable Sensing Devices with Feed Forward Neural Networks and PSO," *Applied Sciences (Switzerland)*, vol. 13, no. 6, 2023, doi: 10.3390/app13063716.

[30] Y. Tian, X. Wang, Y. Geng, Z. Liuand, and L. Chen, "Inertial sensor-based human activity recognition via ensemble extreme learning machines optimized by quantum-behaved particle swarm," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 2, pp. 1443–1453, Feb. 2020, doi: 10.3233/JIFS-179507.

[31] Y. Zhu, J. Yu, F. Hu, Z. Li, and Z. Ling, "Human activity recognition via smart-belt in wireless body area networks," *Int J Distrib Sens Netw*, vol. 15, no. 5, p. 155014771984935, Jan. 2019, doi: 10.1177/1550147719849357.

[32] S. Ankalaki and M. N. Thippeswamy, "Optimized Convolutional Neural Network Using Hierarchical Particle Swarm Optimization for Sensor Based Human Activity Recognition," *SN Comput Sci*, vol. 5, no. 5, 2024, doi: 10.1007/s42979-024-02794-5.

[33] Y. Zhou, R. Wang, Y. Wang, S. Sun, J. Chen, and X. Zhang, "A Swarm Intelligence Assisted IoT-Based Activity Recognition System for Basketball Rookies," *IEEE Trans Emerg Top Comput Intell*, vol. 8, no. 1, pp. 82–94, 2024, doi: 10.1109/TETCI.2023.3319432.

[34] M. Kaur, G. Kaur, P. K. Sharma, A. Jolfaei, and D. Singh, "Binary cuckoo search metaheuristic-based supercomputing framework for human behavior analysis in smart home," *J Supercomput*, vol. 76, no. 4, pp. 2479–2502, Jan. 2020, doi: 10.1007/s11227-019-02998-0.

[35] A. Sarkar, S. K. S. Hossain, and R. Sarkar, "Human activity recognition from sensor data using spatial attention-aided CNN with genetic algorithm," *Neural Comput Appl*, vol. 35, no. 7, pp. 5165–5191, Mar. 2023, doi: 10.1007/s00521-022-07911-0.

[36] A. K. Panja, A. Rayala, A. Agarwala, S. Neogy, and C. Chowdhury, "A hybrid tuple selection pipeline for smartphone based Human Activity Recognition," *Expert Syst Appl*, vol. 217, May 2023, doi: 10.1016/j.eswa.2023.119536.

[37] R. A. Viswambaran, M. Nekooei, G. Chen, and B. Xue, "Evolutionary Design of Long Short Term Memory Networks and Ensembles through Genetic Algorithms," in *2024 IEEE Congress on Evolutionary Computation (CEC)*, Jun. 2024, pp. 1–8. doi: 10.1109/CEC60901.2024.10612126.

Bridging Data and Clinical Insight: Explainable AI for ICU Mortality Risk Prediction

Ali H. Hassan[1], Riza bin Sulaiman[2], Mansoor Abdulhak[3], Hasan Kahtan[4]
Institute of IR 4.0 (IIR4.0), Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia[1, 2]
College of Computer and Cyber Sciences, University of Prince Mugrin, Madinah 41499, Saudi Arabia[1]
Computer Science Department at University of Oklahoma, Norman, Oklahoma 73019[3]
Cardiff School of Technologies, Cardiff Metropolitan University, Cardiff CF5 2YB[4]

***Abstract*—Despite advancements in machine learning within healthcare, the majority of predictive models for ICU mortality lack interpretability, a crucial factor for clinical application. The complexity inherent in high-dimensional healthcare data and models poses a significant barrier to achieving accurate and transparent results, which are vital in fostering trust and enabling practical applications in clinical settings. This study focuses on developing an interpretable machine learning model for intensive care unit (ICU) mortality prediction using explainable AI (XAI) methods. The research aimed to develop a predictive model that could assess mortality risk utilizing the WiDS Datathon 2020 dataset, which includes clinical and physiological data from over 91,000 ICU admissions. The model's development involved extensive data preprocessing, including data cleaning and handling missing values, followed by training six different machine learning algorithms. The Random Forest model ranked as the most effective, with its highest accuracy and robustness to overfitting, making it ideal for clinical decision-making. The importance of this work lies in its potential to enhance patient care by providing healthcare professionals with an interpretable tool that can predict mortality risk, thus aiding in critical decision-making processes in high-acuity environments. The results of this study also emphasize the importance of applying explainable AI methods to ensure AI models are transparent and understandable to end-users, which is crucial in healthcare settings.**

***Keywords*—Explainable AI; healthcare; machine learning; predictive model**

I. Introduction

ICUs (Intensive Care Units) are specialized units that constantly monitor and treat patients with severe or potentially fatal illnesses. Due to the complexity of the ICU environments and the high-stake nature of patient care, accurate prediction of mortality risk in critically ill patients is a key aspect of adequate healthcare management [1], [2]. Machine learning algorithms and predictive models have shown great promise in addressing these needs by mining large amounts of clinical data to predict the risk of patient mortality. Predicting accurately has a significant effect on clinical decisions, allocation of resources, and the management of patients [3]. Machine learning methods have made tremendous progress and opened up new avenues for mortality prediction in the clinical world. To produce predictive insights, machine learning models are trained on complex datasets containing patient demographics, clinical measurements, and historical health records. Machine learning algorithms can explore more complicated relationships between variables than traditional statistical methods allow [4]. One example is through models utilizing extensive, large-scale electronic health record (EHR) data that has improved prediction performance for patient outcomes [5], [6].

Despite advances in machine learning, several challenges remain in accurately predicting mortality risk in ICU patients. For one, low-dimensional and high-dimensional complex healthcare data can increase the risk of overfitting and decrease the generalization performance of models if handled poorly [7], [8]. Also, the explainability of machine learning models is an important issue; medical professionals must have precise and reliable predictions to implement AI tools in clinical practice [9], [10], [11]. Black-box models can be opaque and difficult to interpret, preventing adoption in high-stakes medical contexts. Explainable Artificial Intelligence (XAI) frameworks, such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), address this issue by providing transparent and interpretable explanations of model predictions, enhancing trust and usability [12], [13].

To mitigate these challenges for predictive models, the primary goal of this study is to develop an interpretable mortality risk prediction model incorporating machine learning algorithms with explainable artificial intelligence (XAI) methods. The dataset utilized in this study is derived from the 2020 Women in Data Science (WiDS) Datathon, which provides a rich profile of ICU patients, including various physiological and clinical variables [14]. The use of Explainable AI (XAI) methodologies, such as SHAP (Shapley Additive explanations) and LIME (Local Interpretable Model-agnostic Explanations), is intended to provide transparent insights into the model's decision-making process, thereby facilitating better understanding and trust in the predictive outcomes. This study addresses the limitations of previous studies by combining interpretability, accuracy, and clinical application, covering the way for reliable and actionable AI solutions in critical care settings.

The remainder of this study is structured as follows: Section II presents the background and motivation for the study. Section III provides a detailed description of the dataset's characteristics and explains the preprocessing methods employed in the study. Section IV presents the results of the training and evaluation of the AI model. Section V offers the discussion and considers the limitations of the study. Finally, Section VI concludes the paper,

summarizes the main points, and suggests areas for future research.

II. Background

In recent years, the healthcare industry has rapidly adopted machine learning (ML) and artificial intelligence (AI) approaches to construct predictive models that better decision-making and improve patient outcomes [15]. They excel at interpreting large amounts of medical data, including electronic health records (EHRs), genomic data, and medical images, revealing patterns that humans may overlook [16], [17]. Historically, interpretable methods like linear regression and naïve Bayes have been favored in healthcare domains such as neurology and cardiology, ensuring clarity for medical practitioners. Zhang et al. [18] developed an in silico prediction model for chemical-induced urinary tract toxicity using a naïve Bayes classifier, showcasing its effectiveness in toxicology assessments. Salman [19] explored heart attack mortality prediction by applying various machine learning methods, highlighting the potential of data-driven approaches to enhance clinical decision-making. Obeid et al. [20] introduced a deep learning model for the automated detection of altered mental status in emergency department clinical notes, offering insight into the capability of natural language processing in identifying critical health conditions.

However, while linear regression and naïve Bayes are interpretable by nature, these methods often struggle with complex or non-linear datasets. More recent approaches such as decision trees, K-nearest neighbors (K-NN), and support vector machines (SVMs) allow for more nuanced analyses of the data, which is particularly useful for output for conditions such as diabetes and heart diseases [21]. Abdalrada et al. [22] investigated machine learning models for predicting the co-occurrence of diabetes and cardiovascular disease, highlighting the potential for AI-driven techniques to discover overlapping risk factors and enhance early diagnosis. Karun [23] performed a comparative analysis of heart disease prediction algorithms, assessing the usefulness of several machine learning models, such as decision trees and SVMs, in increasing diagnostic accuracy. Rajkomar et al. [24] investigated the scalability and accuracy of deep learning using electronic health records, revealing how neural networks can analyze large volumes of patient data to improve prediction capacities while remaining adaptable across various medical situations.

While the field has advanced with deep learning algorithms that can deal with complex patterns in large datasets, their black-box nature presents a hurdle to clinical trust and uptake [12], [15], [25]. Interpretability is especially critical in supervised learning, commonly used in healthcare AI to predict specific health outcomes [26]. To counter the above problems, research has focused on developing XAI systems, which combine the complexities of the model and the ease of using it in a clinical setting [27], [28]. XAI controls the complexity of user explanations to those that a medical practitioner can understand, thus helping in forming opinions that can be acted on and trusted [29], [30], [31]. Furthermore, several evaluation methods for interpretability have been proposed, including application-grounded, human-grounded, and functionally grounded approaches. Application-based evaluations, which are performed while undertaking real-life activities along with the experts in the field, are particularly important for validating XAI models in the area of medicine [27]. Human-based evaluations involving laypeople are less expensive and easier to scale, but they may not capture the specific needs of medical practitioners [32]. Functionally grounded evaluations, which do not involve human users, assess interpretability based on proxy metrics like the complexity of decision trees, but they may lack practical relevance [33]. Even with these advances, there are still challenges like obtaining acceptable regulations, integrating them into existing frameworks, and teaching professionals in the field [25].

Despite these challenges, the demand for interpretable AI systems continues to grow as they offer the potential to revolutionize medical diagnostics and treatment planning by providing transparent, trustworthy, and actionable insights [32]. In this regard, this research aims to produce an explainable predictive model for predicting mortality risks in ICUs that use XAI and conventional AI techniques while overcoming inadequate past research in this field.

III. Method

A. Dataset Description

The dataset utilized in this study is sourced from the WiDS Datathon 2020 and comprises de-identified clinical data of 91,713 ICU patients collected over a year. This extensive dataset captures 186 physiological and clinical variables documented within the first 24 hours of ICU admission. The variables include a broad spectrum of demographic information, such as age, gender, and BMI, along with various health metrics like blood pressure, heart rate, and laboratory test results. These variables are essential for predicting patient outcomes, particularly mortality risk, in an ICU setting.

B. Data Preprocessing

1) Data cleaning: The first step in the data preprocessing involved loading the dataset and reducing the number of features from 186 to 28. This reduction was achieved by eliminating irrelevant or redundant features that did not contribute significantly to the prediction task.

2) Handling missing data: Columns containing significant missing data were eliminated to ensure the dataset's integrity and reliability. This step was crucial in ensuring the analysis was based on complete and accurate information. After removing these columns, the dataset was further refined, resulting in a final dataset consisting of 940 rows and 28 columns (Fig. 1, Fig. 2).

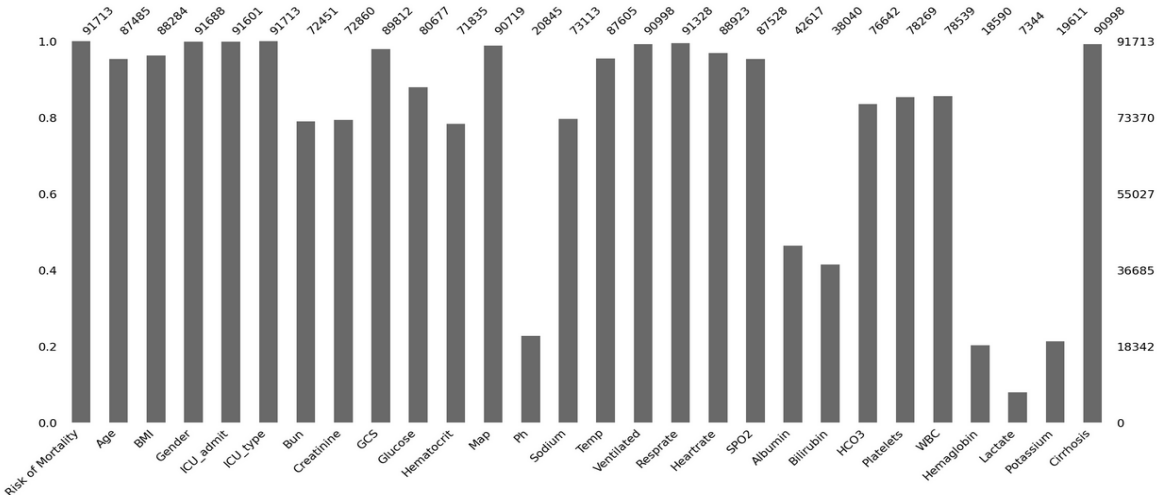


Fig. 1. Distribution of missing data across features.

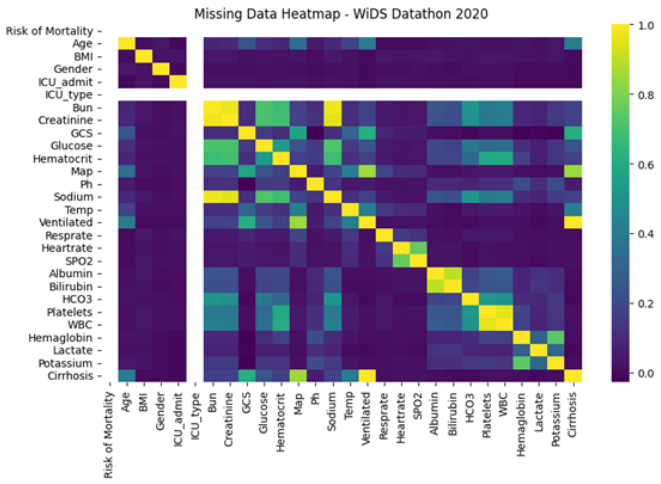


Fig. 2. Heatmap visualizing the correlations of missing values.

First, a correlation matrix of the missing values was generated, and a heatmap was used to identify patterns in the missing data. To address this, we applied the `dropna` function, which successfully removed the missing values and mitigated the impact of outliers. Rather than using imputation, we chose to eliminate entire columns with any missing data, prioritizing precision in healthcare predictions. While removing features with less than 5% missingness might seem counterintuitive, the critical nature of medical decisions made data completeness more important.

3) Data encoding: Data encoding techniques were applied to prepare the categorical variables for machine learning models. Specifically, Label Encoding was used to convert categorical variables such as Gender, ICU_admit, and ICU_type into numerical values, as illustrated in Fig. 3. This transformation allowed the models to process these variables effectively, as most machine learning algorithms require numerical input [34].

- Gender:
  - Original values: 'F', 'M'
  - Encoded values: 0, 1
- ICU_admit:
  - Original values: 'Floor', 'Operating Room / Recovery', 'Accident & Emergency', 'Other Hospital', 'Other ICU'
  - Encoded values: 0, 1, 2, 3, 4
- ICU_type:
  - Original values: 'Med-Surg ICU', 'MICU', 'CTICU', 'Neuro ICU', 'SICU', 'CCU-CTICU', 'Cardiac ICU', 'CSICU'
  - Encoded values: 0, 1, 2, 3, 4, 5, 6, 7

Fig. 3. Label encoding was applied to convert categorical variables.

4) Exploratory Data Analysis (EDA): EDA was conducted to uncover patterns and insights within the dataset. This analysis included visualizing demographic trends and identifying risk patterns among ICU patients. For instance, analysis of patients' age makeup clarified the age distribution. With concentrations in the 60–70 and 70–80 cohorts, age-related dynamics may impact mortality risk. Interestingly, the 50–60 cohort was equally large, demonstrating the sample's age diversity. Fig. 4 shows how the age distribution affects risk assessment. These insights guided the feature selection process and informed the subsequent steps in the analysis.

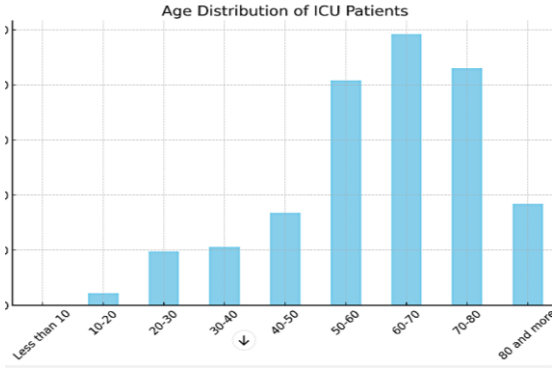


Fig. 4. Distribution of patient age groups.

5) Handling outliers: Outliers can distort the results of data analysis and model predictions. To address this, the Interquartile Range (IQR) technique was employed. Data points below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR were deemed outliers and deleted (Fig. 5, Fig. 6). This step ensured that the data used for model training was free from extreme values that could potentially bias the analysis [35].

6) Feature engineering: Feature engineering was performed to enhance the predictive power of the models. The SelectKBest method was used to identify the most significant features, and these selected features were then standardized using StandardScaler.

7) Balancing dataset: Given the class imbalance in the dataset, where high-risk cases were underrepresented, oversampling techniques were applied to balance the dataset. This step involved increasing the number of high-risk instances to ensure the model could accurately predict high-risk and low-risk cases. Balancing the dataset was vital for improving the model's sensitivity and reducing bias toward the majority class [36].

8) Data splitting: Partitioning the dataset into training and testing sets was the final step in data preprocessing. Eighty percent of the data was used to train the models, with the remaining 20% put aside for testing. This split meant that the models were evaluated using previously unseen data, resulting in a reliable assessment of their performance [37].

C. Model Training and Evaluation

The following six machine learning algorithms were chosen for evaluation: K-Nearest Neighbors, Random Forest, Decision Tree, Gradient Boosting, Logistic Regression, and Support Vector Machine. To optimize performance, GridSearchCV was implemented to tune hyperparameters for each model. The models were evaluated using F1-Score, Accuracy, Precision, and Recall classification measures. The best performing model was chosen for its proficiency in correctly forecasting ICU patient death.

D. Model Explainability

This study employed Explainable Artificial Intelligence (XAI) methods, particularly SHAP and LIME, to enhance model transparency. SHAP was used to interpret the importance of features in the Random Forest model, providing a detailed understanding of how each feature influenced risk predictions. LIME offered localized explanations for individual predictions, aiding in interpreting specific instances. After comparing both methods, SHAP was chosen as the primary tool for generating user-centric explanations, owing to its effectiveness in healthcare contexts [12], [38].

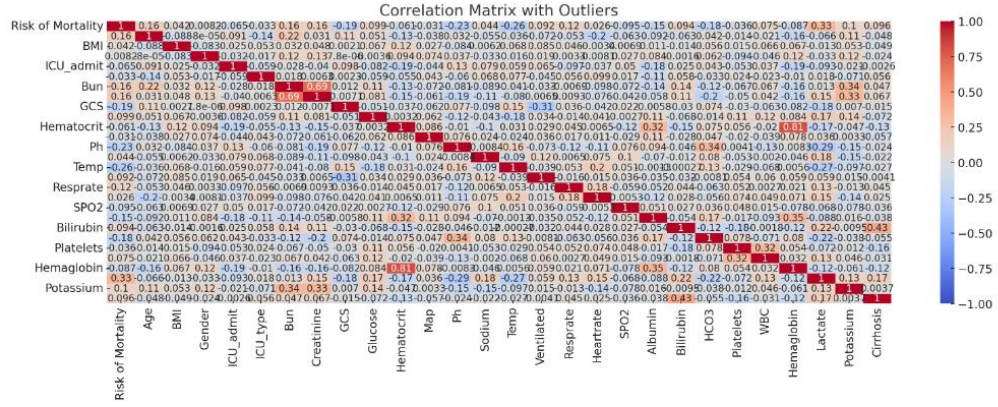


Fig. 5. Correlation matrix with outliers.

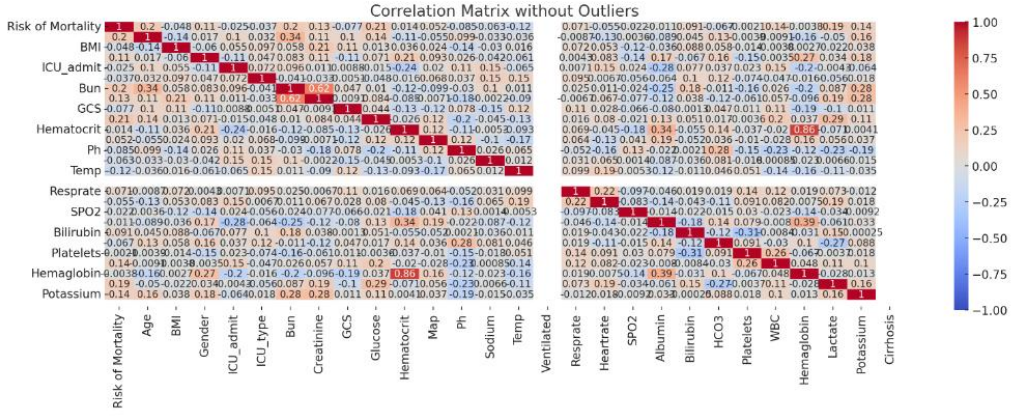


Fig. 6. Correlation matrix without outliers.

IV. Results

A. Model Performance

The performance of the six machine learning algorithms was compared based on their ability to predict mortality risk. As shown in Table I, the Random Forest algorithm outperformed the other models across all evaluation metrics, with an accuracy of 0.8511, precision of 0.8485, recall of 0.855, and an F1-Score of 0.8517.

TABLE I. Comparison of Machine Learning Techniques

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.7252	0.7185	0.7405	0.7293
Random Forest	**0.8511**	**0.8485**	**0.855**	**0.8517**
Support Vector Machine	0.8321	0.8372	0.8244	0.8308
Gradient Boosting	0.8206	0.8088	0.8397	0.824
K-Nearest Neighbors	0.8244	0.7931	0.8779	0.8333
Decision Tree	0.7443	0.7133	0.8168	0.7616

B. Accuracy of Models

Fig. 7 illustrates the accuracy of the six machine learning algorithms. The Random Forest model proved the best accuracy, while the Support Vector Machine, K-Nearest Neighbors, and Gradient Boosting models performed comparably. In comparison, the Decision Tree and Logistic Regression models had slightly lower accuracy values.

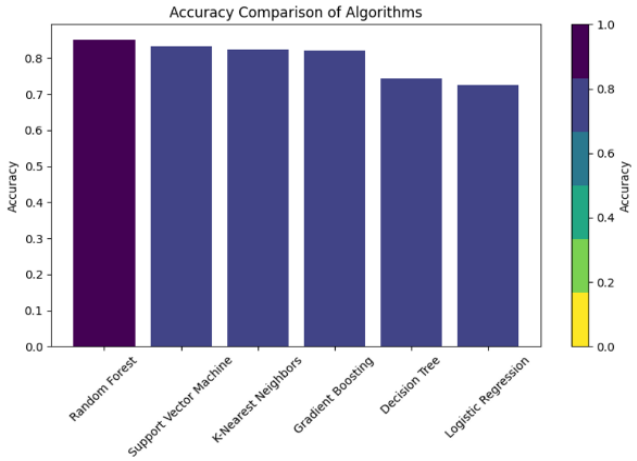


Fig. 7. Comparison of algorithms showing random forest model has the best accuracy.

1) ROC curve analysis: The ROC curve analysis emphasized the Random Forest model's superior performance. As shown in Fig. 8, the Random Forest model had the most significant AUC of 0.94, showing an excellent capacity to distinguish between high-risk and low-risk mortality cases. The Support Vector Machine model was next with an AUC of 0.92, while the Decision Tree model had the lowest AUC of 0.82.

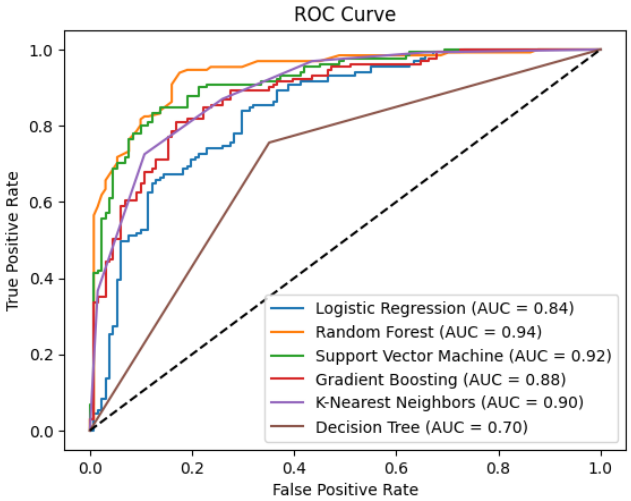


Fig. 8. ROC curve comparison for six machine learning algorithms.

2) Confusion matrix: The confusion matrix for each model provided insights into their classification performance. Fig. 9 illustrates the number of correct and incorrect predictions made by each model. The Random Forest model had the highest counts of true positives and true negatives, indicating its superior ability to classify both high-risk and low-risk mortality cases correctly.

C. Feature Importance Analysis

To enhance the interpretability of the model, Explainable Artificial Intelligence (XAI) methodologies, specifically SHAP and LIME, were employed. These techniques were valuable in making the model's decision-making process more transparent and understandable for healthcare practitioners.

1) SHapley Additive exPlanations (SHAP): SHAP was used to interpret the feature importance of the trained Random Forest model. A SHAP summary plot (Fig. 10) was generated using the TreeExplainer, which showed that features such as age, SOFA score, and lactate levels were significant contributors to the model's predictions. The SHAP values offered a clear grasp of how each characteristic affects mortality risk predictions, with greater SOFA scores and elevated lactate levels related to increased mortality risk, as acknowledged by studies such as [12], [38], [39].

2) Local Interpretable Model-agnostic Explanations (LIME): LIME was employed to explain individual predictions by generating local explanations for specific instances. Fig. 11 illustrates a detailed LIME explanation, showing the contribution of each feature to the prediction. For instance, one risk-predicting factor was high lactate level with significantly low temperature, with other factors like GCS, SPO2, ICU type, creatinine levels, age, and gender also playing roles. By breaking down predictions into these specific details, LIME enables healthcare practitioners to understand the rationale behind individual predictions, fostering trust in the model's outputs [13], [40], [41].

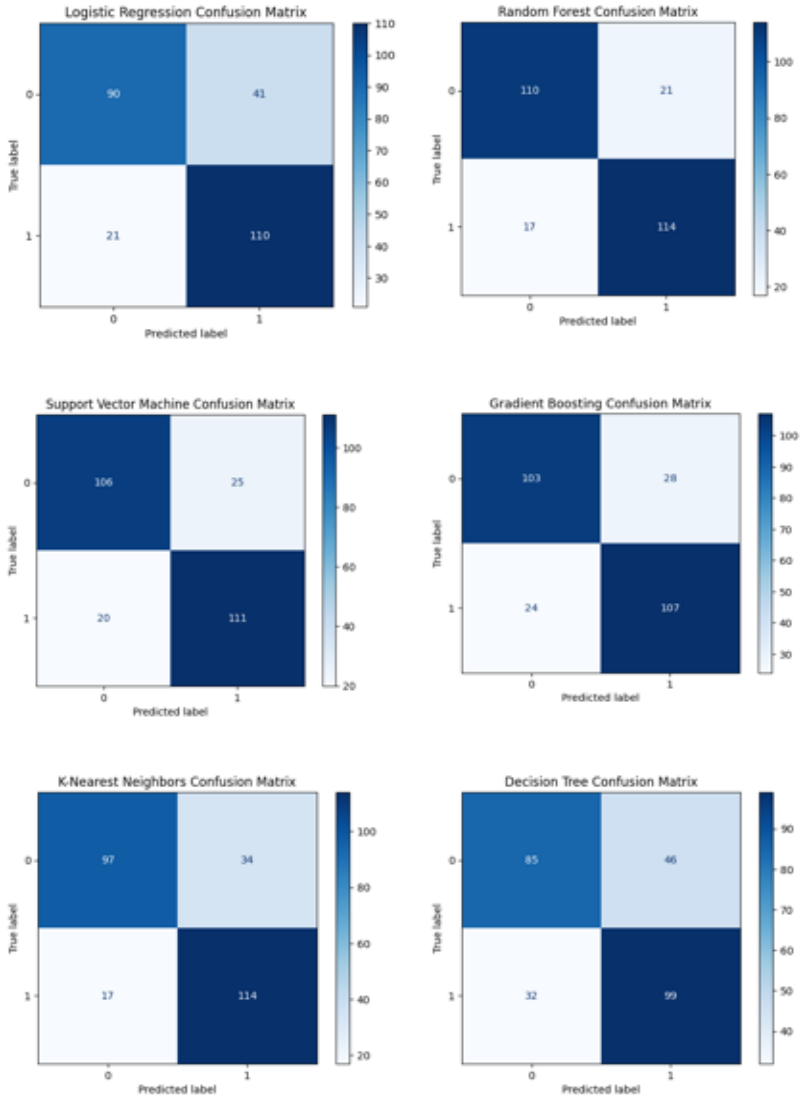


Fig. 9. Confusion matrix showing the number of correct and incorrect predictions made by each model.

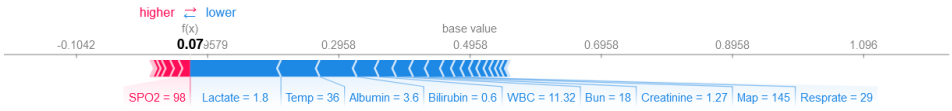


Fig. 10. SHAP summary plot.

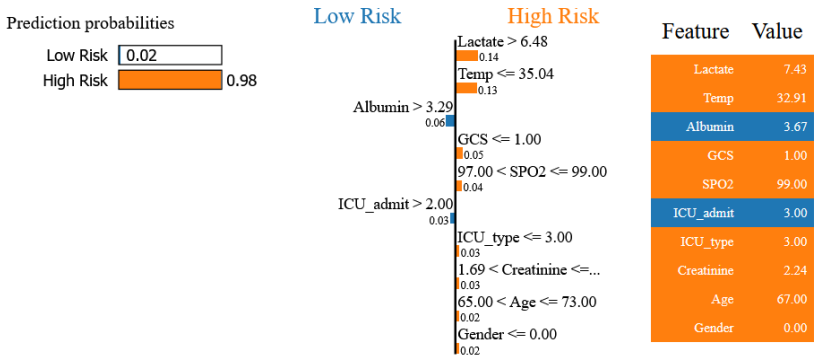


Fig. 11. LIME explanation.

V. Discussion

The results of this study are consistent with previous research that has demonstrated the effectiveness of Random Forest in predicting mortality risk in ICU patients. The Random Forest model's outstanding performance is primarily due to its capacity to handle large datasets and capture complicated correlations between variables. This study's findings align with previous research showing the usefulness of Random Forest in predicting mortality risk in ICU patients. The Random Forest model's exceptional success is primarily due to its ability to handle large datasets and identify the intricate connections between variables. The decision to balance the dataset by oversampling was crucial for improving the model's accuracy and reducing bias, a method supported by other studies, including that of [36]. Although [4] showed how effectively deep learning models capture non-linear correlations in complicated datasets, Random Forest was selected for this study because of its exceptional interpretability and widespread use in clinical settings. The use of SHAP and LIME provided valuable insights into the model's decision-making process. SHAP allowed for a global interpretation of feature importance, highlighting the significant role of clinical markers such as age, SOFA score, and lactate levels. These results are consistent with clinical expectations and established literature, thereby validating the model's predictions [12], [38], [39], [42]. The implications of this study for clinical practice are substantial. Implementing a Random Forest-based mortality risk prediction model enables healthcare professionals to more accurately identify patients at elevated risk of mortality and allocate resources more efficiently. This methodology may result in enhanced individualized care plans and superior patient outcomes. Furthermore, the model's ability to incorporate a wide range of clinical variables makes it adaptable to different ICU settings and patient populations.

While the results are promising, the study is not without limitations. The dataset is comprehensive; hence, the generalizability of the model may be restricted to different ICU populations. Despite the dataset is large, it only comes from one source. Hence, the generalizability of the model may be restricted to different ICU populations. This emphasizes the need for external validation over several datasets representing various geographical and clinical settings. Also, relying on these individual attributes raises concerns about potential biases. For example, demographic factors such as age or gender may potentially introduce systemic biases, limiting the model's generalizability across different patient populations [43]. Additionally, even though Random Forest performed more effectively than the other models in this study, investigating advanced methods—like ensemble approaches that combine deep learning with XAI tools—could improve the models' predictive abilities and interpretability.

VI. Conclusion

This study developed and evaluated a machine learning model to predict ICU patient mortality risk using the WiDS Datathon 2020 dataset. The Random Forest algorithm emerged as the most effective model, outperforming other algorithms in accuracy, precision, recall, and F1-Score. The AI mortality risk prediction model showed strong performance, with SHAP and LIME providing essential tools for enhancing model explainability. SHAP was particularly effective in offering clear and actionable explanations, making it the preferred method for developing user-centric explanations in this study. The study's findings highlight the potential of machine learning to improve clinical decision-making in ICU settings. Future research should focus on validating the model across multiple datasets from diverse geographic and clinical settings. Additionally, exploring integrating other machine learning techniques, such as deep learning, could further enhance the model's predictive accuracy.

References

[1] S. Barbieri, J. Kemp, O. Perez-concha, and S. Kotwal, “Benchmarking Deep Learning Architectures for Predicting Readmission to the ICU and Describing Patients-at-Risk,” Sci Rep, pp. 1–10, 2020, doi: 10.1038/s41598-020-58053-z.

[2] C. V Cosgriff, L. A. Celi, and D. J. Stone, “Critical Care, Critical Data,” Biomed Eng Comput Biol, vol. 10, p. 117959721985656, Jan. 2019, doi: 10.1177/1179597219856564.

[3] A. Hassan, R. bin Sulaiman, M. A. Abdulgabber, and H. Kahtan, “Balancing Technological Advances with User Needs: User-centered Principles for AI-Driven Smart City Healthcare Monitoring,” International Journal of Advanced Computer Science and Applications, vol. 14, no. 3, p. 2023, 2023, doi: 10.14569/IJACSA.2023.0140341.

[4] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, “Deep learning for healthcare: Review, opportunities and challenges,” Brief Bioinform, vol. 19, no. 6, pp. 1236–1246, 2017, doi: 10.1093/bib/bbx044.

[5] S. M. Lauritsen et al., “Explainable artificial intelligence model to predict acute critical illness from electronic health records,” Nat Commun, vol. 11, no. 1, pp. 1–11, 2020, doi: 10.1038/s41467-020-17431-x.

[6] S. Boughorbel, F. Jarray, N. Venugopal, S. Moosa, H. Elhadi, and M. Makhlouf, “Federated Uncertainty-Aware Learning for Distributed Hospital EHR Data,” pp. 1–6, 2019, [Online]. Available: http://arxiv.org/abs/1910.12191

[7] O. Efthimiou, M. Seo, K. Chalkou, T. Debray, M. Egger, and G. Salanti, “Developing clinical prediction models: a step-by-step guide,” BMJ, p. e078276, Sep. 2024, doi: 10.1136/bmj-2023-078276.

[8] M. A. Rahman Bhuiyan, M. R. Ullah, and A. K. Das, “iHealthcare: Predictive model analysis concerning big data applications for interactive healthcare systems,” Applied Sciences (Switzerland), vol. 9, no. 16, 2019, doi: 10.3390/app9163365.

[9] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, “Intelligible Models for HealthCare,” in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, Aug. 2015, pp. 1721–1730. doi: 10.1145/2783258.2788613.

[10] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” Nat Mach Intell, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[11] M. Cheng, X. Li, and J. Xu, “Promoting Healthcare Workers’ Adoption Intention of Artificial-Intelligence-Assisted Diagnosis and Treatment: The Chain Mediation of Social Influence and Human–Computer Trust,” Int J Environ Res Public Health, vol. 19, no. 20, 2022, doi: 10.3390/ijerph192013311.

[12] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” Adv Neural Inf Process Syst, vol. 2017-Decem, no. Section 2, pp. 4766–4775, May 2017, [Online]. Available: http://arxiv.org/abs/1705.07874

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, “Model-Agnostic Interpretability of Machine Learning,” no. Whi, Jun. 2016, [Online]. Available: http://arxiv.org/abs/1606.05386

[14] “WiDS Datathon 2020.” Accessed: Aug. 17, 2023. [Online]. Available: https://www.kaggle.com/c/widsdatathon2020/data

[15] M. A. Ahmad, A. Teredesai, and C. Eckert, “Interpretable machine learning in healthcare,” in Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018, IEEE, Jun. 2018, p. 447. doi: 10.1109/ICHI.2018.00095.

[16] J. H. Chen and S. M. Asch, "Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations," New England Journal of Medicine, vol. 376, no. 26, pp. 2507–2509, Jun. 2017, doi: 10.1056/NEJMp1702071.

[17] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," 2019. doi: 10.1038/s41591-018-0300-7.

[18] H. Zhang, J. X. Ren, J. X. Ma, and L. Ding, "Development of an in silico prediction model for chemical-induced urinary tract toxicity by using naïve Bayes classifier," Mol Divers, vol. 23, no. 2, pp. 381–392, 2019, doi: 10.1007/s11030-018-9882-8.

[19] I. Salman, "Heart attack mortality prediction: An application of machine learning methods," Turkish Journal of Electrical Engineering and Computer Sciences, vol. 27, no. 6, pp. 4378–4389, 2019, doi: 10.3906/ELK-1811-4.

[20] J. S. Obeid et al., "Automated detection of altered mental status in emergency department clinical notes: A deep learning approach," BMC Med Inform Decis Mak, vol. 19, no. 1, pp. 1–9, 2019, doi: 10.1186/s12911-019-0894-9.

[21] D. D. Miller, "The Big Health Data–Intelligent Machine Paradox," Am J Med, vol. 131, no. 11, pp. 1272–1275, Nov. 2018, doi: 10.1016/j.amjmed.2018.05.038.

[22] A. S. Abdalrada, J. Abawajy, T. Al-Quraishi, and S. M. S. Islam, "Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: a retrospective cohort study," J Diabetes Metab Disord, vol. 21, no. 1, 2022, doi: 10.1007/s40200-021-00968-z.

[23] I. Karun, Comparative Analysis of Prediction Algorithms for Heart Diseases, vol. 1158. Springer Singapore, 2021. doi: 10.1007/978-981-15-4409-5_53.

[24] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," NPJ Digit Med, vol. 1, no. 1, 2018, doi: 10.1038/s41746-018-0029-1.

[25] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," Future Healthc J, vol. 6, no. 2, pp. 94–98, Jun. 2019, doi: 10.7861/futurehosp.6-2-94.

[26] H. Lu, Y. Li, M. Chen, H. Kim, and S. Serikawa, "Brain Intelligence: Go beyond Artificial Intelligence," Mobile Networks and Applications, vol. 23, no. 2, pp. 368–375, 2018, doi: 10.1007/s11036-017-0932-8.

[27] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," Feb. 2017, [Online]. Available: http://arxiv.org/abs/1702.08608

[28] R. Dwivedi et al., "Explainable AI (XAI): Core Ideas, Techniques, and Solutions," ACM Comput Surv, vol. 55, no. 9, 2023, doi: 10.1145/3561048.

[29] A. Hassan, R. Sulaiman, M. A. Abdulgabber, and H. Kahtan, "Towards User-Centric Explanations For Explainable Models: A Review," Journal of Information System and Technology Management, vol. 6, no. 22, pp. 36–50, Sep. 2021, doi: 10.35631/JISTM.622004.

[30] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. van Moorsel, "The relationship between trust in AI and trustworthy machine learning technologies," FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 272–283, 2020, doi: 10.1145/3351095.3372834.

[31] L. Yu and Y. Li, "Artificial Intelligence Decision-Making Transparency and Employees' Trust: The Parallel Multiple Mediating Effect of Effectiveness and Discomfort," Behavioral Sciences, vol. 12, no. 5, 2022, doi: 10.3390/bs12050127.

[32] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine Learning Interpretability: A Survey on Methods and Metrics," Electronics (Basel), vol. 8, no. 8, p. 832, Jul. 2019, doi: 10.3390/electronics8080832.

[33] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," in 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, Oct. 2018, pp. 80–89. doi: 10.1109/DSAA.2018.00018.

[34] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data for neural networks," J Big Data, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00305-w.

[35] H. P. Vinutha, B. Poornima, and B. M. Sagar, "Detection of outliers using interquartile range technique from intrusion dataset," in Advances in Intelligent Systems and Computing, 2018. doi: 10.1007/978-981-10-7563-6_53.

[36] K. Rasheed, A. Qayyum, M. Ghaly, A. Al-Fuqaha, A. Razi, and J. Qadir, "Explainable, trustworthy, and ethical machine learning for healthcare: A survey," 2022. doi: 10.1016/j.compbiomed.2022.106043.

[37] Z. Ashfaq et al., "Embedded AI-Based Digi-Healthcare," Applied Sciences (Switzerland), vol. 12, no. 1, 2022, doi: 10.3390/app12010519.

[38] S. Das, M. Sultana, S. Bhattacharya, D. Sengupta, and D. De, "XAI–reduct: accuracy preservation despite dimensionality reduction for heart disease classification using explainable AI," Journal of Supercomputing, 2023, doi: 10.1007/s11227-023-05356-3.

[39] N. Barr Kumarakulasinghe, T. Blomberg, J. Liu, A. Saraiva Leao, and P. Papapetrou, "Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models," in Proceedings - IEEE Symposium on Computer-Based Medical Systems, 2020. doi: 10.1109/CBMS49503.2020.00009.

[40] R. K. Sheu and M. S. Pardeshi, "A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System," 2022. doi: 10.3390/s22208068.

[41] R. Rodríguez-Pérez and J. Bajorath, "Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions," J Comput Aided Mol Des, vol. 34, no. 10, 2020, doi: 10.1007/s10822-020-00314-0.

[42] S. D. Mohanty, D. Lekan, T. P. McCoy, M. Jenkins, and P. Manda, "Machine learning for predicting readmission risk among the frail: Explainable AI for healthcare," Patterns, vol. 3, no. 1, 2022, doi: 10.1016/j.patter.2021.100395.

[43] R. R. Fletcher, A. Nakeshimana, and O. Olubeko, "Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health," Front Artif Intell, vol. 3, 2021, doi: 10.3389/frai.2020.561802.

Comparative Analysis of Undersampling, Oversampling, and SMOTE Techniques for Addressing Class Imbalance in Phishing Website Detection

Kamal Omari[1], Chaimae Taoussi[2], Ayoub Oukhatar[3]
Computer Science Department-Polydisciplinary Faculty of Ouarzazate, University Ibn Zohr, Ouarzazate, Morocco[1]
Laboratory of Process Engineering Computer Science and Mathematics,
University Sultan Moulay Slimane, Beni Mellal, Morocco[2]
Computer Science Department-Higher School of Technology Ouarzazate, University Ibn Zohr, Ouarzazate, Morocco[3]

***Abstract*—Since this is one of the most challenging tasks in cyber security, many of them are affected by class imbalance when it comes to the performance of machine learning. This paper evaluates various strategies using a number of resampling-based approaches: ROS, RUS, and SMOTE-based methods in conjunction with XGBoost classifier techniques to solve such an imbalanced dataset. Key performance measures include precision, F1 score, recall, precision, ROC-AUC, and geometric mean score. Among the methods, the highest was found with regard to the SMOTE-NC-XGB with precision equal to 98.0% and a recall of 98.5%, thus ensuring an effective trade-off between sensitivity and specificity. Although the stand-alone XGB model performs really well, adding resampling techniques makes its efficiency much higher, especially in cases of evident imbalance between classes. These results also revealed that resampling techniques are really helpful to enhance detection performance; hence, the SMOTE-NC-XGB is found out as the best among all of these. It will be of great contribution for future works in order to enhance the development of phishing detection systems and investigate other new hybrid resampling methods.**

Keywords*—Phishing website detection; class imbalance; XGBoost; SMOTE-NC

I. Introduction

While the cyberworld has expanded infinitely, phishing assaults have evolved into increasingly sophisticated attacks that target individuals and entities by assuming the identity of legitimate websites for the purpose of retrieving sensitive information [1]. Detection of these malicious sites is an important challenge in cybersecurity which is further complicated by the gross class imbalance contained in phishing collections. In these datasets, legitimate websites far exceed phishing websites, and machine learning models cannot detect phishing attacks effectively—a key process of mitigating cyber attacks [2].

Accuracy and reliability of machine learning models heavily depend on the quality and consistency of training data. Proper data preparation—cleaning, processing, validation, and transformation of raw data—is essential to building good models [3]. Class imbalance handling is one of the most important tasks in this work. Class imbalance, as a condition of underrepresented minority class, is common in applications such as fraud detection, health, and phishing website detection [4, 5]. In phishing detection, underrepresentation of phishing sites in datasets generally results in model bias towards the majority class, increasing the risk of false negatives, where phishing sites are classified as normal [6].

Machine learning models are negatively affected by imbalanced datasets with skewed class distributions, as this causes the models to be biased towards the majority class. This diminishes their generalization capability and hinders their applicability in real-world situations [7]. Traditional classification techniques are biased towards the majority class, which worsens these problems and highlights the necessity of specialized methods to counteract class imbalance.

In order to address this problem, researchers have put forward various resampling methods, i.e., undersampling, oversampling, and the Synthetic Minority Over-sampling Technique (SMOTE) [8]. All these methods assist in rebalancing data by reducing the majority class size or enhancing the minority class size and, therefore, the model becomes more capable of distinguishing between authentic sites and phishing websites. Here, we will attempt to combine these resampling techniques with the XGBoost classifier, a widely recognized algorithm for its high performance and scalability [9]. Despite its demonstrated effectiveness, inadequate detailed comparative studies in the literature have investigated the impacts of using different resampling techniques when coupled with XGBoost for detecting phishing websites.

This research aims to bridge this gap by providing a comparative analysis of undersampling, oversampling, and SMOTE methods under the XGBoost algorithm for phishing website detection optimization. From real datasets, we compare how these resampling methods affect the performance of the XGBoost classifier in handling class imbalance. The results not only show the relative merits and drawbacks of every resampling method but also give specific guidelines for researchers and practitioners interested in constructing more robust phishing detection systems.

The rest of this paper is organized as follows: Section II describes the importance of class imbalance in phishing website detection and how resampling methods can be utilized to alleviate it. Section III compares the performance of the XGBoost classifier when combined with these resampling methods, comparing their impacts on key performance metrics. Section IV outlines the experimental design, data set, preprocessing, and application of the XGBoost algorithm. Section V presents the results and discusses the impact of resampling methods on model performance. Finally, Section VI concludes with a summary of the major findings and suggesting potential areas for further study.

II. Literature Review

Preparation of data is an important phase of the data mining process, typically consuming 70–80% of the total time invested in data science activities [10]. The "garbage in, garbage out" (GIGO) principle emphasizes the extreme importance of high-quality input data in obtaining accurate and useful output. Good data preparation is more than data cleansing to include missing value handling, duplicate elimination, and outlier handling, all of which are required to derive actionable insights from raw and usually unstructured data [11].

In detecting phishing sites, datasets show complex patterns when it comes to characteristics such as URL patterns, hosting characteristics, and user activities. These patterns can be effective inputs for predictive models to identify phishing attacks [12]. However, these patterns have to be unearthed via repeated testing and knowledge of domains [13]. Therefore, data preparation is an important task in cybersecurity that transforms raw data into formats where it can be acted upon, being sensitive to machine learning algorithms.

Among the biggest problems for detecting phishing is the extreme class imbalance, where genuine websites heavily outnumber phishing websites. The issue tends to bias model performance and has a propensity to result in bias towards the majority class and higher false negative likelihoods. This research seeks to mitigate this through experimentally testing resampling methods including undersampling, oversampling, and an in-house approach called SMOTE-NC (Synthetic Minority Over-sampling Technique for Nominal and Continuous variables). These methods are intended to introduce balance to the dataset so that the model can better identify phishing sites accurately and eliminate the adverse effect of class imbalance.

A. Undersampling Techniques to Handle Class Imbalance Undersampling reduces the majority class in such a way that the dataset is balanced and machine learning algorithms can learn from minority class samples without bias. For instance, downsampling 12,000 majority class samples to balance 2,000 minority class samples creates a balanced 1:1 ratio. Although this technique makes the dataset easy to train, it must selectively choose majority class samples to maintain the integrity and predictive power of the dataset (Fig. 1).

A. Undersampling Techniques for Addressing Class Imbalance

Undersampling decreases the majority class to a level where the dataset is balanced, allowing the machine learning models to learn from the instances in the minority class without bias. For instance, decreasing 12,000 majority class samples to equal 2,000 minority class samples creates a 1:1 balanced dataset. Although this technique minimizes the complexity of the dataset during training, it needs to sample the majority class samples aggressively so that it does not compromise the integrity and forecastability of the dataset (Fig. 1).

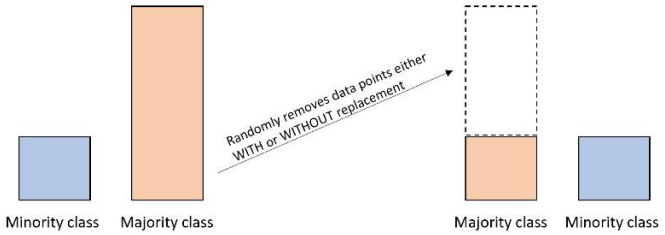


Fig. 1. Random undersampling process [14].

1) Random Undersampling (RUS): Random Undersampling (RUS) is the most basic and widely used undersampling technique. It does this by randomly selecting a subset of samples from the majority class to build a balanced dataset. This reduces the majority class bias, thereby improving the sensitivity of the model to phishing websites [15].

RUS has been used in a wide range of applications that encompass anomaly detection, fraud detection, and cyber security. When used in the detection of phishing websites, it reduces false negatives and enhances model performance by stabilizing class representation. Its simplicity comes at a price, though: the potential loss of informative information from the discarded majority class samples, which can limit the model's generalization capability in real-world scenarios.

B. Oversampling Techniques for Addressing Class Imbalance

Oversampling compensates for class imbalance by increasing the size of the minority class in the data set. The current minority class samples are replicated or new synthetic samples are created. Since the data set is balanced, oversampling eliminates the majority class bias and enhances the performance of the model in identifying phishing websites (Fig. 2).

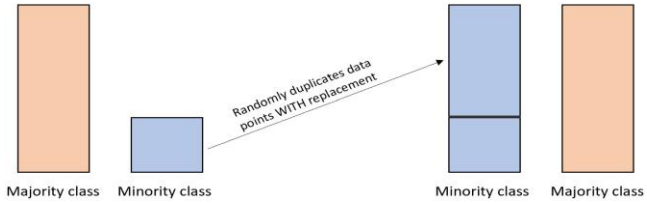


Fig. 2. Random oversampling process [14].

1) Random Oversampling (ROS): Random Oversampling (ROS) evens out the dataset by duplicating instances from the minority class to be identical to the majority class instances. As effective as this process treats class imbalance, it would also lead to overfitting based on the duplication of the same instances, hence reduced robustness and ability to generalize to new data [16].

In order to counteract these issues, variants like distributional random oversampling, where the synthetic samples are created from the statistical distribution of features, and stratified random oversampling, where generation of varied samples is guaranteed, have been suggested. Though having its own drawbacks, ROS has been found to be highly beneficial for phishing website detection by enhancing the representation of the minority class and model performance during training.

C. SMOTE-NC for Handling Class Imbalance

SMOTE-NC is an extension of the original SMOTE algorithm for the treatment of datasets with mixed data types, i.e., categorical and numerical variables. SMOTE-NC creates the synthetic samples in such a manner that it preserves the inherent features and interrelation among the data and effectively addresses class imbalance without compromising data integrity (Fig. 3).

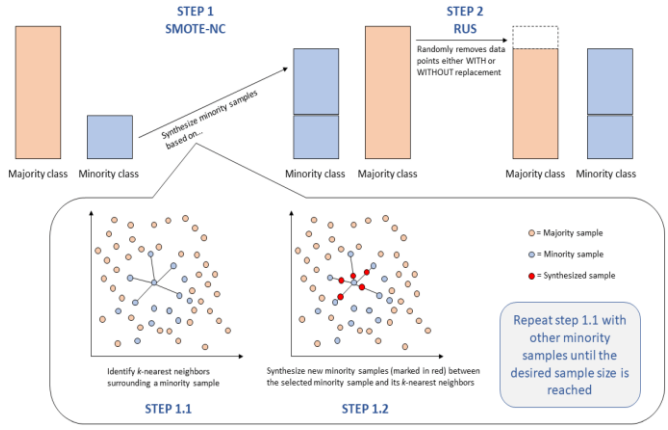


Fig. 3. SMOTE-NC process [14].

SMOTE-NC algorithm generates samples by interpolating between the minority class instances and their k-nearest neighbors in the numerical features. In categorical features, it finds the most frequent category among the neighbors so that the synthetic samples preserve the inherent categorical data distributions. This approach preserves the integrity of the mixed data types while improving the model performance, particularly in the detection of phishing websites [17].

However, SMOTE-NC is not limited [18]:

- Problems with high-dimensional data: Distance measures are meaningless in high-dimensional space.
- Introduction of noise: Synthetic samples may deviate from the true minority class distribution.
- Fixed k-neighbors: The k-neighbors parameter is applied uniformly, without considering the local variation of the data.

Despite the limitations outlined, SMOTE-NC has proven to be a valuable tool when used in phishing website detection, particularly when operating in datasets that involve categorical features such as domain-pattern-based or URL feature-based classifications. Through the preservation of integrity of the categorical variables and handling the class imbalance problem, SMOTE-NC operates to enhance the model towards effective phishing website detection.

Overall, undersampling, oversampling, and hybrid approaches each possess their own strength in class imbalance handling for phishing website detection. Selection of an effective resampling approach must be guided by the characteristics of the dataset and the requirements of the task at hand. Effective data preprocessing, particularly when working with imbalanced datasets, is critical to developing stable and actionable machine learning models with the ability to counteract the dynamic nature of phishing attacks.

III. Methods

This chapter introduces the research framework that seeks to enhance phishing website detection using a combination of state-of-the-art resampling methods and the XGBoost classifier. This chapter begins with introducing a step-by-step description of the dataset, as well as a heatmap visualization of feature correlation, which gives valuable information regarding variable correlation. The article then goes on to outline various class imbalance learning techniques, i.e., Random Oversampling (ROS), Random Undersampling (RUS), and SMOTE-NC, and outlines their settings and respective contributions to the resampling process. The XGBoost classifier is then outlined in detail, including its setup, tuning, and why it is suitable for the phishing detection task. This integrated paradigm facilitates systematic study of resampling's methods and their combined impact when used along with XGBoost, thereby preventing the consequences of class imbalance in phishing site detection.

A. Dataset Description

The dataset used in this work is taken from the UCI Machine Learning Repository [19]. It contains 11,055 records with 30 website attributes along with a class label indicating websites as phishing (1) and legitimate (-1). Table I summarizes the makeup of the dataset.

TABLE I. Dataset Description

Total number of instances	Total of Features	Class Variable Phishing website	Class Variable Legitimate website
11055	30	6157	4898

1) Visualizing the dataset: Heatmap of feature correlations: To understand relationships between features, a heatmap visualization was generated, identifying patterns and correlations crucial for feature engineering and selection [20] (Fig. 4).

In the heat map, each cell is colored according to a correlation value that goes from dark to light, indicating respectively a strong positive correlation and a strong negative one. The closer the correlation value is to +1, the stronger the positive relationship. The closer the value is to -1, the stronger the negative relationship. A correlation close to 0 means there is little or no linear relationship between the compared features. This visualization has given a better idea of interfeature dependencies; hence, crucial toward understanding the feature relevancy in respect of phishing website detection.

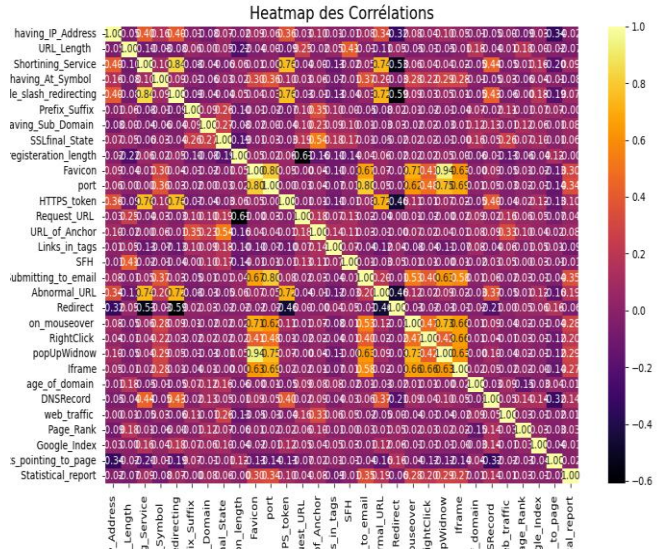


Fig. 4. The heatmap of dataset features.

B. Imbalance Learning Techniques

In this paper, ROS, RUS, and a hybrid approach with SMOTE-NC are used for performance comparisons of different resampling methods. SMOTE-NC with five nearest neighbors uses a resampling ratio of 0.8 to generate synthetic samples, while ROS and RUS apply random sampling processes to adjust the class distribution. These resampling techniques were implemented using the imbalanced-learn Python package [21], ensuring the reproducibility and consistency of the results.

1) Random Oversampling-XGBoost (ROS-XGB): ROS is a technique of oversampling the minority class samples to make the dataset balanced [22]. The ROS-XGB method, which combines the XGBoost classifier with random oversampling, enhances the model performance in phishing website detection by overcoming the problem of class imbalance. The hyperparameters of ROS-XGB are tuned for better performance. The details are shown in Table II. The parameters include the learning rate, maximum depth, number of estimators, and subsampling rate, which are tuned carefully for better accuracy and robustness of the model.

TABLE II. ROS-XGB PARAMETERS

Parameter	Value	Description
learning_rate	0.5	Resampling the minority class to balance the dataset.
max_depth	7	Controls tree depth for model complexity.
n_estimators	100	Number of boosting rounds.
subsample	1.0	Fraction of samples used for fitting each tree.

ROS-XGB improves recall and overall performance by mitigating class imbalance, making it effective at detecting phishing websites while maintaining high precision.

2) Random Undersampling-XGBoost (RUS-XGB): Random Undersampling (RUS) reduces the size of the majority class to that of the minority class, ensuring a balanced dataset [23]. When combined with the XGBoost classifier, the RUS-XGB approach enhances phishing website detection by giving equal importance to both classes. The hyperparameters for RUS-XGB are shown in Table III, which presents the specific configurations used to optimize model performance.

TABLE III. RUS-XGB PARAMETERS

Parameter	Value	Description
learning_rate	0.5	Resampling the minority class to balance the dataset.
max_depth	8	Tree depth to control model complexity.
n_estimators	300	Number of boosting rounds.
subsample	0.8	Fraction of samples used for fitting each tree.

RUS-XGB helps address class imbalance by reducing the influence of the majority class, thus improving recall and overall detection accuracy.

3) SMOTE-NC-XGBoost (SMOTE-NC-XGB): SMOTE-NC [24] is the Synthetic Minority Over-sampling Technique for Nominal and Continuous features that will generate synthetic samples for the minority class without affecting the structure of categorical features. XGB integrated with SMOTE-NC, denoted as SMOTE-NC-XGB, addresses class imbalance by generating more balanced data. The detailed hyperparameter tuning for SMOTE-NC-XGB on the number of nearest neighbors and resampling ratio is listed in Table IV.

TABLE IV. SMOTE-NC-XGB PARAMETERS

Parameter	Value	Description
learning_rate	0.1	Resampling the minority class to balance the dataset.
max_depth	7	Tree depth to control model complexity.
n_estimators	300	Number of boosting rounds.
subsample	0.8	Fraction of samples used for fitting each tree.

SMOTE-NC-XGB generates synthetic samples while maintaining the integrity of both categorical and continuous features, boosting recall, precision, and overall model performance in phishing website detection.

C. The XGBoost Classifier

XGBoost was selected because of its superior performance in classification tasks, especially for complex and imbalanced datasets, which is often the case in phishing website detection. Being a gradient boosting algorithm, XGBoost creates an ensemble of decision trees, improving predictive accuracy by iteratively adding models in a sequence [25]. It is particularly suited for phishing detection, where the challenge lies in identifying a minority class of phishing websites within a larger majority class of legitimate websites. XGBoost has, by default, mechanisms that handle class imbalances by adjusting the class weights, thus making the model focus more on the minority class. Moreover, in XGBoost, there is a possibility of extensive hyperparameter tuning, including but not limited to learning rate and tree depth,

which can be optimized for a particular dataset [26]. Its parallel processing ability combined with regularization techniques enhances the efficiency of the algorithm and helps avoid overfitting, therefore enhancing the robustness of the model [27]. These features together make XGBoost a very effective tool in combating class imbalance for the improvement of the performance of a phishing website detector.

IV. Proposed Framework

The presented work investigates some of the state-of-the-art class imbalance mitigation strategies in the phishing website detection problem by performing performance evaluation with respect to three different resampling techniques: Random Oversampling, Random Undersampling, and SMOTE-NC, when used together with the XGBoost classifier. Its main purpose is to look at the results that can be expected regarding major model performance metrics with such resampling techniques.

The proposed framework describes, in detail, the handling of class imbalance and the measurement of performance of the XGBoost classifier. Fig. 5 depicts the step-by-step data acquisition from the UCI database, analysis of distribution of classes for checking possible imbalances. Then, resampling techniques such as SMOTE-NC, ROS, and RUS are further used to balance the dataset.

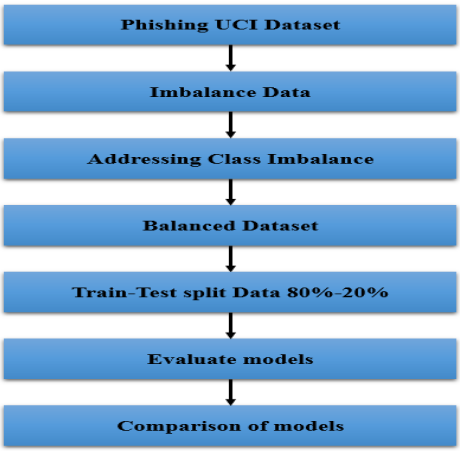


Fig. 5. Proposed framework.

After balancing, the dataset will then be divided into training and testing subsets. Here, the training subset is used in optimizing and training the XGBoost model while the test subset will be used in determining the classification accuracy of the model. Thereafter, after the training is done, the model classifies a new instance as either a phishing or a legitimate website. The key performance metrics used to measure the performance of the approach include accuracy, F1 score, recall, precision, ROC-AUC, and geometric mean score. These metrics also serve to underline the varying impact different resampling techniques have on overall model performance.

A. Class Distribution Analysis Before and After Resampling

Table V presents the class distributions of the original dataset and the balanced datasets after applying the resampling techniques.

TABLE V. The Class Distribution

Class	Legitimate	Phishing
Original Samples	4898	6157
Distribution (%)	44.30%	55.70%
ROS Samples	4926	4926
Distribution (%)	50.00%	50.00%
RUS Samples	3918	3918
Distribution (%)	50.00%	50.00%
SMOTE-NC Samples	4926	4926
Distribution (%)	50.00%	50.00%

This table is representing the balance of the imbalanced dataset in which phishing samples had a minority participation compared to the legitimate dataset. It is balanced using ROS, RUS, and SMOTE-NC. These all techniques are implemented to address the class imbalance such that the performance of the classifier XGBoost would improve in phishing web site detection so as to make a fair deal for both the classes: minority as well as the majority.

B. Evaluation Metrics

Evaluation metrics play an important role in machine learning for the measurement of model performance. It provides a quantitative measure of the performance of the model using various metrics: accuracy, precision, recall, F1 score, ROC-AUC, and Geometric Mean score. These metrics help the researchers in choosing the best approach for a particular task and hence ensure that the chosen model will effectively address the unique challenges of the problem.

1) Accuracy: Accuracy is the measure of the percentage of correctly predicted instances. Although useful, it can be misleading for imbalanced data sets as high accuracy might overemphasize the performance on the majority class.

$$Accuracy = \frac{Number\ of\ correctly\ predicted\ instances}{Total\ number\ of\ instances}$$

2) F1 Score: F1-score is the harmonic mean of precision and recall, effectively balancing the trade-off between false positives and false negatives.

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

3) Recall: Recall measures the proportion of actual positives correctly identified, making it crucial when minimizing false negatives is a priority.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

4) Precision: Precision quantifies the proportion of true positives among all predicted positives, and is particularly important when minimizing false positives is critical.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

5) ROC-AUC: ROC-AUC measures the model's ability to distinguish between classes. A higher value indicates better performance in differentiating between the positive and negative classes.

$$ROC - AUC = \int_0^1 TPR(FPR)d(FPR)$$

6) Geometric Mean Score (G-Mean): The geometric mean score equilibrates performance on both classes and is hence useful in the case of imbalanced datasets. It is calculated as the square root of the product of recall and specificity, offering a more balanced evaluation when class distribution is skewed.

$$\text{G} - \text{Mean} = \sqrt{Recall \times Specificity}$$

V. Findings and Analysis

This paper investigates the performance of three resampling techniques, namely ROS, RUS, and SMOTE-NC, integrated with the XGBoost classifier for phishing website detection. The key objective is to investigate how these techniques affect class distribution, improve predictive accuracy, and optimize key evaluation metrics.

Table VI provides the performance evaluation of the discussed models with respect to six performance metrics: accuracy, F1 score, recall, precision, ROC-AUC, and the Geometric Mean score.

TABLE VI. Evaluation Results in (%)

Classifier	Accuracy	F1 score	Recall	Precision	ROC-AUC	The Geometric Mean score
XGB	0.977	0.979	0.983	0.976	0.998	0.976
ROS-XGB	0.976	0.978	0.979	0.977	0.998	0.975
RUS-XGB	0.974	0.976	0.974	0.979	0.998	0.974
SMOTE-NC-XGB	0.980	0.982	0.985	0.979	0.998	0.979

SMOTE-NC-XGB gives the best general performance among the tested models, with the highest accuracy of 0.980, F1 score of 0.982, recall of 0.985, and Geometric Mean score of 0.979. It means that the SMOTE-NC-XGB model provides very well-balanced sensitivity and specificity while handling class imbalance. The XGB baseline also performs quite well, with high recall of 0.983 and a high ROC-AUC of 0.998. However, resampling techniques enormously increased the generalization of the model to imbalanced datasets.

- ROS-XGB slightly improved precision, at 0.977, from the baseline XGB, though in most metrics it performed the same as the SMOTE-NC-XGB model.
- RUS-XGB has better precision, 0.979, but it decreases the recall to 0.974 due to reduced false positives.
- SMOTE-NC-XGB outperforms all the other methods since it strikes the best balance among all metrics; thus, it is the most suitable approach for phishing website detection.

These findings really pinpoint the very crucial role that resampling techniques play in improving model performance, especially when dealing with imbalanced datasets. The results clearly indicate that SMOTE-NC-XGB emerges as the optimal method for phishing detection in this study.

VI. Conclusion

This paper discusses the performance comparison of undersampling, oversampling, and SMOTE-based techniques combined with the XGBoost classifier for class imbalance in phishing website detection. The experimental results illustrate that resampling methods greatly enhance the capability of a model in dealing with imbalanced datasets and improving the key evaluation metrics, including accuracy, F1 score, recall, precision, ROC-AUC, and Geometric Mean score.

Among the models tested, SMOTE-NC-XGB has always been at the top for all metrics with an accuracy of 98.0% and a recall of 98.5%, values which are rather high, therefore, it is able to balance sensitivity and specificity, which is the very key problem in phishing website detection; either false positives or false negatives can cause huge losses.

While the XGB baseline is also good to go, at 99.8% ROC-AUC, the introduction of resampling techniques into its training gives the model extra strength, particularly under strong class imbalance. Both Random Oversampling (ROS) and Random Undersampling (RUS) turn in competitive performances; the former slightly improves precision, while the latter is able to strongly reduce false positives.

In a nutshell, the results of this work highlight the key contribution of resampling techniques in enhancing the performance of models on imbalanced datasets. Among the considered methods, SMOTE-NC-XGB is the most balanced and reliable, hence presenting a promising solution for the enhancement of phishing website detection systems. Future work may consider hybrid resampling methods or more advanced methods to further optimize detection performance.

References

[1] M. S. Kheruddin, M. A. E. M. Zuber, and M. M. M. Radzai, "Phishing Attacks: Unraveling Tactics, Threats, and Defenses in the Cybersecurity Landscape," TechRxiv, Jan. 15, 2024. DOI: 10.22541/au.170534654.48067877/v1.

[2] I. Araf, A. Idri, and I. Chairi, "Cost-sensitive learning for imbalanced medical data: A review," Artificial Intelligence Review, vol. 57, no. 4, p. 80, 2024. DOI: 10.1007/s10462-023-10652-8.

[3] A. Aldoseri, K. N. Al-Khalifa, and A. M. Hamouda, "Re-thinking data strategy and integration for artificial intelligence: Concepts, opportunities, and challenges," Applied Sciences, vol. 13, no. 12, p. 7082, 2023. DOI: 10.3390/app13127082.

[4] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," Progress in Artificial Intelligence, vol. 5, no. 4, pp. 221–232, 2016. DOI: 10.1007/s13748-016-0094-0.

[5] M. Lokanan, "Exploring Resampling Techniques in Credit Card Default Prediction," Research Square, Preprint, 15 Mar. 2024. DOI: 10.21203/rs.3.rs-4087259/v1.

[6] Q. Wei and R. L. Dunbrack Jr., "The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics," PLoS ONE, vol. 8, no. 7, p. e67863, Jul. 2013. DOI: 10.1371/journal.pone.0067863.

[7] M. Galar, A. Fernández, E. Barrenechea, and F. Herrera, "EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling," Pattern Recognition, vol. 46, no. 12, pp. 3460-3471, 2013. DOI: 10.1016/j.patcog.2013.05.006.

[8] M. Shelke, P. R. Deshmukh, and V. K. Shandilya, "A review on imbalanced data handling using undersampling and oversampling technique," International Journal of Recent Trends in Engineering, 2017. DOI: 10.3883/IJRTER.2017.3168.0UWXM.

[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, Aug. 2016, pp. 785-794. Association for Computing Machinery, doi: 10.1145/2939672.2939785.

[10] L. Yu, S. Wang, and K. K. Lai, "An integrated data preparation scheme for neural network data analysis," IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 2, pp. 217–230, 2006, doi: 10.1109/TKDE.2006.22.

[11] S. Angra and S. Ahuja, "Machine learning and its applications: A review," in 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), 2017, pp. 57–60, doi: 10.1109/ICBDACI.2017.8070809.

[12] M. J. Nigrini, "The patterns of the numbers used in occupational fraud schemes," *Managerial Auditing Journal*, vol. 34, no. 5, pp. 606–626, 2019, doi: 10.1108/MAJ-11-2017-1717.

[13] V. Mirchevska, M. Luštrek, and M. Gams, "Combining domain knowledge and machine learning for robust fall detection," *Expert Systems*, vol. 31, no. 2, pp. 163–175, 2014, doi: 10.1111/exsy.12019.

[14] T. Wongvorachan, S. He, and O. Bulut, "A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining," *Information*, vol. 14, no. 1, p. 54, 2023, doi: 10.3390/info14010054.

[15] G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data," *Data Mining and Knowledge Discovery*, vol. 28, no. 1, pp. 92–122, 2014.

[16] S. J. Dattagupta, "A performance comparison of oversampling methods for data generation in imbalanced learning tasks," Ph.D. thesis, Universidade Nova de Lisboa, Lisbon, Portugal, 2018.

[17] M. Mukherjee and M. Khushi, "SMOTE-ENC: A Novel SMOTE-Based Method to Generate Synthetic Data for Nominal and Continuous Features," *Applied System Innovation*, vol. 4, no. 1, p. 18, 2021, doi: 10.3390/asi4010018.

[18] W. W. Islahulhaq and I. D. Ratih, "Classification of Non-Performing Financing Using Logistic Regression and Synthetic Minority Oversampling Technique-Nominal Continuous (SMOTE-NC)," *International Journal of Advanced Soft Computing and Its Applications*, vol. 13, pp. 116–128, 2021.

[19] UCI Machine Learning Repository, "Phishing Websites Data Set," Available online: https://archive.ics.uci.edu/ml/datasets/phishing+websites.

[20] K. Omari, "Comparative study of machine learning algorithms for phishing website detection," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 9, 2023.

[21] Lema, G., "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2017.

[22] More, A., "Survey of resampling techniques for improving classification performance in unbalanced datasets," *arXiv preprint*, 2016. doi: 10.48550/ARXIV.1608.06048.

[23] R. Zuech, J. Hancock, and T. M. Khoshgoftaar, "Detecting Web Attacks Using Random Undersampling and Ensemble Learners," *Journal of Big Data*, vol. 8, no. 1, p. 75, 2021. doi: 10.1186/s40537-021-00460-8.

[24] Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P., "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[25] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794. Association for Computing Machinery, doi: 10.1145/2939672.2939785.

[26] T. Kavzoglu and A. Teke, "Advanced hyperparameter optimization for improved spatial prediction of shallow landslides using extreme gradient boosting (XGBoost)," *Bulletin of Engineering Geology and the Environment*, vol. 81, Article 201, 2022. doi: 10.1007/s10064-022-02708-w.

[27] S. S. Dhaliwal, A. A. Nahid, and R. Abbas, "Effective Intrusion Detection System Using XGBoost," *Information*, vol. 9, no. 7, p. 149, 2018. doi: 10.3390/info9070149.

Deep Learning-Driven Detection of Terrorism Threats from Tweets Using DistilBERT and DNN

Divya S[1]*, B Ben Sujitha[2]
Research Scholar-Department of Computer Science and Engineering,
Noorul Islam Centre for Higher Education, Kumaracoil, Thuckalay, Tamil Nadu, India[1]
Professor-Department of Computer Science and Engineering,
Noorul Islam Centre for Higher Education, Kumaracoil, Thuckalay, Tamil Nadu, India[2]

***Abstract*—As globalization accelerates, the threat of terrorist attacks poses serious challenges to national security and public safety. Traditional detection methods rely heavily on manual monitoring and rule-based surveillance, which lack scalability, adaptability, and efficiency in handling large volumes of real-time social media data. These approaches often struggle with identifying evolving threats, processing unstructured text, and distinguishing between genuine threats and misleading information, leading to delays in response and potential security lapses. To address these challenges, this study presents an advanced terrorism threat detection model that leverages DistilBERT with a Deep Neural Network (DNN) to classify Twitter data. The proposed approach efficiently extracts contextual and semantic information from textual content, enhancing the identification of potential terrorist threats. DistilBERT, a lightweight variant of BERT, is employed for its ability to process large volumes of text while maintaining high accuracy. The extracted embeddings are further analyzed using a Dense Neural Network, which excels at recognizing complex patterns. The model was trained and evaluated on a labeled dataset of tweets, achieving an impressive 93% accuracy. Experimental results demonstrate the model's reliability in distinguishing between threatening and non-threatening tweets, making it an effective tool for early detection and real-time surveillance of terrorism-related content on social media. The findings highlight the potential of deep learning and natural language processing (NLP) in automated threat identification, surpassing traditional machine learning approaches. By integrating advanced NLP techniques, this model contributes to enhancing public safety, national security, and counter-terrorism efforts.**

***Keywords*—Terrorism; global safety; terrorist attacks; data mining; artificial intelligence; natural language processing; DistilBERT; deep neural network**

I. INTRODUCTION

An act of violence against individuals committed to further political or ideological goals is frequently referred to as terrorism [1]. Terrorism aims to erode the rule of law, democracy, and human rights. Terrorism has a direct influence on several human rights, including life, liberty, and bodily integrity [2]. The study of terrorism does not have strict boundaries. It covers the full spectrum of human emotions, attitudes, and behaviours, incorporating a diversity of perspectives along with the associated fears and inclinations.

There has been terrorism- related fatalities in almost every part of the world. Terrorism may harm civil society, impair social and economic advancements, destroy governments, and threaten peace and security [3]. Each of these elements greatly impacts the enjoyment of human rights. Recently, states have implemented counterterrorism measures that have presented significant challenges to human rights and the rule of law [4]. As a counterterrorism strategy, certain regimes have employed torture along with other cruel treatment, often ignoring the legal and practical protections against assaults, like regular, unbiased inspections of detection facilities. Certain states have violated the international legal obligations of non-refoulement by returning suspected terrorists to countries where they truly face the risk of torture or other grave violations of human rights. In certain situations, the use of special courts to try civilians has undermined the independence of the judiciary and reduced the efficacy of traditional court systems. Repressive measures have suppressed the voices of journalists, human rights advocates, indigenous populations, minorities, and civil society. The economic, social, and cultural rights of many people have suffered as a result of the security sector receiving funds that were frequently intended for social initiatives and development assistance.

The adversary does not publicly announce or carry out the attack, and no enemy aircraft fly over the targets. Instead, all that is left are the horrific images of building collapses, plane hijacking, and innocent people murdered or wounded in bombings and shootings that have caused fear and panic around the world. Attacks of this kind have been referred to as "terrorism" [5]. This is not a struggle between states: rather it is a new kind of terror warfare. The enemy may not be a nation-state but rather a small terrorist group based mostly in a developing country. For the sake of their cause, they are willing to give their lives. Thus, it is clear that they are inaccessible to the police and too elusive for the military to hunt down. Conventional forces are designed to fight, they are not designed for unexpected peaceful situations.

Terrorist attack detection has become essential in the contemporary global context due to the growing frequency, sophistication, and unpredictability of terrorist attacks, which pose major threats to national security, public safety, and economic stability. Rapidly developing communication technologies, like the internet and social media have provided terrorists with platforms to plot and execute attacks, therefore,

early discovery is crucial to reducing such dangers. Traditional methods of detecting terrorists' activity, such as surveillance and intelligence gathering often fall short because of the volume of information, the complexity of attack preparation, and the secretive character of modern terrorist activities. Real-time analysis of large datasets, such as text, images, videos and audio has demonstrated extraordinary proficiency in identifying patterns suggestive of terrorist operations using Deep Learning (DL) algorithms [6]. DL is capable of efficiently analyzing unstructured data and producing insights that were previously unavailable using traditional techniques. This aids security forces in stopping attacks before they are carried out and makes it easier to identify terrorists' activities quickly. DL can manage complex, multidimensional data and provide immediate practical insights in a constantly shifting danger scenario, making it an essential tool for terrorist attack detection in the current era.

The detection of terrorism threats from tweets has been transformed by the combination of DL and Natural Language Processing (NLP), which allows for the accurate and efficient analysis of large volumes of unstructured text data [7]. NLP methods enable the extraction of context, sentiment, and linguistic patterns, whereas DL models are excellent at identifying subtleties and intricate relationships in text. These methods are essential for recognizing keywords, spotting extreme language, and quickly detecting hidden risks. Their capacity to adjust and pick up on changing linguistic patterns improves prediction skills, which is crucial for counterterrorism operations and public safety. This paper proposes an effective terrorism threat detection model utilizing the advantages of DL and NLP. The major contributions of the paper include:

- To prepare tweet data using efficient text preprocessing methods, ensuring high- quality input for the detection algorithm.
- To effectively tokenize and extract features from tweet data by utilizing the transformer- based architecture of DistilBERT tokenizer, which captures contextual meaning and relationships.
- To design a system for detecting terrorism threats by combining DistilBERT embeddings with a Deep Neural Network (DNN) model, with the goal of accurately classifying tweets as either non- threats or threats.

The remaining sections of the research are arranged as follows: Section II gives a summary of the current studies, highlighting areas that need more investigation. Section III offers a comprehensive explanation of the methodology. Section IV offers the findings derived from the suggested methodology in detail. A discussion is provided in Section V and finally, a summary of the findings is included in Section VI, which gives a conclusion to the paper.

II. Literature Review

Shinde et al. [8] employed data from the Global Terrorism Database (GTD) to create a model for displaying the aspects of terrorist acts. The study used two different graph embedding methods and seven Machine Learning (ML) models to sort the data into groups. These models included K- Nearest Neighbour (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF) and adaboost. The model achieved 90% accuracy. The study did not include the model's generalizability and scalability. An et al. [9] conducted a study to predict and examine the impact of microblogging on terrorist incidents. Emotion analysis was conducted utilizing user, time and feature content and topics were created using a combination of Word2Vec and K means clustering. The proposed Logistic Regression (LR) based approach outperformed six existing classification models with an accuracy of 85% and was able to identify high influence attributes. Jyothilinga [10] used the GTD's historical terrorist attack data to predict future occurrences and causalities. The study made use of ML techniques such as RF, Multilayer Perceptron (MLP) regressors for predicting deaths, DT and MLP classifiers for classifying offenders, and KNN classifiers for classifying weapons. Using time series analysis to predict terrorists' attacks, KNN achieved 92.25% accuracy, while DT and MLP both produced 90 % accuracy for classifying offenders. The study recognized certain limitations, though including the intricacy of terrorist incidents and a lack of thorough data necessary for precise model validation and training.

Gaikwad et al. [11] created the Merged Islamic State of Iraq and Syria/ jihadist White Supremacist (MIWS) dataset with the goal of examining the impact of extremist groups on social media. The effective use of NLP with pretrained models such as BERT, RoBERTa and DistilBERT demonstrated the effective application of DL for identifying extremist content across many approaches with BERT achieved an F1- score of 0.72. Nevertheless, it was pointed out that the dataset relies on terms that can result in false positives from non- extremist sources such as journalists and critics. Yi Feng et al. [12] designed RP-GA-XGBoost, a causality prediction system based on XGBoost to determine if terrorist acts will cause the deaths of innocent people. The proposed approach integrated Principal Component Analysis (PCA) and RF for feature selection and utilized a Genetic Algorithm (GA) to optimize the hyperparameters of XGBoost. The findings showed that, when compared to some- well known ML techniques, the suggested approach performed the best. Ruifang Zhao et al. [13] combined Inverse Distance Weighting (IDW) with a Multilabel K- Nearest (I-MLKNN) based evaluation framework in order to assess terrorist activities. The dimensions of the features of terrorist attack categories were reduced using the Locally Linear Embedding (LLE) dimension reduction technique. Before and after the dimension reduction, the correlation between the attributes was assessed using the Maximal Information Coefficient (MIC). The inverse distance weighting method was subsequently integrated with the MLKNN multi-label classification algorithm on a grid framework. The outcomes of the simulation showed that the I-MLKNN multi-label classification approach is a perfect tool for determining the geographic distribution of terrorist acts.

Lanjun Luo and Chao Qi [14] focused on determining important markers that influence the likelihood of terrorist attacks from a predictive approach. Factors at the event level and the root cause are considered, and they are qualified using

28 indicators. A recursive feature elimination technique using RF kernels was suggested to find the most important ones among the 28 initial signs. The simulation findings demonstrated that the indicators severely degrade when the bare minimum of input indicators was used before the prediction performance. A method for identifying statistically significant change points in time series connected to terrorism that could indicate the onset of events that require attention was developed by Theodosiadou et al. [15]. Before and after the dimension reduction, the correlation between the attributes was assessed using the Maximal Information Coefficient (MIC). The potential of the suggested methodology in successfully identifying such changes was demonstrated by applying it to a real- world dataset. Change point detection techniques can be used to estimate statistically significant changes in the structural behavior of the aforementioned indicators at particular time points by analyzing the resulting time series. Shynar Mussiraliyeva et al. [16] suggested text categorization methods for identifying extremist activity using MLP technologies. The produced corpus was divided into two sections: 3000 words of postings with extremist intent and 15,000 words of non- extremist posts, including news portals and religious materials. According to the simulation findings, the suggested approach classified extremist texts from the collected corpus with a high degree of accuracy.

Ghada M. A. Soliman and Tarek H. M. Abou- El- Enien [17] designed a hybrid computationally intelligent system to assist in making decisions on the subject of terrorism. The suggested hybrid prediction algorithm combined Data Mining (DM) approaches with various Operations Research (OR) and decision support tools. The development, implementation and evaluation of the suggested system have utilized a variety of assessment metrics. The experimental findings showed that the suggested method identify the terrorist group that is responsible for terror strikes in different places. Georgios Koutidis et al. [18] developed a framework based on ML and DL models for predicting the probability of terrorist occurrences in a target region over a certain time frame. Based on RF models, a feature selection procedure was suggested in which each feature's predictive power is determined and only features that are thought to have adequate predictive power are then taken into consideration. The simulation findings showed that the suggested framework improved the predictive ability of the ML and DL models with the suggested feature selection in predicting the probability of terrorist attacks. Ben Chaabene et al. [19] developed a computational model that uses a variety of ML and recommender system techniques to predict how terrorists' actions will affect social networks based on the text and image data. The suggested method comprises two models: the text classification framework and the image classification system. The text classification model comprises LR, NB and SVM. The model for classifying images was based on CNN. The simulation results showed that the suggested model successfully identified and predicted the behaviour of militant terrorists on twitter.

The field of ML and DL techniques for terrorist attack prediction has advanced significantly in recent years, yet there is still a notable gap in the existing literature. Nations with insufficient or missing data find it difficult to understand terrorist behaviour and highlighting particular groups or ideas results in mispredictions. The socio- political, economic and psychological elements that influence terrorism are not adequately represented by current models. Despite their higher accuracy, many models generate false positives, misclassify harmless behaviours associated with terrorism, and create ethical concerns like discriminating against the community. Moreover, it is challenging to comprehend ML models decision- making processes because of their complexity. Models that can adjust and learn from new data in real time are necessary due to the dynamic nature of terrorist actions, but many previous studies rely on static dataset, which reduces their applicability in situations that change rapidly. Overcoming these gaps is crucial to improving the effectiveness and moral applications of DL and ML in predicting terrorist activities, which will ultimately lead to safer communities across the world. So, in order to rectify the above-mentioned limitations, a novel NLP model based on DistilBERT and DNN for the detection of terrorism threats from tweets has been proposed in this paper.

III. Materials and Methods

The proposed approach combines DistilBERT with Deep Neural Network (DNN) in a structured pipeline to identify terrorist threats in tweets. To standardize and clean the input text, raw twitter data is initially collected and processed through preprocessing approaches that include removing URLs, mentions, hashtags, numbers, punctuation and lowercase texts, as well as stop words. The DistilBERT tokenizer tokenizes the processed text, transforming it into attention masks and input IDs. A DistilBERT model with numerous transformer layer, pre- trained using these representations, captures contextual embeddings of the input text. The DistilBERT output embeddings are sent to a DNN classification head, which is made up of Fully Connected (FC) layers that are intended to train discriminative features for classification. Finally, the model provides a reliable and effective method for detecting terrorist threats by predicting whether a tweet is a threat or not. The detailed block diagram of the suggested terrorism threat detection model is visualized in Fig. 1.

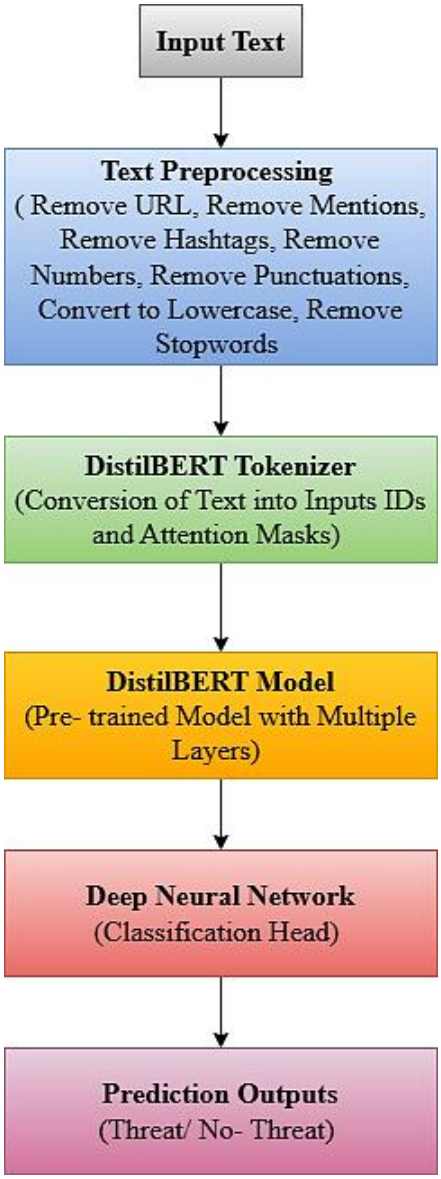


Fig. 1. Block diagram of suggested terrorism threat detection system.

A. *Dataset Description and Exploratory Data Analysis*

The textual data in the dataset was gathered from social media, particularly tweets and was classified as either indicating non- threating or threatening material. The almost equal distribution of these classes is crucial for maintaining the balance of classification tasks. Preprocessing eliminates features like mentions (@username), hashtags (#hashtag) and URLs for a cleaner analysis. The tweets range significantly in length from mere lines to extensive paragraphs. The dataset also includes linguistic variables that aid in distinguishing between threatening and non- threatening tweets, such as average word length and frequency. The dataset is downloaded from GitHub [20]. Fig. 2 shows the dataset sample.

Dataset Head:

	Tweet	Class
0	PT announced his death. In current tweet shows...	1
1	The Muslim Brotherhood of #Iraq condemns the k...	1
2	RT @TedNugent: Gun control talks with gunrunni...	0
3	Come home from work to Bridesmaids, Twinkies a...	0
4	RT @Ghostmanzzz: #Breaking #US airstrikes rock...	1

Fig. 2. Sample dataset.

Each class in the text dataset for terrorist threat detection contains the same number of samples, indicating a balanced class distribution. The number of tweets classified as threatening (Class 1) and non- threatening (Class 0) is 5265 respectively. This balanced distribution promotes accurate predictions for both threatening and non- threatening tweets and allows for equitable learning during training by preventing the model from being biased toward any one class. The class distribution is plotted in Fig. 3.

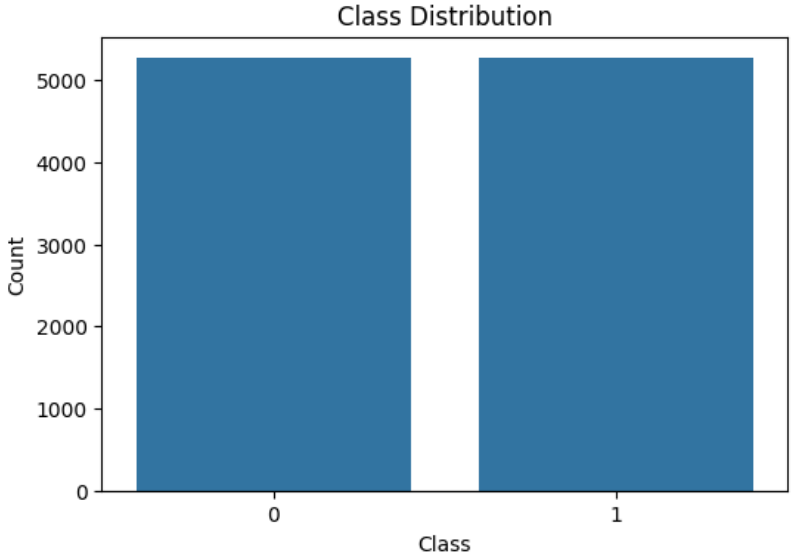
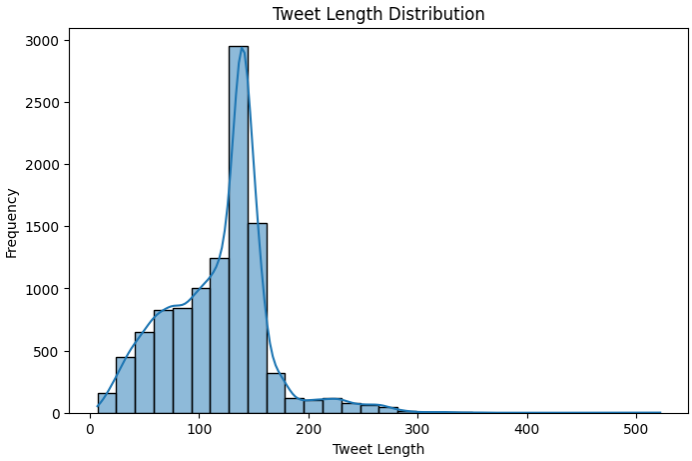
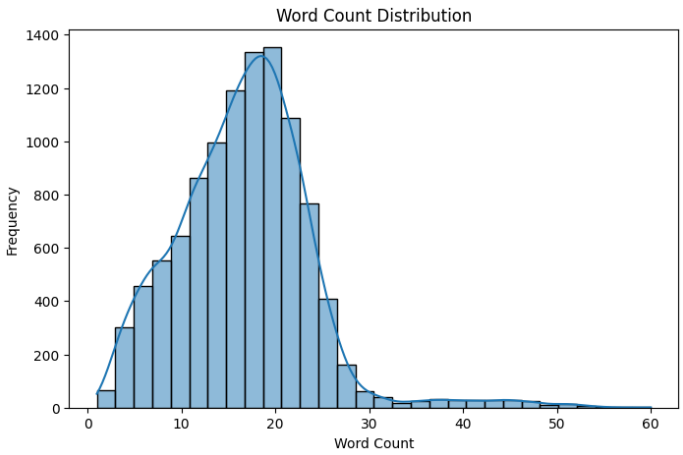


Fig. 3. Class distribution.

The statistical distribution of the character count in each tweet in the dataset is known as the “Tweet Length Distribution” and is displayed in Fig. 4 (a). It provides information about the range of tweet sizes by displaying how lengthy or short the tweets are. The statistical distribution of the word count in each tweet within the dataset is referred to as the “Word Count Distribution” and is displayed in Fig. 4 (b). It provides insights into the typical word count in a tweet, ranging from brief tweets with few words to longer ones with numerous words. The "Average Word Length Distribution" is the distribution of the average number of characters per word in each tweet in the dataset, as shown in Fig. 4(c). It gives information about the complexity of the language used in the tweets, including whether individuals typically use longer, more complicated phrases or shorter, simpler ones. As seen in Fig. 4 (d), the “Hashtag Count Distribution” represents the distribution of hashtags used in each tweet across the datasets. It shows the frequency with which people add hashtags to their tweets, ranging from one to multiple per tweet. The “Mention Count Distribution” represents the distribution of mentions (i.e., tagging another user with “@username”) in each tweet in the dataset, as illustrated in Fig. 4 (e). This distribution shows how frequently users mention or engage with other people in their tweets.



(a)



(b)

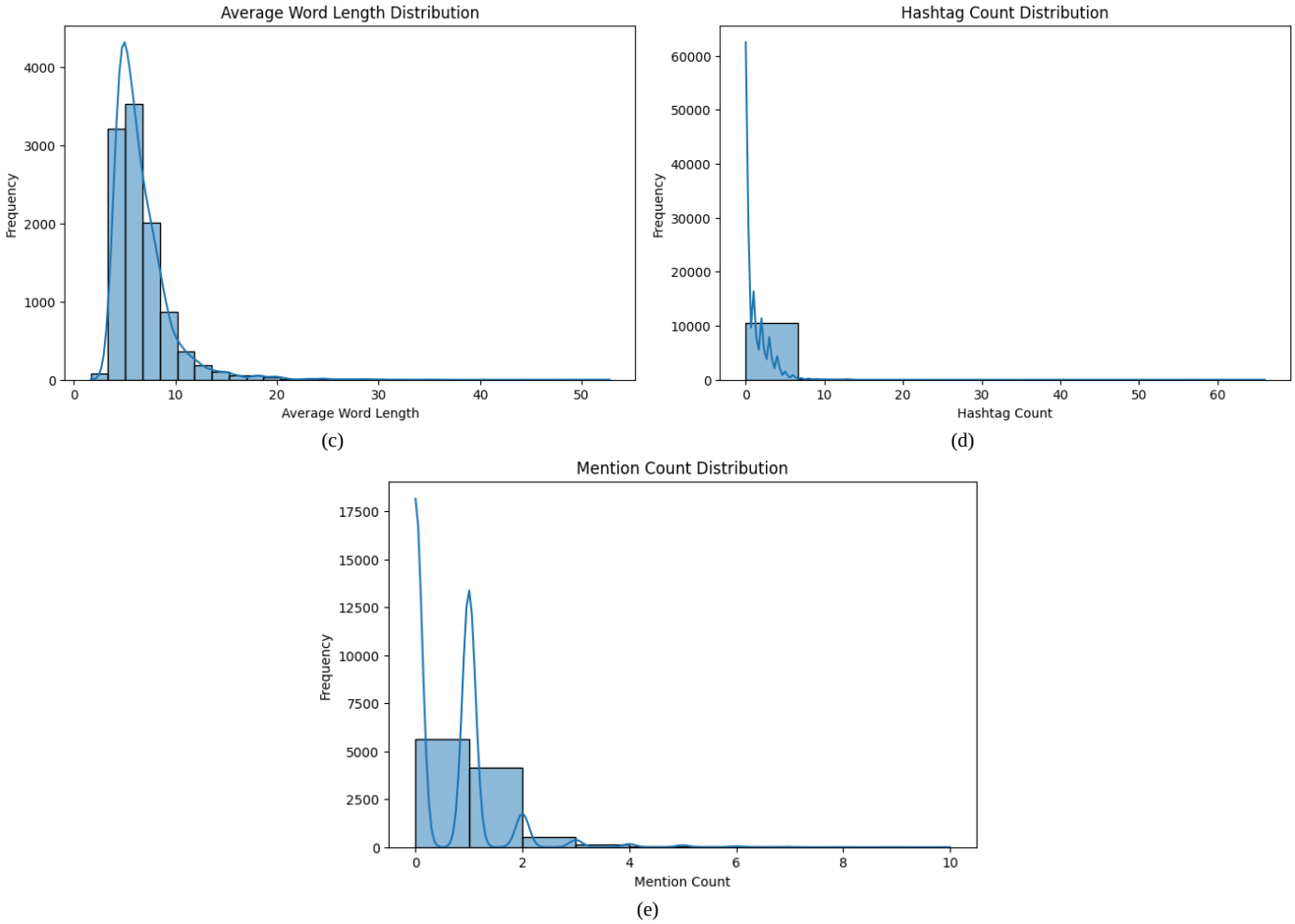
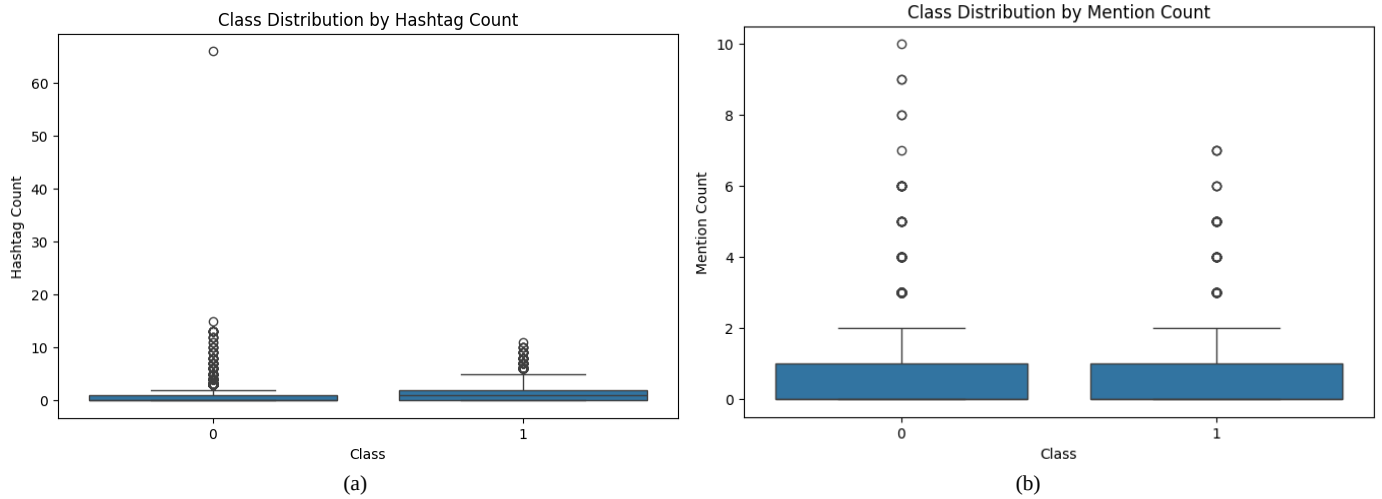
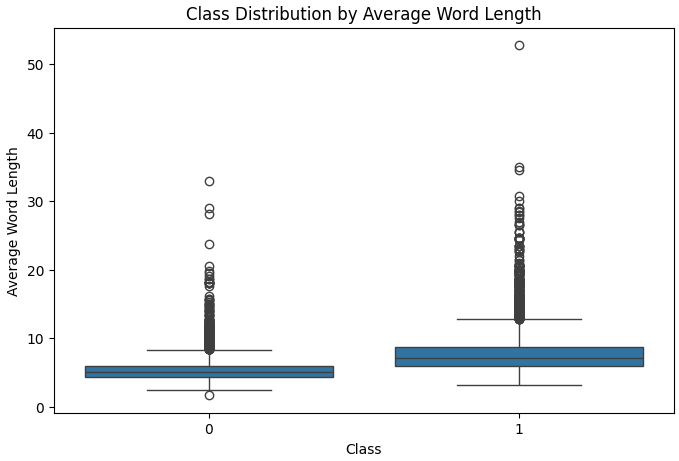


Fig. 4. (a) Tweet length distribution (b) Word count distribution (c) Average word length distribution (d) Hashtag count distribution (e) Mention count distribution.

As illustrated in Fig. 5 (a), "Class Distribution by Hashtag Count" is an analysis that looks at the relationship between a tweet's hashtag count and its classification as "threat" or "non-threat". Box plots commonly compare the quantity of hashtags for each class labels when displaying this kind of distribution.





(c)

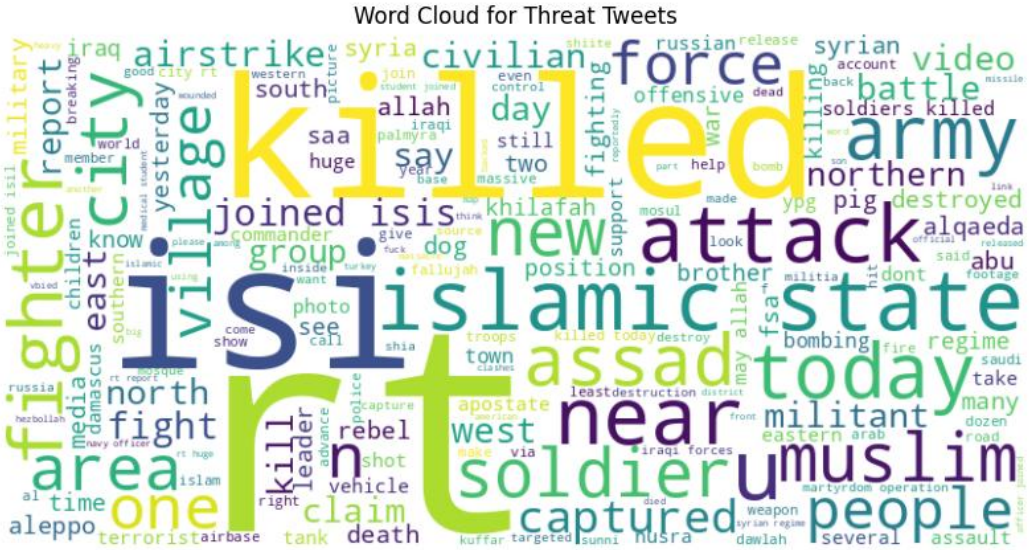
Fig. 5. (a) Class distribution by hashtag count (b) Class distribution by mention count (c) Class distribution by average word length.

It provides information about whether tweets with a threat label typically contain more or fewer hashtags compared to tweets without a threat label. The analysis known as "Class Distribution by Mention Count" examines the frequency of mentions of a user, such as "@username", in tweets classified as "threat" or "non- threat". Box plots, as seen in Fig. 5 (b) are frequently used to illustrate this kind of distribution. They display the median, quartiles, and overall range of mention counts for each class. An association between the quantity of mentions in a tweet and its probability of being categorized as threatening can be ascertained by comparing these distributions. "Class Distribution by Average Word Length" examines the differences in average word length between the two classes of tweets. Fig. 5 (c) display this distribution using box plots. They show the median, quartiles and range of average word lengths for each class. It is feasible to determine whether the complexity of language used in threatening versus non- threatening tweets differs by comparing these distributions.

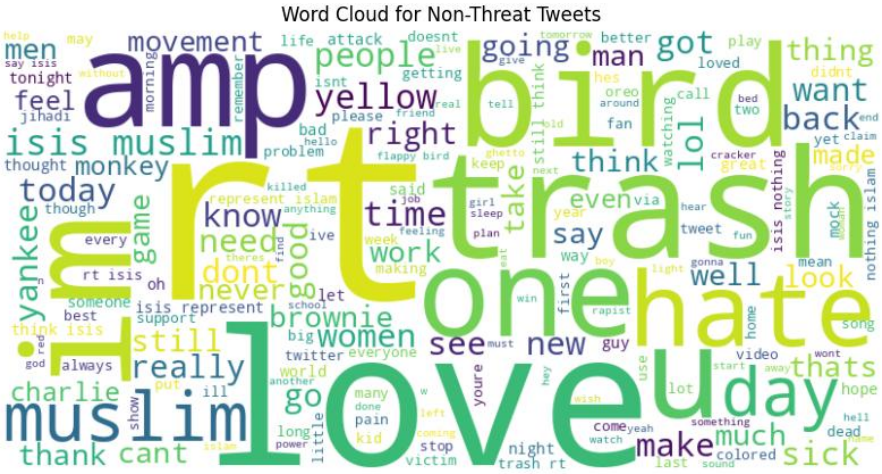
The most common word in a corpus of text is represented visually in a word cloud, where the size of each word indicates its significance or frequency. It is possible to create distinct word clouds for threat and non- threat tweets in order to emphasize the unique terminology utilized in each category. The word cloud for threat and non- threat tweets is displayed in Fig. 6 (a) and Fig. 6 (b).

B. *Text Preprocessing*

In NLP, text preprocessing is an essential step that cleans and gets raw data ready for analysis. A variety of methods standardize and simplify the text. URLs, hashtags, and mentions are eliminated in order to remove unnecessary components that are not beneficial to the semantic meaning. Eliminating punctuations and numbers ensures that the words stay the main emphasis, which lowers noise. Text conversion to lowercase standardizes data and prevents case- sensitivity induced duplication. Stopwords like "is", "the" and "and" which are often used words that don't significantly advance the analysis, can be removed. When combined, these strategies improve the text's quality and make it more appropriate for NLP tasks.



(a) Threat tweet.



(b) Non-threat tweet.

Fig. 6. Word cloud.

C. Tokenization using DistilBERT Tokenizer

The process of tokenization involves converting unprocessed text data, such as tweets, into a format that the DistilBERT model understands [21]. Since DistilBERT is a transformer- based model, incoming text must be divided into tokens, which are effectively smaller subwords or units. The DistilBERT tokenizer manages this process by dividing words into tokens, assigning a token ID to each token and generating input sequences that sum up the tweet. In addition, the tokenizer uses truncation to eliminate extra tokens that exceed the allowable length and padding to ensure that every sequence has a constant length. This phase is crucial for preparing the data for feeding into the model, which will enable the transformer to retrieve the text effectively and comprehend the context of each tweet. An example of tokenization operation is visualized in Fig. 7.

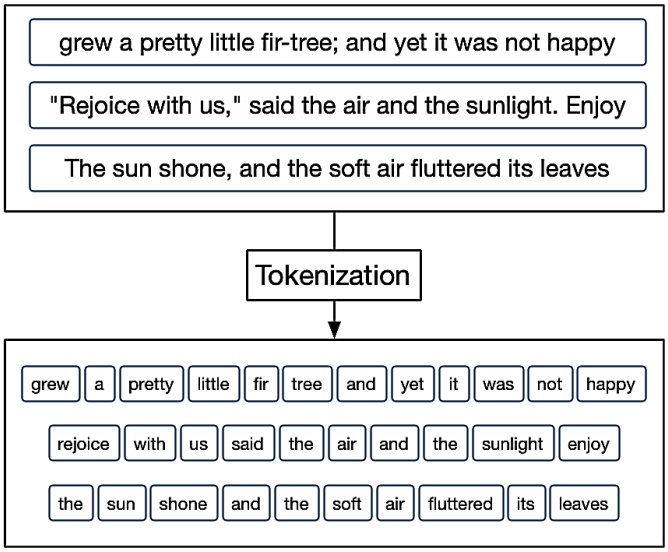


Fig. 7. An illustration of tokenization.

D. Proposed Terrorism Threat Detection Model Based on DistilBERT and DNN

DistilBERT is a faster and more compact version of Bidirectional Encoder Representations from Transformer (BERT) designed to be more efficient on devices with limited processing power. Using a method called knowledge distillation, the BERT model is shrunk by teaching a smaller model to behave similarly to a bigger one [22]. DistilBERT is more efficient due to its reduced size and simplified training procedure. It uses a simplified transformer architecture, notably removing the token type embeddings from the input and the pooling layer from the output. Fig. 8 visualizes the DistilBERT architecture.

The primary goal of DistilBERT is to preserve the majority of BERT's performance while lowering its memory and computational costs. DistilBERT accomplishes this through:

- DistilBERT employs a smaller architecture than BERT with fewer layers and hidden components. In particular, it features 6 layers rather than 12 and 2048 hidden units rather than 7680 for the base model.
- Training through Knowledge Distillation: DistilBERT is trained to replicate the behaviors of the larger BERT model via a strategy called knowledge distillation. To train DistilBERT, the outputs of BERT are employed as soft targets.

The general structure of DistilBERT is similar to that of BERT, despite having fewer layers and hidden components. A feed-forward neural network with a self-attention mechanism is included within each layer of DistilBERT's multi-layer bidirectional transformer encoder. The self-attention mechanism enables the model to focus on different components of the input sequence and identify connections among words. DistilBERT first embeds a sequence of tokens in a high dimensional vector space as its input [23]. The transformer encoder creates contextualized representations for every token by further processing these embeddings. A classification or regression layer processes the contextualized representations to predict the outcome.

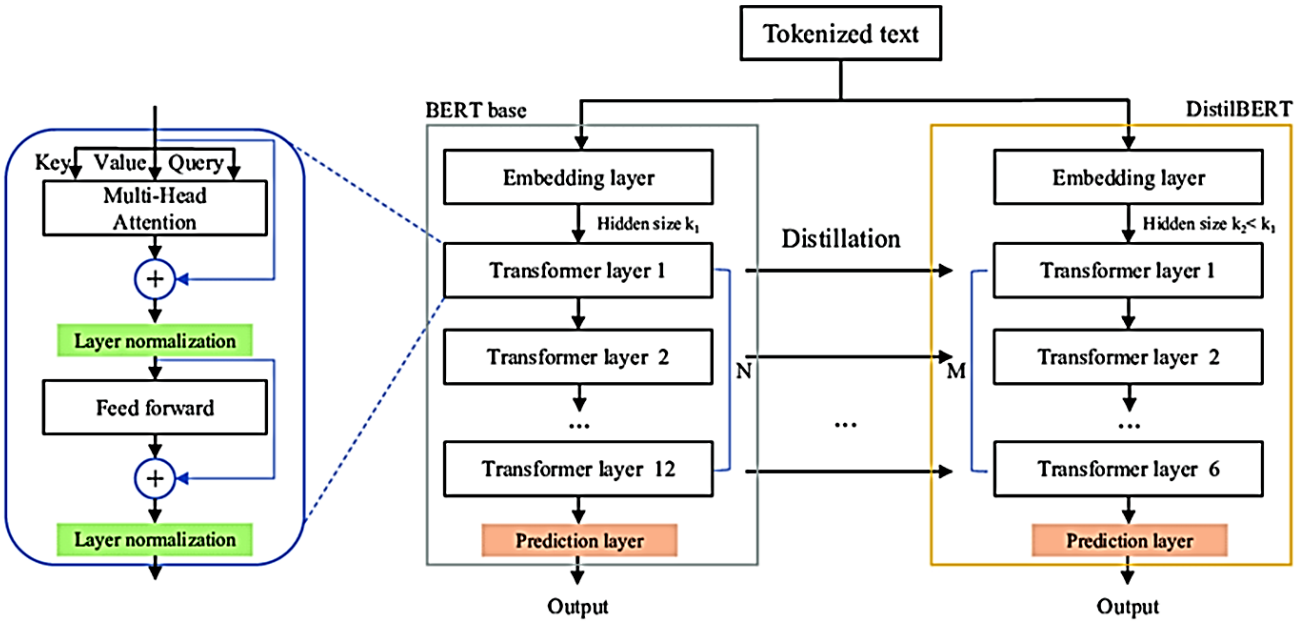


Fig. 8. DistilBERT architecture.

The tokenized input text is fed into both DistilBERT and BERT. The tokenization process divides the input text into smaller tokens, transforms them into IDs and ensures that each input sequence has a constant length. This uniform input is necessary for both BERT and DistilBERT to process the data effectively. The transformer layer and the embedding layer are two of the layers that compose the BERT base architecture. The embedding layer provides numerical representations of words or subwords by transforming input tokens into dense vectors with specified dimensions. BERT comprises 12 transformer layers, including feed- forward, multi- head attention and layer normalization [24]. Each transformer layer can focus on different parts of the sequence at the same time due to a multi- head attention mechanism that accepts three inputs: key, query and value. Layer normalization normalizes the output of the multi- head attention to maintain stability throughout the training process. The feed- forward layer extracts higher level characteristics from the output of the attention mechanism by applying a FC layer. The feed-forward procedure is followed by another normalization step. The structure of DistilBERT is similar to that of BERT, although there are a few significant differences. Similar to BERT, an embedding layer transforms tokens into dense vectors. DistilBERT employs six transformer layers instead of twelve, reducing the number of levels. Its distillation training approach allows it to preserve most of the BERTs context-handling capabilities despite having fewer layers. A major factor in DistilBERT's lightweight design is that its hidden size (k_2) is usually smaller than BERT' s (k_1).

Eq. (1) calculates the scaled dot-product attention, the fundamental component of the attention process.

$$Attention\ (\ Q,K,V) = softmax\ \left(\frac{QK^T}{\sqrt{d_k}}\right) V \qquad (1)$$

The input generates three separate representations: Q, K, and V. The three variables possess dimensions of $n \times d$, with d representing the dimensionality of each token and n indicating the sequence length (number of tokens). The model produces attention scores by evaluating each query against every key. The attention scores are evaluated employing the dot product of the Q and K metrics and the square root of the dimensionality d_k. These scores highlight the level of attention that each component of the input should receive when producing the output. The attention mechanism then uses the attention scores to determine the weighted sum of the values. The system may retrieve all types of relationships in the data because multiple heads produce queries, keys, and values concurrently. The multi- head attention mechanism then creates the final output by combining the individual outputs of each attention head. A prediction layer is applied to the output of the transformer layer. This layer is utilized for a variety of NLP tasks. The output layer of DistilBERT is essentially the same as that of BERT, but because of its simplified architecture and fewer transformer layer, it utilizes fewer parameters.

Specifically for binary classification tasks, the output layer of the DistilBERT is the DNN with the sigmoid layer, which generates a final prediction [25]. Fig. 9 illustrates that a dense layer, also referred to as a FC layer, connects every input node to every output node in a neural network. This dense layer in the DistilBERT context receives the output from the global average pooling operation or earlier transformation layers. The transformer layers extract high- level characteristics and send them to the dense layer, which applies an activation function after a linear transformation.

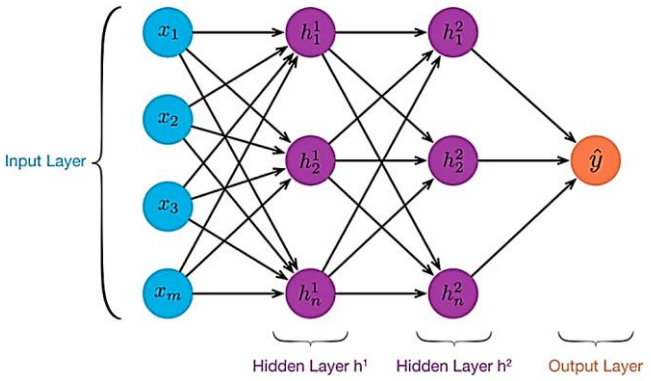


Fig. 9. Deep neural network.

The output layer uses the sigmoid activation function to classify data into binary categories. The sigmoid function is appropriate for situations where the objective is to categorize the input into one of two groups since it converts the result to a number between 0 and 1. In the classification of threats versus non- threats, a sigmoid output around 1 would suggest a high likelihood that the input tweet is a threat, whereas a value near 0 would suggest a non- threat. The sigmoid function is mathematically expressed as in Eq. (2).

$$Sigmoid\ (x) = \frac{1}{1+e^{-x}} \quad (2)$$

The embedding layer and transformer levels receive tokenized text data (viz multi- head attention and feed- forward sublayers). Next, a global average pooling layer usually shrinks the sequence of vectors into a single vector of fixed length that represents the whole sequence. Next, a dense layer employing a sigmoid activation function receives this vector. The dense layer then produces a single probability score that indicates the probability that the input fits to a specific category. The model learns to modify its parameters during training so that it generates output values that are nearly equal to 0 for non-threat class and 1 for threat class. The sigmoid output is subjected to a decision threshold (usually 0.5) during inference in order to classify the input. The model classifies the output as a threat if it exceeds 0.5 and as a non- threat otherwise.

In order to ensure uniformity in the input sequence, the suggested model starts with a DistilBERT tokenizer, which tokenizes and pads input tweets to a constant length of 100 tokens. A custom DistilBERT layer receives the tokenized outputs, which comprise input_ids and attention_mask. This layer captures the semantic complexities of the tweets and uses the pre- trained DistilBERT model to extract contextual embeddings. To reflect the deep contextual relationships in the text, the DistilBERT model creates a series of embeddings for every token in the input. The resultant embeddings are then subjected to a global average pooling layer. This layer aggregates the token- level embeddings into a fixed- size vector, simplifying the representation while keeping the essential elements. A dropout layer with 20% rate is applied to the pooled output to further reduce overfitting by randomly turning off some neurons during training. A deep layer with 128 neurons and ReLU activation enables the network to subsequently discover complex patterns in the data. Regularization is boosted by adding a second dropout layer at the same rate. The output layer, comprising one neuron with a sigmoid activation function, finishes the proposed system by providing a binary classification for deciding whether the tweet contains information relevant to terrorism. Fig. 10 shows the proposed model architecture.

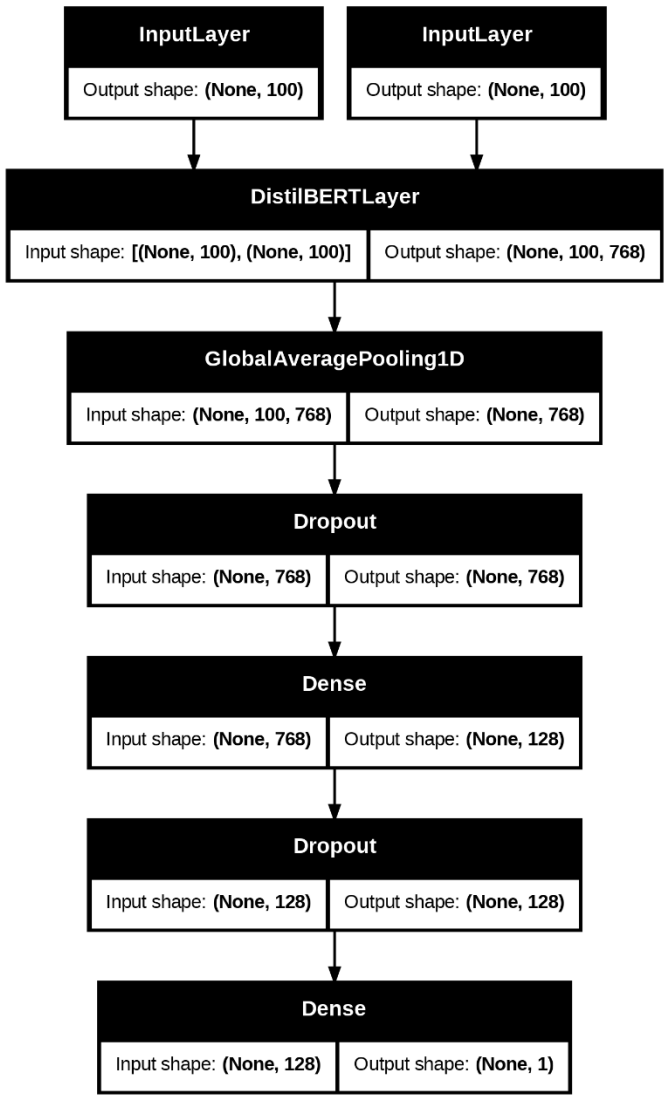


Fig. 10. Proposed terrorism attack detection model architecture.

The detailed algorithm of the proposed terrorism attack detection model is explained below.

Algorithm: A Novel NLP Model Based on DistilBERT and DNN for the Detection of Terrorism Threat from Tweets

Input: *Text data related to terrorism attacks, Pre- trained DistilBERT Model*
Output: *Efficient Terrorism Attack Detection Model*

Begin:

- *Collect Dataset: $\mathcal{D}$= {(X, y), where X is a tweet and y∈ {0,1}, (0 for non- threat tweet and 1 for threat tweet)*
- *Text Preprocessing:*
  - *URLs Removal*
  - *Mentions Removal*
  - *Hashtags Removal*
  - *Numbers Removal*
  - *Punctuations Removal*
  - *Convert to lowercase*
  - *Stopwords Removal*
- *Dataset Splitting:*
  - *X = df['Cleaned_Tweet']*
  - *y = df['Class']*
  - *X_train, X_test, y_train, y_test = train_test_split (X, y, test_size=0.2, random_state=42)*
- *Tokenization using DistilBERT Tokenizer:*
  - *Tokenized_Input=tokenizer (X, padding='max_length', truncation=True, max_length=100)*
  - *Tokenizer returns:*
    - *input_ids ∈ $Z^{n\times 100}$: Tokenized IDs for each word.*
    - *attention_mask ∈ $\{0,1\}^{n\times 100}$: Masks indicating relevant tokens.*
- *Proposed DistilBERT- DNN Model Creation:*
  - *distilbert_model = TFDistilBertModel.from_pretrained ('distilbert-base uncased')*
  - *input_ids = Input (shape= (100,), dtype=tf.int32, name='input_ids')*
  - *attention_mask = Input (shape= (100,), dtype=tf.int32, name='attention_mask')*
  - *bert_output = DistilBERTLayer () ([input_ids, attention_mask])*
  - *x = GlobalAveragePooling1D () (bert_output)*
  - *x = Dropout (0.2) (x)*
  - *x = Dense (128, activation='relu') (x)*
  - *x = Dropout (0.2) (x)*
  - *output = Dense (1, activation='sigmoid') (x)*
  - *M = Model (inputs= [input_ids, attention_mask], outputs=output)*
- *Model Compilation and Training:*
  - *M.compile(optimizer=Adam(learning_rate=0.0001), loss='binary_crossentropy', metrics=['accuracy'])*
  - *M.fit([X_train_tokens['input_ids'], X_train_tokens['attention_mask']], y_train, epochs=100, batch_size=32, validation_split=0.2)*
- *Model Evaluation:*
  - *y_pred_prob=M.predict([X_test_tokens['input_ids'], X_test_tokens['attention_mask']])*
  - *y_pred = (y_pred_prob > 0.5). astype(int). flatten ()*
  - *accuracy = accuracy_score (y_test, y_pred)*
  - *report = classification_report (y_test, y_pred)*
  - *conf_matrix = confusion_matrix (y_test, y_pred)*
  - *roc_auc = roc_auc_score (y_test, y_pred_prob)*
- *Save the Model:*

End

E. Simulation Setup

The proposed terrorism attack detection model has been designed and implemented on Google Collaboratory platform with Python language. Gradient estimation stability and computing performance were balanced by using a batch size of 32. The Adam optimizer was selected to speed up convergence because of its popularity for adaptable learning. The model underwent training for 100 epochs, providing enough iterations to discover intricate patterns in the data. The learning rate was chosen to be 0.001 to allow a fine-grained change to model weights during training, limiting overshooting and ensuring stable convergence. The loss function, binary cross entropy was appropriate for binary classification tasks. Table I offers various hyperparameters utilized by the model.

TABLE I. HYPERPARAMETERS

Hyperparameters	Values
Batch Size	32
Optimizer	Adam
Number of Epochs	100
Loss Function	Binary Crossentropy
Learning Rate	0.0001

A batch size of 32 was selected in order to balance model generalization with computational efficiency. This implies that the model processes 32 datasets at a time before updating the weights. The model successfully improved its parameters repeatedly throughout several dataset iterations during the 100

training epochs. This binary classification problem is well-fitted by the binary cross-entropy loss function, which counts the discrepancy between expected probability and actual binary class labels. The Adam optimizer was utilized to optimize the weights, combining the advantages of adaptive learning rates and momentum for effective training. The step size during weight updates was controlled by setting the learning rate to 0.0001, which allowed for consistent convergence.

IV. Experimental Results

The performance of a model during training and validation can be assessed using the accuracy plot and loss plot. The accuracy plot provides insight into the effectiveness of the model in generalizing to unknown data by illustrating changes in its capacity to accurately categorize samples over different epochs. A steadily increasing accuracy curve demonstrates effective learning. The loss plot on the other hand, displays the model's error over epochs and a decreasing loss indicates a reduction in the difference between predicted and assigned labels. When combined, these plots offer a thorough understanding of how the model learns and provide optimization techniques for enhanced efficiency. Fig. 11(a) and Fig. 11(b) visualizes the accuracy plot and loss plot of the model. The model performs much better in the first epochs with training accuracy increasing from 77% in epoch 1 to over 93% by epoch 50 and training loss decreasing in parallel. After epoch 30, validation accuracy steadily increases between 91 and 92%, while validation loss levels off at 0.22, suggesting efficient learning and little overfitting. The model may have achieved an ideal balance if both training and validation measures exhibit convergence.

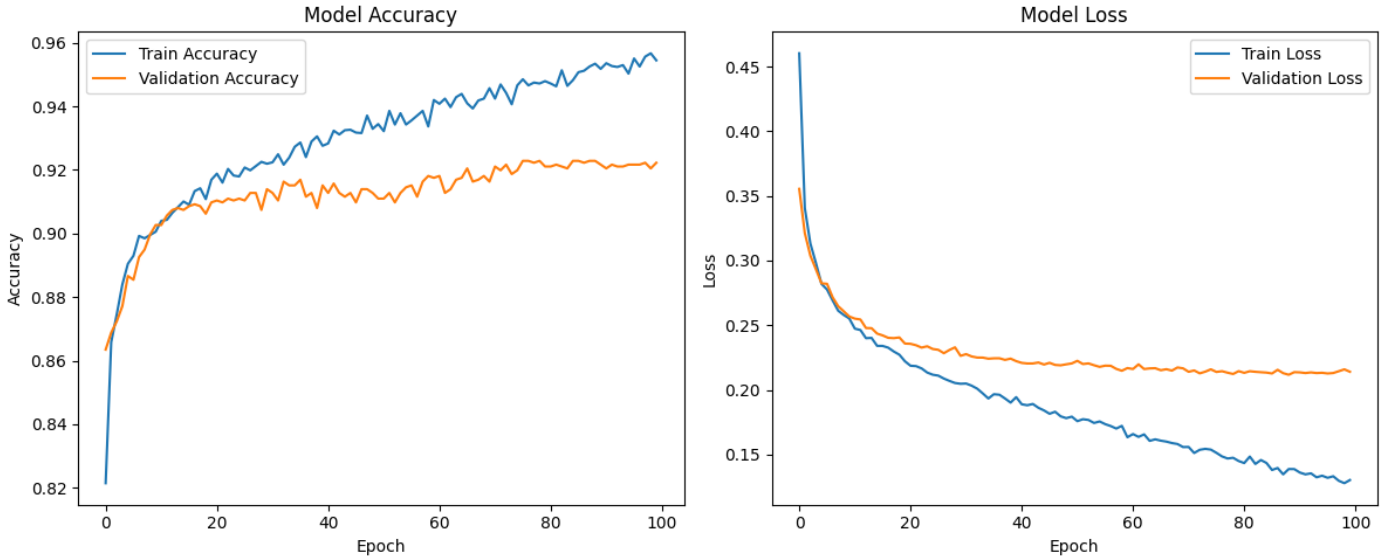


Fig. 11. (a) Accuracy plot (b) Loss plot of proposed terrorism attack detection model.

Accuracy is a crucial performance static in the context of terrorism attack detection from twitter data. It quantifies the percentage of correctly recognized occurrences among all occurrences. It is computed using Eq. (3).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

The term true positive (TP) refers to the quantity of threat-related tweets correctly identified. True negativity (TN) refers to the quantity of accurately identified, non-threatening tweets. False positives (FP) refer to the number of non-threat-related tweets that receive incorrect classification as threats. The number of tweets about threat that are inaccurately classified as non- threat is known as false negative (FN). The proposed model for terrorism attack detection accomplished a remarkable performance with an accuracy of 93%.

A classification report is a thorough performance assessment tool that offers important indicators for evaluating the effectiveness of a detection model. It comprises support for every class, F1- score, precision and recall. The F1 score offers a balanced measure of precision and recall, with recall demonstrating the model's ability to identify all pertinent events and precision quantifying the accuracy of positive predictions. The number of instances of each class in the dataset is known as support. Table II presents the classification report for the suggested model.

TABLE II. Classification Report of Suggested Model

Class	Precision	Recall	F1- Score	Support
0 (Non- Threat)	0.93	0.93	0.93	1074
1 (Threat)	0.92	0.92	0.92	1032
Accuracy			0.93	2106
Macro Average	0.93	0.93	0.93	2106
Weighted Average	0.93	0.93	0.93	2106

The classification report shows that the terrorism attack detection model performs well in both the threat and non-threat classes. For non- threat class, the model's precision, recall and F1- score of 0.93 show that it is quite accurate in detecting non- threats without producing a lot of false positives. The precision, recall and F1- score for the threat class are 0.92, but they still show a high level of threat

detection accuracy. The classification between the two classes is well balanced with an overall model accuracy of 93%. The macro average, which calculates the measures without taking into account class imbalance shows consistent performance across classes, getting a strong 0.93 for precision, recall and F1- score. Additionally, the weighted average supports the efficacy of the model by confirming constant performance while accounting for the number of cases in each class.

An essential tool for assessing the effectiveness of detection models is the confusion matrix, which offers a thorough analysis of the prediction model. The confusion matrix aids in locating class imbalances and directs model enhancements, such as threshold adjustments, regularization or class distribution balancing. Fig. 12 visualizes the confusion matrix of the suggested model.

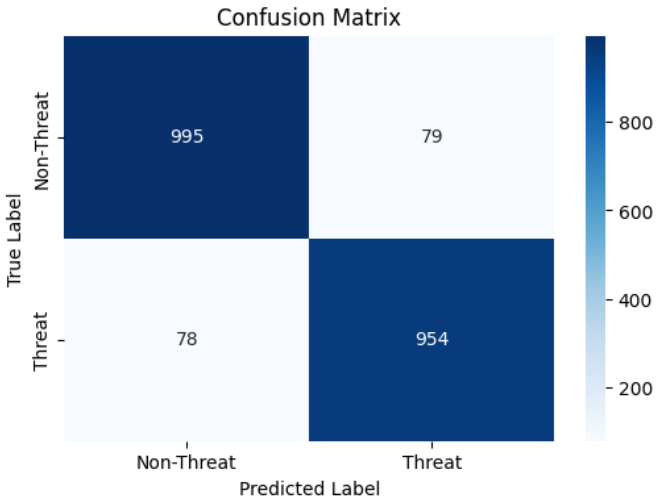


Fig. 12. Confusion matrix of suggested system.

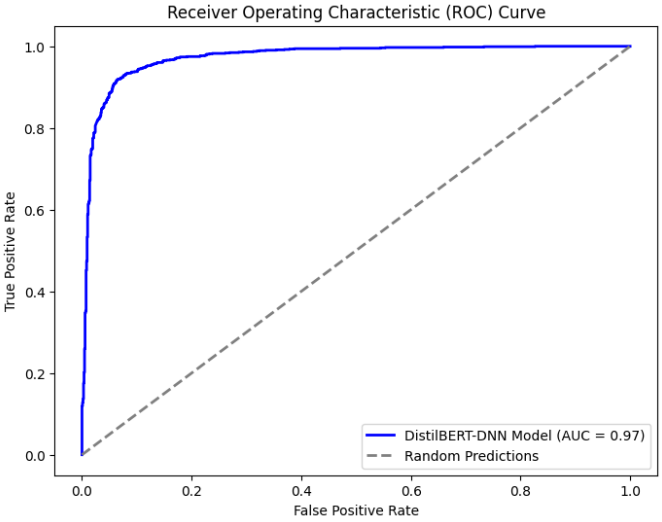


Fig. 13. ROC Curve of proposed model.

The confusion matrix illustrates that the model has a high degree of accuracy in both classifications, properly classifying 995 non- threat cases as non- threats and 954 threat cases as threats. However, there are 78 false negatives where real threats were incorrectly classified as non- threats and 79 false positives where non- threats were incorrectly classified as threats. The effectiveness of detection models is evaluated using a performance evaluation tool called the Receiver Operating Characteristic (ROC) curve, which is shown in Fig. 13. The rates of false positives and true positives are shown at different thresholds.

The ROC curve illustrates the balance between minimizing false positives and identifying true positives. The efficacy of a model is frequently evaluated by the area under the ROC curve, known as AUC. A greater AUC denotes better model performance. Some of the prediction outputs are displayed in Fig. 14.

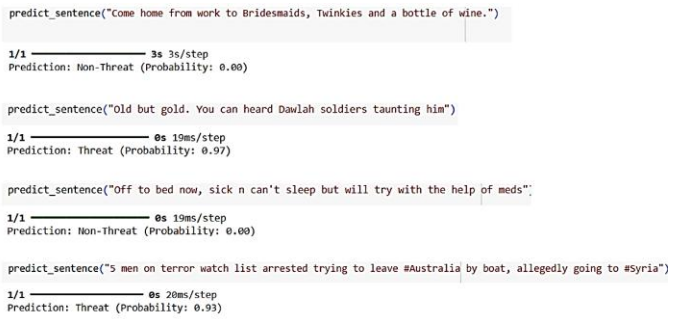


Fig. 14. Prediction outputs.

V. DISCUSSION

Table III provides the performance comparison of the suggested terrorist threat detection system with existing approaches. The suggested DistilBERT- DNN model substantially outperformed a number of current approaches in the field of terrorist threat identification obtaining an incredible 93% accuracy rate. The suggested model shows the benefits of advanced NLP techniques with a significant improvement over the NB classifier, maximum entropy classifier and SVM, which all reached an accuracy of 73.33%. This highlights the effectiveness of merging lightweight transformer-based embeddings with a DNN.

The DistilBERT- DNN model also outperformed the SVM based technique, which obtained 84% and ensemble methods, including RF, KNN, DT, SVM and Adaboost, which achieved 90%. Additionally, the model outperformed Word2Vec in gradient boosting with Word2Vec (87.9%) and K- means clustering (85.8%). Additionally, it outperforms conventional and neural based methods like ANN at 75 % and BERT and its variants at 72%. This demonstrates how well the DistilBERT-DNN model captures the semantic context of tweets for accurate threat identification, making it an innovative approach in this field. Fig. 15 shows the performance comparison of the suggested model with state-of-art methodologies.

TABLE III. PERFORMANCE COMPARISON

Authors	Proposed Methodology	Detection Accuracy
Mayur Gaikwad et al. [11]	BERT and its Variants	72%
Kuljeet Kaur [26]	NB Classifier, Maximum Entropy Classifier and SVM	73.33%
Ghada M.A. Soliman and Tarek H.M. Abou-El-Enien [17]	Artificial Neural Network	75%
Waqas Sharif et al. [27]	SVM	84 %
Lu An et al. [9]	Word2Vec with the K-means clustering	85.8%
Shynar Mussiraliyeva et al. [16]	Gradient boosting with word2vec	89%
Sagar Shinde et al. [8]	RF, KNN, DT, SVM, Adaboost	90%
Proposed DistilBERT- DNN Model		**93%**

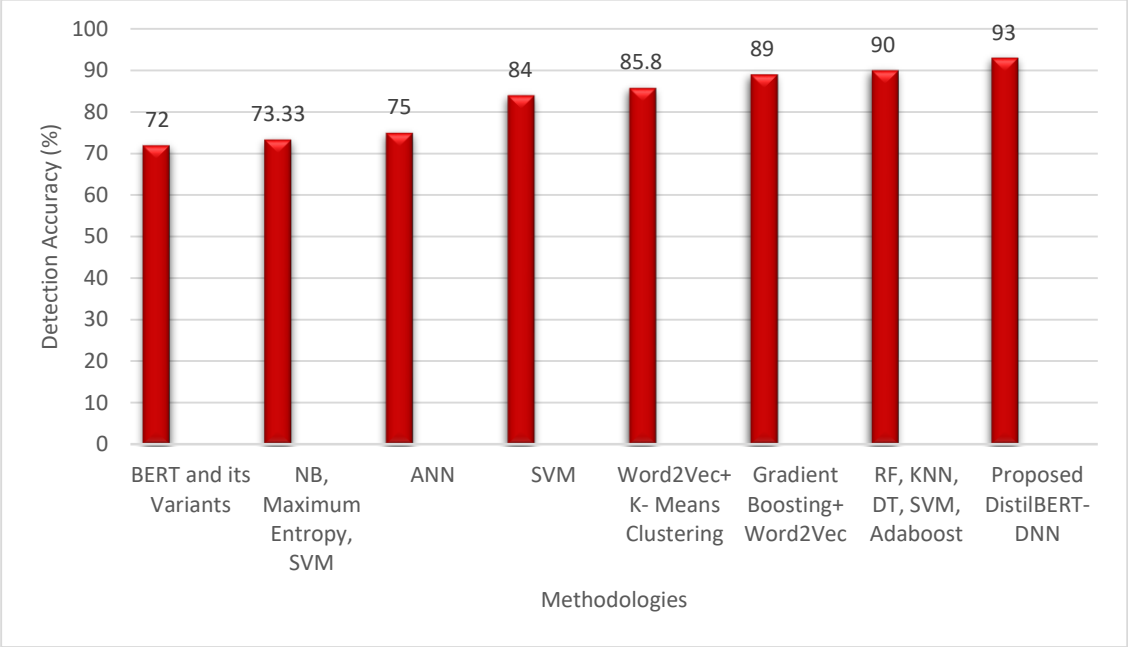


Fig. 15. Graphical illustration of performance comparison with existing methodologies.

VI. CONCLUSION

Global terrorist activity has increased significantly in recent years, causing major risks to national security and public safety. Conventional methods of detecting and preventing terrorist attacks often rely on human contact and existing surveillance systems, which limits their monitoring capabilities and response measures. The intricacy, speed and extent of modern terrorism renders these conventional techniques insufficient, as it often involves decentralized operations, digital traces and secret conversations. The development of increasingly advanced automated systems that can identify possible terrorist actions with greater accuracy and speed is required due to the technological advancements, particularly the growth of digital media and communication platforms. This paper proposed an effective terrorism threat detection model using DistilBERT with Deep Neural Network from twitter data. The model ensures accurate and efficient representation of text data by utilizing DistilBERT's transformer-based architecture, which is lightweight and powerful to extract rich semantic information from tweets. A Dense Neural Network, which is excellent at spotting intricate patterns and relationship in the data, processes these embeddings further to accurately classify tweets as either threats or non- threat. The DistilBERT-DNN model has demonstrated remarkable efficacy in detecting terrorist threats from twitter data with a 93% accuracy. The methodology surpasses existing neural network-based approaches and conventional machine learning techniques by effectively combining the advantages of deep learning and natural language processing. This demonstrates how advanced NLP techniques can improve the effectiveness and reliability of automated systems for detecting terrorist threats. Future research can focus on integrating multimodal data sources, such as image and video analysis, to improve threat detection beyond textual data. Additionally, incorporating real-time streaming capabilities can enable proactive identification and mitigation of threats as they emerge. Moreover, leveraging graph-based neural networks can help analyze network structures and relationships between suspicious accounts to identify coordinated terrorist activities. These future enhancements will contribute to the development of a more comprehensive, accurate, and real-time terrorism threat detection system, strengthening national security and public safety.

REFERENCES

[1] Schmid, A. P. (2011). The definition of terrorism. In The Routledge handbook of terrorism research (pp. 39-157). Routledge.

[2] Nasution, A. R. (2017, December). Acts of terrorism as a crime against humanity in the aspect of law and human rights. In 2nd International Conference on Social and Political Development (ICOSOP 2017) (pp. 346-353). Atlantis Press.

[3] Wilkinson, P. (2006). Terrorism versus democracy: The liberal state response. Routledge.

[4] Huang, R. (2008). Counterterrorism and the Rule of Law. Civil War and the Rule of Law: Security, Development, Human Rights, edited by A. Hurwitz and R. Huang, 261-284.

[5] Turk, A. T. (2004). Sociology of terrorism. Annu. Rev. Sociol., 30(1), 271-286.

[6] Kelleher, J. D. (2019). Deep learning. MIT press.

[7] Chowdhary, K., & Chowdhary, K. R. (2020). Natural language processing. Fundamentals of artificial intelligence, 603-649.

[8] Shinde, S., Khoje, S., Raj, A., Wadhwa, L., & Shaikha, A. S. (2024). Artificial intelligence approach for terror attacks prediction through machine learning. Multidisciplinary Science Journal, 6(1), 2024011-2024011.

[9] An, L., Han, Y., Yi, X., Li, G., & Yu, C. (2023). Prediction and evolution of the influence of microblog entries in the context of terrorist events. Social Science Computer Review, 41(1), 64-82.

[10] Jyothilinga, S. (2023). Analysis and Prediction of Terrorist Attacks using Supervised Machine Learning and Deep Learning Techniques (Doctoral dissertation, Dublin, National College of Ireland).

[11] Gaikwad, M., Ahirrao, S., Kotecha, K., & Abraham, A. (2022). Multi-ideology multi-class extremism classification using deep learning techniques. IEEE Access, 10, 104829-104843.

[12] Feng, Y., Wang, D., Yin, Y., Li, Z., & Hu, Z. (2020). An XGBoost-based casualty prediction method for terrorist attacks. Complex & Intelligent Systems, 6, 721-740.

[13] Zhao, R., Xie, X., Zhang, X., Jin, M., & Hao, M. (2021). Spatial distribution assessment of terrorist attack types based on I-MLKNN model. ISPRS International Journal of Geo-Information, 10(8), 547.

[14] Luo, L., & Qi, C. (2021). An analysis of the crucial indicators impacting the risk of terrorist attacks: A predictive perspective. Safety science, 144, 105442.

[15] Theodosiadou, O., Pantelidou, K., Bastas, N., Chatzakou, D., Tsikrika, T., Vrochidis, S., & Kompatsiaris, I. (2021). Change point detection in terrorism-related online content using deep learning derived indicators. Information, 12(7), 274.

[16] Mussiraliyeva, S., Bolatbek, M., Omarov, B., & Bagitova, K. (2020). Detection of extremist ideation on social media using machine learning techniques. In Computational Collective Intelligence: 12th International Conference, ICCCI 2020, Da Nang, Vietnam, November 30–December 3, 2020, Proceedings 12 (pp. 743-752). Springer International Publishing.

[17] Soliman, G. M., & Abou-El-Enien, T. H. (2019). Terrorism Prediction Using Artificial Neural Network. Rev. d'Intelligence Artif., 33(2), 81-87.

[18] Koutidis, G., Loumponias, K., Tsikrika, T., Vrochidis, S., & Kompatsiaris, I. (2024). Towards AI-driven Prediction of Terrorism Risk based on the Analysis of Localised Web News. IEEE Access.

[19] Chaabene, N. E. H. B., Bouzeghoub, A., Guetari, R., & Ghezala, H. H. B. (2021, October). Applying machine learning models for detecting and predicting militant terrorists behaviour in Twitter. In 2021 IEEE international conference on systems, man, and cybernetics (SMC) (pp. 309-314). IEEE.

[20] https://github.com/juanbetancur96/TerrorismDetectionTweeter/blob/main/Dataset.xlsx

[21] A. Mullen, L., Benoit, K., Keyes, O., Selivanov, D., & Arnold, J. (2018). Fast, consistent tokenization of natural language text. Journal of Open-Source Software, 3(23), 655.

[22] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. ArXiv preprint arXiv:1503.02531.

[23] Sanh, V. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[24] Ghojogh, B., & Ghodsi, A. (2020). Attention mechanism, transformers, BERT, and GPT: tutorial and survey.

[25] Kriegeskorte, N., & Golan, T. (2019). Neural network models and deep learning. Current Biology, 29(7), R231-R236.

[26] Kaur, K. (2016, March). Development of a framework for analyzing terrorism actions via Twitter lists. In 2016 International conference on computational techniques in information and communication technologies (ICCTICT) (pp. 19-24). IEEE.

[27] Sharif, W., Mumtaz, S., Shafiq, Z., Riaz, O., Ali, T., Husnain, M., & Choi, G. S. (2019). An empirical approach for extreme behavior identification through tweets using machine learning. Applied Sciences, 9(18), 3723.

Utilizing NLP to Optimize Municipal Services Delivery Using a Novel Municipal Arabic Dataset

Homod Hamed Alaloye, Ahmad B. Alkhodre, Emad Nabil
Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah, Saudi Arabia

***Abstract*—The natural language processing paradigm has emerged as a vital tool for addressing complex business challenges, mainly due to advancements in machine learning (ML), deep learning (DL), and Generative AI. Advanced NLP models have significantly enhanced the efficiency and effectiveness of NLP applications, enabling the seamless integration of various business processes to improve decision-making. In the municipal sector, the Kingdom of Saudi Arabia is trying to harness the power of NLP to promote urban development, city planning, and infrastructure enhancements, ultimately elevating the quality of life for its residents. In the municipal sector, approximately 300 services are available through multiple channels, including the Baladi application, unified communication services, WhatsApp, a dedicated beneficiary center (serving citizens and residents), and social media accounts. These channels are supported by a dedicated team that operates 24/7. This paper examines the implementation of ML and DL methods to categorize requests and suggestions submitted by residents for various municipal services in the Kingdom of Saudi Arabia. The primary aim of this work is to enhance service quality and reduce response times to community inquiries. However, a significant challenge arises from the lack of Arabic datasets specifically tailored to the municipal sector for training purposes, which limits meaningful progress. To address this issue, we have created a novel dataset consisting of 3,714 manually classified requests and suggestions collected from the X platform. This dataset is organized into eight classes: tree maintenance, lighting, construction waste, old and neglected assets, road conditions, visual pollution, billboards, and cleanliness. Our findings indicate that ML models, particularly when optimized with hyperparameters and appropriate pre-processing, outperformed DL models, achieving an F1 score of 90% compared to 88%. By releasing this novel Arabic dataset, which will be open sourced for the scientific community, we believe this work provides a foundational reference for further research and significantly contributes to improving the municipal sector's service delivery.**

Keywords*—Arabic text classification; machine learning (ML); deep learning (DL); hyperparameter optimization; municipal services

I. Introduction

Machine learning (ML) and deep learning (DL) models have significantly advanced our understanding of natural language, creating new research opportunities in areas such as text classification, translation, summarization, generation, and question-answering. However, the effectiveness of these models can vary depending on the specific task, particularly in text classification.

The authors in [1] have applied ensemble learning techniques to improve accuracy, which are especially effective for dealing with complex data with varied features. Ensemble learning combines multiple models with diverse parameters, enhancing the analysis and classification of intricate data patterns. As a result, these techniques prove beneficial for achieving higher accuracy in complex text classification tasks.

The authors in [2] utilize neural network architectures with multiple layers to effectively identify and interpret patterns in language-related tasks, such as question answering, translation, and text classification. This effectiveness is further enhanced by embedding techniques, including word and character embeddings, which are essential for tasks like text classification, knowledge extraction, and question-answering. Additionally, hyperparameter optimization (HPO) is crucial in improving model performance, as systematic hyperparameter tuning can significantly enhance accuracy and efficiency.

The authors in [3] describe a strategy for selecting hyperparameter values incorporating techniques such as Grid Search, Random Search, Bayesian Optimization, Particle Swarm Optimization, and Genetic Algorithm Metaheuristics.

ML and DL models have transformed how business applications and services are enhanced and developed; many applications and services currently use machine learning (ML) and deep learning (DL) models for various purposes, including task automation, improved decision-making, enhanced customer experience, predictions, and automatic text classification. In Saudi Arabia, the municipal sector provides over 300 essential services for improving the quality of life through urban development, city planning, and infrastructure improvement. These services cover streets, parks, lighting, green spaces, sidewalks, public health, sanitation, and public transportation management. They are delivered through various channels, including the Balady application, unified contact numbers, WhatsApp service, beneficiary centers, and the municipal sector's social media accounts, all supported by a dedicated team working around the clock.

Using machine learning (ML) and deep learning (DL) models in municipal service procedures can improve how citizen and resident requests and suggestions are classified. This enhancement leads to better service performance, lower financial and human costs, and increased overall satisfaction with municipal services.

These models require a robust dataset for training, which is crucial for effectively extracting knowledge, analyzing and classifying data, recognizing patterns, and turning raw data into actionable information. However, our review of existing literature reveals a significant gap in the availability of training datasets, especially in Arabic. Most datasets for training learning

models are available in English, while there is a shortage of sufficient datasets in Arabic.

This paper significantly contributes to natural language processing, particularly concerning municipal services in the Kingdom of Saudi Arabia. It introduces a comprehensive, specialized Arabic dataset that captures citizens' requests and suggestions related to these services. This dataset enhances the resources available for researchers and practitioners in classifying and analyzing Arabic texts. Furthermore, it is a valuable tool for advancing the development of machine learning (ML) and deep learning (DL) models tailored to the Arabic language's unique linguistic and cultural characteristics. Ultimately, this will facilitate more effective data-driven decision-making in municipal service delivery. The dataset will be open-sourced for the community, and we will also provide a performance baseline, allowing others to use it for performance comparison in their own work.

The paper is organized into six sections. Section II provides a comprehensive literature review, offering insights into previous research. Section III details the paper's key contributions, elaborating on the proposed methodology. Section IV presents the experimental results, followed by a discussion in Section V. Finally, Section VI concludes the paper by summarizing the key insights and future directions.

II. Literature Review

The text classification problem has many real-life applications, and there are many techniques for tackling it. Fig. 1 depicts the major methods.

With the global rise of social media platforms like Twitter, Facebook, and Instagram, coupled with greater engagement with hashtags, there has been a significant increase in Arabic-language tweets that contain racist, bullying, or hateful content. This trend has prompted researchers to classify Arabic text into multi-class and binary categories to tackle these issues.

In a related study [4], the authors employed binary classification to categorize millions of articles and blogs as either real or fake news using the Multivariate Bernoulli Naïve Bayes model, demonstrating the effectiveness of ML techniques in authenticating Arabic content. These studies highlight the wide-ranging applicability of ML methods in Arabic-language tasks, from hate speech detection to verifying the authenticity of information.

In [5], the authors contributed to reducing the time researchers spend searching for relevant papers by employing classification models such as SVM, Naïve Bayes, K-nearest Neighbor, and Decision Tree to categorize a set of 107 research papers across various fields, including science, business, and social sciences.

In [6], the authors examined the impact of various stemming methods and word representations on Arabic text classification using RNN architectures, specifically LSTM, Bi-LSTM, GRU, and Bi-GRU models. They compared the accuracy among different combinations of stemming approaches (no stemming, root-based stemming, and light-based stemming) and word representations (Word2Vec, FastText, and GloVe). The highest accuracy was found with no stemming using Word2Vec for word representation, along with the Bi-GRU architecture as the classifier.

Similarly, the authors in [7] developed a dataset for dialectal Arabic tweet acts by annotating a subset of the existing Arabic Sentiment Analysis Dataset (ASAD) based on six speech act categories. The ArSAS dataset was classified using BERT, and the results were compared with various Arabic BERT models. The findings suggest that the top-performing model was araBERTv2-Twitter.

The authors in [8] classified numerous policies published by governments and institutions, each containing several paragraphs indicating a specific topic. The study relied on ensemble learning techniques, particularly Bagging, integrated with DL methods, specifically Convolutional Neural Networks (CNN), to classify the policies based on the predetermined topics of the paragraphs.

In a related study [9], the authors demonstrated that ensemble learning outperformed traditional ML approaches by 6% in classifying Arabic data, which is characterized by its complexity arising from multiple dialects and variations in letter formation.

In [10], the authors employed four deep learning models—Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Convolutional LSTM (CNN-LSTM), and Bidirectional LSTM (Bi-LSTM)—to classify millions of Arabic texts across diverse domains, including cultural, scientific, political, and health-related topics, sourced from websites, social media platforms, academic publications, and news outlets. The Bi-LSTM model achieved the highest accuracy result (94.02).

The authors in [11] applied multi-label classification models to categorize the topics of each verse in the Quran, enhancing understanding of its content. They developed a majority voting approach, achieving optimal performance through a combination of Random Forest, Support Vector Machine (SVM), and Adaptive Boosting models. This ensemble achieved a micro F1 score of 0.5143 and a Hamming loss of 0.0878. Additionally, the majority voting method demonstrated improved performance compared to the individual models.

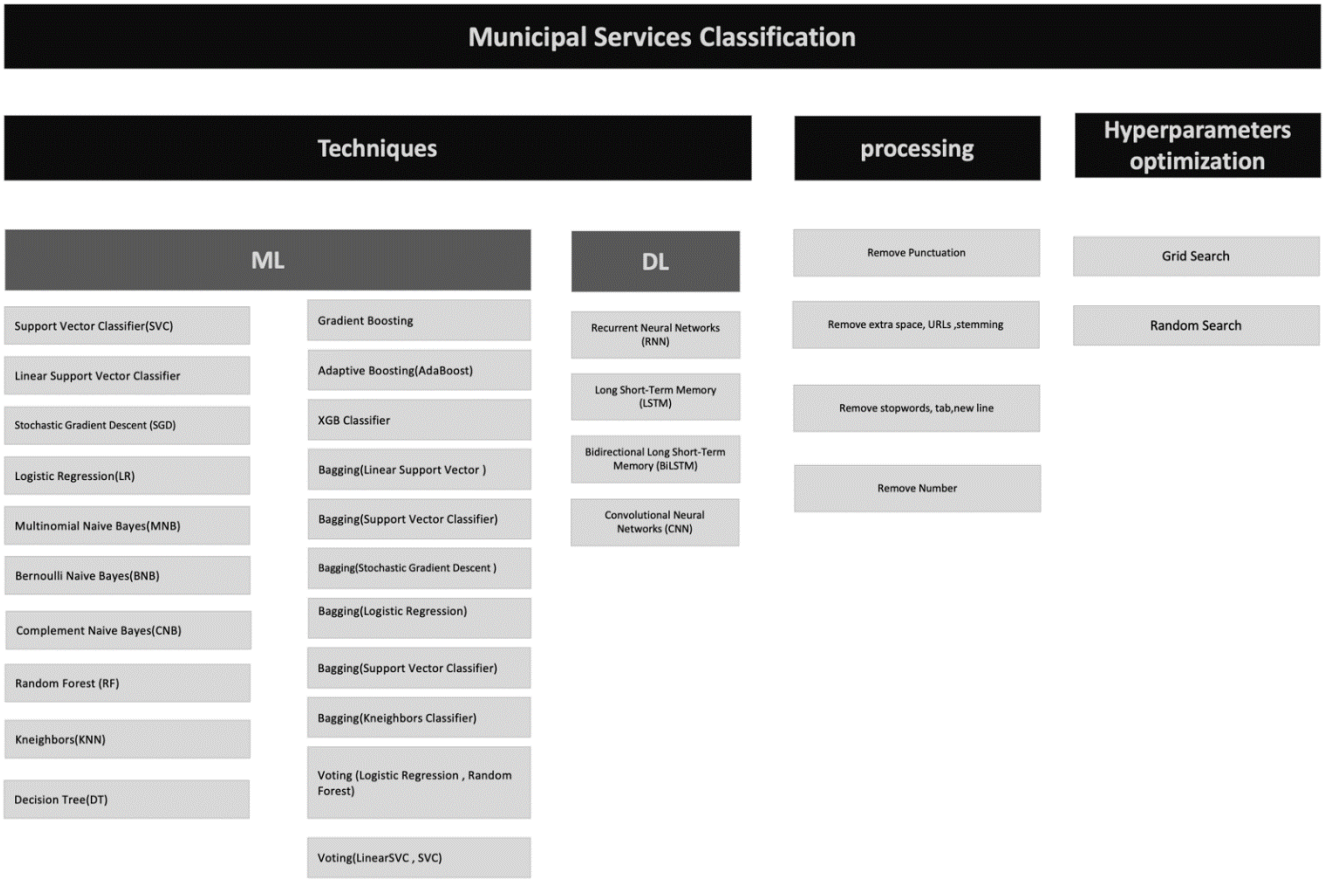


Fig. 1. NLP Classification: Techniques, preprocessing steps, and hyperparameter optimization.

The authors in [12] focused on classifying SMS messages as either spam or not spam, emphasizing the importance of this medium for communication and verification by banks, institutions, and e-platforms. They employed ensemble learning techniques for message classification, achieving an impressive accuracy of 99.91%.

The authors in [13] examined the classification of a large dataset containing 290,000 multi-tagged articles across various categories, including business, technology, sports, and politics. This research aimed to analyze the dataset to identify patterns, which would assist in summarizing and classifying the articles. Classifying news stories that belong to multiple categories posed significant challenges. However, logistic regression achieved an accuracy of 81.3%, while XGBoost reached 84.7%. Notably, a combination of ten neural network architectures—namely CNN, CLSTM, LSTM, BILSTM, GRU, CGRU, BIGRU, HANGRU, CRF-BILSTM, and HANLSTM—performed exceptionally well as a multilabel classifier, achieving an accuracy of 94.85%.

In a related study [14], the authors experimentally investigated the application of Convolutional Neural Networks (CNNs) to text classification; using FastText for word embeddings, the CNN model was trained on publicly different datasets. The authors in [15] compared the effects of classical and contextualized word embeddings on model performance. They evaluated classical embeddings, such as GloVe, Word2Vec, and FastText, on the HARD, Khooli, AJGT, ArSAS, and ASTD datasets using BiLSTM and CNN models. However, the contextualized transformer-based embedding model BERT outperformed the classical embeddings, achieving the highest classification accuracy across all datasets.

The authors in [16] concentrated on classifying fake Arabic news on Twitter by using various Arabic contextualized embedding models, such as AraBert, Arabic-BERT, ArBert, MARBert, Araelectra, and QaribBert, which together achieved an impressive classification accuracy of 98%.

The authors in [17] implemented various models, including Convolutional-GRU and Attention-GRU, using Word2Vec embeddings for Arabic text classification. They compared model performance on the SANAD and NADiA datasets, achieving an accuracy of 91.18% on SANAD and 96.94% on NADiA.

The authors in [18] applied a Genetic Algorithm to optimize CNN model parameters, improving Arabic text classification accuracy to 98%.

The authors in [19] utilized Word2Vec and TF-IDF with models such as SVC, Random Forest, and Gaussian Naive Bayes to classify Arabic tweets, reporting high model accuracy rates of 98.09% and 98.14%.

The authors in [20] demonstrated that the performance of the large language model BERT, when applied without fine-tuning its parameters, surpassed that of models utilizing fine-tuned parameters.

The authors in [21] used the NLTK package to improve machine learning performance in classifying Arabic news (DAFN) as either fake or real.

The authors in [22] collected a dataset using the Twitter API, which they manually labeled to ensure its reliability. They employed two techniques for feature extraction: N-gram models (including unigrams, bigrams, and char-grams) to enhance the classification of advertisements and illegal content through both machine learning (ML) and deep learning (DL) approaches.

The authors in [23] utilized a meta-heuristic genetic algorithm (G.A.) to optimize hyperparameters and feature extraction techniques, specifically the bag of words (BOW) and term frequency-inverse document frequency (TF-IDF) methods. They assessed the performance of several ensemble learning models, including the Extra Trees Classifier (ETC), Random Forest Classifier (RFC), and Logistic Regression Classifier (LRC). Their efforts led to an F1 score of 80.88% for sentiment classification and 68.76% for multilabel classification tasks.

The authors in [24] achieved a classification accuracy of 94.75% by grouping similar Arabic words generated by the Word2Vec model. Each group of similar words utilized the Word2Vec-based method RARF, which enhanced the model's effectiveness in accurately classifying Arabic text.

The authors in [25] utilized the ARABERT4TWC model, which features a custom-built Arabic BERT, transformation layers, a self-attention layer, and a classification layer to improve the classification of Arabic tweets. Using datasets such as SANAD for Arabic news stories, HARD, AJGT, ASTD, and ArsenTD–Lev, the model achieved an impressive accuracy of 98%.

The authors in [26] utilized ensemble learning alongside N-gram features (Unigram, Bi-Gram, and Tri-Gram) to improve the classification accuracy of Arabic spam reviews, achieving maximum accuracies of 95% and 99.98% across two datasets.

In a related study [27], the authors emphasized the importance of hyperparameter tuning techniques in improving the performance of machine learning (ML) and deep learning (DL) algorithms applied to Arabic text. They utilized random search and grid search methods for tuning in classification tasks sourced from cnnarabic.com. The findings indicated that random search outperformed grid search, achieving accuracies of 95.16% for Artificial Neural Networks (ANN), 94.57% for Multinomial Logistic Regression (MLR), and 94.28% for Support Vector Machines (SVM).

The authors in [28] used the Particle Swarm Optimization (PSO) algorithm for feature selection, resulting in a document classification accuracy of 97%.

The authors in [29] compared various Arabic BERT models used for text classification to assess and summarize each model's performance.

In [30], a related study, the authors introduced the BERT-Mini Model (ABMM) to classify hate speech in Arabic social media posts into standard, abusive, and hate speech. The ABMM model achieved an impressive accuracy score of 0.986.

The authors in [31] collected opinions from Arabic-speaking users about ChatGPT technology using the Tweepy and SNscrape Python libraries. They categorized the data into three sentiment classes: negative, positive, and neutral. By employing classification models such as RoBERTa, they achieved a recall accuracy of 99%.

The authors in [32] performed a comparative analysis of ChatGPT-4 and Google Gemini for detecting and classifying messages as spam, showcasing advancements in AI language models for content moderation.

The authors in [33] presented a method for improving the classification of cognitive distortions in Arabic content on Twitter. They incorporated an unlabeled dataset using the unsupervised learning model BERTopic. This approach combined the processed unlabeled data with a labeled dataset, which enhanced the overall classification performance using machine learning techniques. A concise summary of the literature is presented in Table I.

III. METHODOLOGY

The proposed methodology, illustrated in Fig. 2, presents a multi-method approach aimed at enhancing the automatic classification of citizens' and residents' requests and suggestions regarding municipal services in the Kingdom of Saudi Arabia. This approach incorporates various machine learning (ML) and deep learning (DL) models, along with hyperparameter tuning and data processing techniques, to improve classification accuracy. It includes a comparison of model performance before and after applying these hyperparameters and data preprocessing, with the objective of identifying the best-performing model for this task. The ultimate goal is to streamline the classification of public requests and suggestions related to municipal services, thus facilitating more efficient service delivery and quicker response times.

TABLE I. REVIEW OF THE PRIMARY STUDIES IN THE LITERATURE

Index	Reference	Main Idea	Used Techniques	Dataset
1	Alzanin et al.	Multilabel classification of Arabic content in the Arabic language.	Ensemble learning models (Extra Trees Classifier, Random Forest Classifier, Logistic Regression Classifier), genetic algorithm (GA), Bag of Words (BOW), and Term Frequency-Inverse Document Frequency (TF-IDF).	MAWQIF
2	Sabri et al.	Arabic text classification	SVM,NB, K-NN, FS(PCA,LDA chi-square, mutual information, RARF),TF-IDF, Word2Vec	Khaleej-2004 Arabic, Watan-2004 is a large Arabic

3	Alruily et al.	Classification of Arabic Long Tweets Using Transfer Learning with BERT	BERT - ARABERT4TWC	SANAD of Arabic news stories, HARD, AJGT, ASTD, ArsenTD–Lev
4	Alhaj et al.	Classification through the introduction of Optimal Configuration Determination for Arabic Text Classification (OCATC).	LR, RF, KNN, DT, NN, SVC, LSVM, and SGD	The dataset created by Abuaiadah, 2700 documents, dataset by collect Saad, CNN Dataset
5	Alammary et al.	Applying BERT Models for Arabic Text Classification	BERT, AraBERT, MARBERT, ArabicBERT, ARBERT, XLM-RoBERTa, QARiB, Arabic ALBERT	1780 articles from Google Scholar
6	Alwehaibi et. al.	They are using deep learning embedding methods to optimize sentiment classification of dialectal Arabic short texts at the document level.	LSTM, CNN, and ensemble model	A dataset of Standard Arabic and colloquial Arabic collected from Twitter, AraSenTi
7	Saeed et al.	Various classification methods for spam detection in Arabic opinion texts were investigated. The methods examined include rule-based approaches, classical machine learning techniques, majority voting ensemble, and stacking ensemble methods. Additionally, N-gram features were used to enhance the classification process.	N-gram feature, voting ensemble classifier, Stacking ensemble classifier	DOSC dataset is translated from English language to Arabic language, HARD
8	Nassif et al.	Arabic Fake News Classification	Arabert,Arabic-Bert, ArBert, MARBert, Araelectra and QaribBert	Single-class (SANAD), multi-class (NADiA)
9	Almaliki et. al.	Classification of Arabic hate speech on social media platforms, specifically Twitter, using the Arabic BERT-Mini Model (ABMM).	BERT-Mini Model (ABMM), Bert, LSTM, CNN+LTSM, Linear SVC, SVC, SGD	
10	Mujahid et.al	Automatically classify tweets and opinions of Arab users about ChatGPT technology using large models.	RoBERTa, XLNet , DistilBERT	Tweepy SNscrape Python library using various hash-tags such as # Chat-GPT, #OpenAI, #Chatbot, Chat-GPT3 collect 27000 tweet
11	Mardiansyah et. al.	Compare the performance of ChatGPT-4 and Google Gemini in spam detection.	ChatGPT-4, Google Gemini	SpamAssassin public mail corpus
12	Alhaj et. al.	Enhancing the classification of cognitive distortions in Arabic content on Twitter.	BERTopic, DT, SVM, R.F., KNN, XGBoost, Bagging, and Stacking	Public Therapist Q&A, collected from different resources, and most of them are not publicly available
13	Hassan et. al.	A Comparison of Learning Model Performance in Text Classification Across Various Datasets	SVM,K-NN -L.R. -MNB RF	IMDB dataset, SPAM dataset
14	Elnagar et al.	Comparison of Various Deep Learning Models for Arabic Text Categorization	convolutional-GRU - attention-GRU	Single-label (SANAD), multilabel (NADiA)
15	Alsaleh et al.	Optimize the parameters of the Convolutional Neural Network (CNN) to enhance Arabic text classification using a genetic algorithm.	CNN with GA	AlRiyadh Newspaper and Saudi Press Agency (SPA)
16	Samah M.Alzanin et. al.	Apply feature extraction using Word2vec and TF-IDF on multiple models to classify Arabic tweets and compare the accuracy of the best models.	GNB- SVM- RF	Thirty-five thousand six hundred twenty-seven collected tweets were manually annotated into five categories.
17	Galal et al.	Classification of Arabic texts using BERT as a feature extractor without fine-tuning performance.	BERT, MARBERT, Qarib, ARBERT, GigaBERT	ArSarcasm-v2 Dataset for Arabic sarcasm
18	Mahmoudi et. al.	Classifying Arabic news as fake or real by creating a machine learning model.	R.F., L.R., Linear SVM, Gradient Boosting Classifier	Arabic fake news (DAFN)
19	Kaddoura et. al.	Classify fake advertisements and illegal content on Twitter using machine learning (ML) and deep learning (DL).	ML – DL	Twitter API and labeled manually to get a reliable dataset

20	Elgeldawi et al.	A thorough comparative analysis of different hyperparameter tuning techniques.	hyperparameter tuning by using techniques, including Grid Search, Random Search, Bayesian Optimization, Particle Swarm Optimization (PSO), Genetic Algorithm (GA), six machine learning algorithms LR, RC, SVC, DT), RF, NB) classifier	Booking.com website to collect hotel reviews. A total of 3500 positive, 3500 negative
21	Chowdhury et al.	Extract meaningful information from published abstracts and classify it into three fields: Science, Business, and Social Science.	Support Vector Machines, Naïve Bayes, K-Nearest Neighbour and Decision Tree	107 research abstracts are collected to build the dataset
22	Decui et al.	Work on classifying paragraphs into various topics within policy.	ensemble convolution neural network (CNN)	policy in China
23	Onan et al.	utilized several keyword extraction methods because they are important for capturing the content of the document.	most frequent measure-based keyword extraction, comparing base learning algorithms (Naïve Bayes, support vector machines, logistic regression, and Random Forest) with five widely utilized ensemble methods (AdaBoost, Bagging, Dagging, Random Subspace, and Majority Voting)	Reuters-21578 document collection, ACM document collection
24	Husain	Impact of using single learner machine learning approaches compared to ensemble machine learning approaches for text classification in Arabic.	single learner machine learning approach and ensemble machine learning	Open-Source Arabic Corpora and Corpora Processing Tools (OSACT) in Language Resources and Evaluation Conference (LREC) 2020
25	Ghourabi	The text focuses on developing an ensemble learning strategy that uses an optimized weighted voting technique to improve classification results. This approach incorporates a text embedding method based on a pre-trained language model, aiming to enhance performance in classification tasks.	Embedding technique based on the GPT-3 Transformer is used to represent text messages as dense numerical vectors. An ensemble of classifiers, including SVM, KNN, CNN, and LightGBM, is created through a voting mechanism.	SMS Spam Collection Dataset
26	Al-Qerem	Classifying Arabic news articles based on their content using deep learning techniques.	RNN, LSTM, CNN-LSTM, and Bi-LSTM	data set of news articles Arabic
27	Hariyadi	Classify the Al-Quran into various topics like faith, deeds, and science to enhance understanding of the text.	majority voting approach	Al-Quran
28	Xu, S., Li, Y	Classifying documents with machine learning.	Multinomial NB, Bayesian Multinomial Classifier	20 newsgroup data contain 18,828
29	Bahbib	Investigating the effect of different stemming methods and word representations on Arabic text classification.	Word2Vec, FastText, and GloVe, LSTM, Bi-LSTM, GRU, and Bi-GRU	
30	Alshehri	A classification approach for identifying dialectal Arabic speech acts on Twitter using transformer-based deep learning models.	various BERT models,	Arabic sentiment analysis dataset (ASAD) , Arabic Tweet Act dataset (ArSAS)
31	Al-Fuqaha'a	The text is centered on creating a model for multi-class text classification specifically within the Arabic healthcare sector.	feature extraction techniques (TF-IDF, Word2Vec), LR, RF, MNB, SGD, SVM),. Ensemble methods(stacking and bagging) and AraBERT and MarBERT	Altibbi dataset

A. Data Preparation

In this study, we collected a dataset from the social media platform (X) to analyze requests and suggestions from citizens and residents about the quality and performance of municipal services across various cities in the Kingdom of Saudi Arabia.

After gathering this feedback, we sorted, filtered, and selected comments that specifically addressed improvements to municipal services to help develop an effective classification model for these requests and suggestions. The dataset includes a total of 3,714 comments and suggestions, with most contributions focusing on critical areas such as infrastructure services, including roads, sidewalks, and lighting; issues related to water leakage and sidewalk subsidence; concerns regarding street excavations; and requests for enhancing green spaces, including tree planting and park development in residential areas. Details of the dataset can be found in Table II, Table III, and Fig. 3.

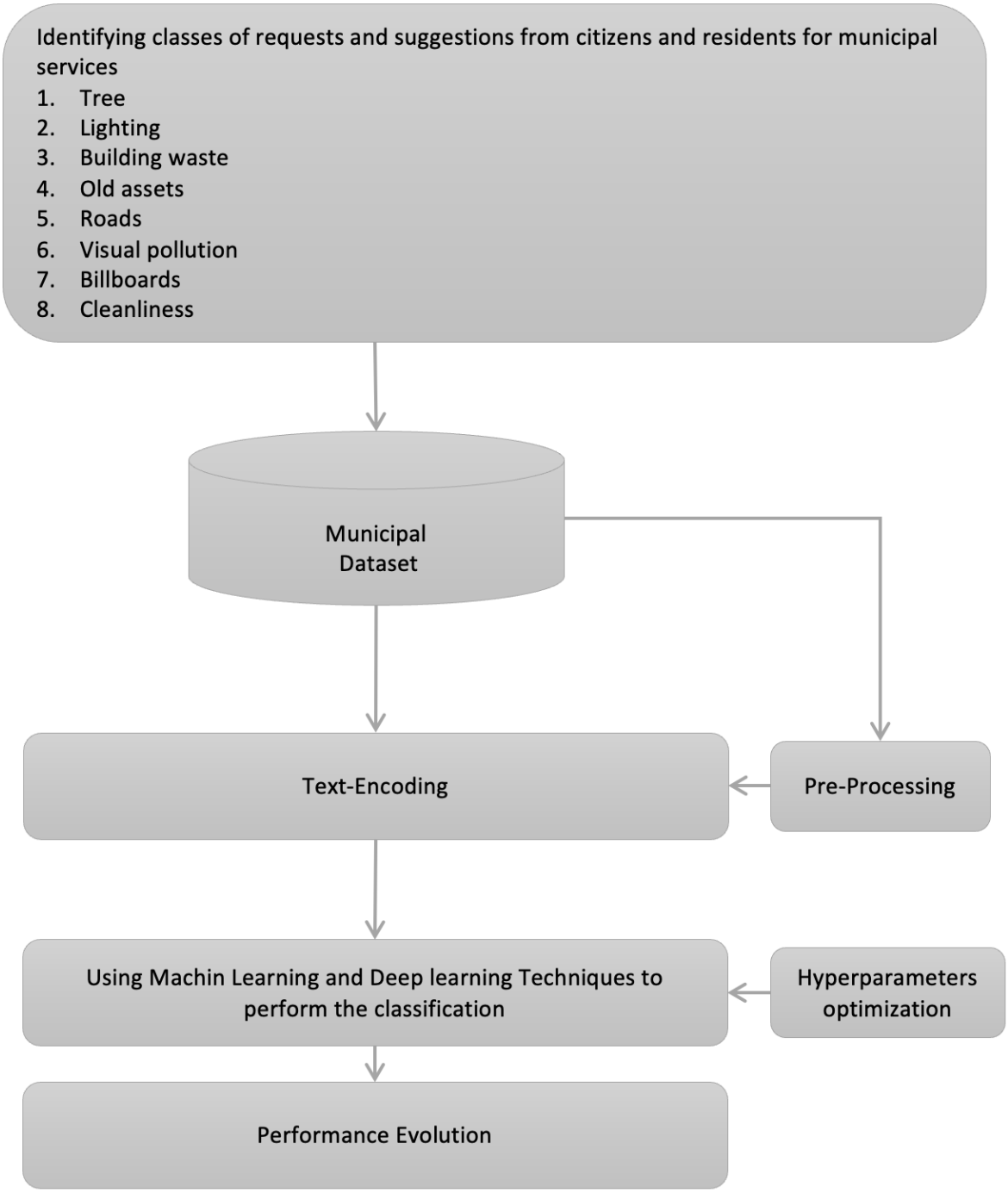


Fig. 2. Block diagram of the proposed methodology.

TABLE II. DATASET AND CLASS DESCRIPTIONS

Class	Description	No. of records
Tree	This category includes tweets that focus on suggestions for planting more trees, trimming existing ones, maintaining tree health, addressing insect infestations, and enhancing the care of green spaces. These tweets reflect public interest in sustainable urban development and emphasize the importance of green areas in improving environmental quality and community well-being.	400
Lighting	The Lighting category in this dataset includes tweets from citizens and residents requesting the installation or improvement of lighting in various municipal areas to enhance visibility and safety. Many tweets also emphasize the need for ongoing maintenance of public lighting systems, highlighting issues such as dim or non-functional streetlights and insufficient illumination in certain public spaces. This systematic categorization of tweets allows for a structured analysis of public feedback regarding lighting infrastructure, helping municipal authorities identify specific areas that need attention and improvement in lighting services.	550
Building Waste	Tweets in this category primarily focus on reporting construction waste in urban neighborhoods and include questions about available disposal methods. These comments reflect the interest of citizens and residents in managing construction debris and seeking solutions for waste disposal within residential areas to promote more sustainable urban environments. By analyzing this feedback, municipal authorities can gain valuable insights into community needs and develop targeted strategies to improve construction waste management practices.	337
Old assets	Tweets in this category consist of questions from citizens and residents about abandoned vehicles in residential neighborhoods and concerns regarding other neglected assets in these areas. These observations demonstrate the community's interest in understanding how to report old and abandoned vehicles. By analyzing these tweets, the municipal sector can effectively allocate follow-up teams to address cases of abandoned vehicles and facilitate their removal.	400
Roads	The tweets express the demands of citizens and residents for improvements in infrastructure. They specifically highlight the need for resurfacing roads, addressing subsidence issues affecting roads and sidewalks, and installing speed bumps to reduce vehicle speeds in residential neighborhoods. These comments underscore the essential role of local government in enhancing and developing infrastructure. By analyzing these tweets, municipal authorities can more effectively identify areas in need of maintenance, allowing them to prioritize initiatives that will improve urban living conditions.	916
Visual pollution	Tweets in this category consist of requests from citizens and residents urging the removal of graffiti from the walls of public buildings. This feedback highlights the community's desire for clean and well-maintained public spaces, emphasizing the importance of aesthetic upkeep in urban environments. By addressing these requests, municipal authorities can improve the appearance of public facilities and foster a greater sense of pride within the community.	251
Billboards	Tweets in this category address requests for the removal of unlicensed advertising posters and billboards located at the entrances of residential buildings, on sidewalks, and in public areas. These unauthorized displays can disrupt the visual appeal of the environment and violate legal standards. By responding to these concerns, municipal authorities can help maintain the integrity of public spaces and ensure compliance with local regulations, fostering a more aesthetically pleasing community.	76
Cleanliness	This category includes tweets that focus on the cleanliness of public spaces. Common topics cover the condition of roads, sidewalks, and waste management. Discussions often highlight issues such as littering, the frequency of waste collection, and the overall maintenance of public areas. These tweets reflect public sentiment regarding the effectiveness of municipal services in maintaining a clean and hygienic environment and may include suggestions for improvement.	600

TABLE III. SAMPLES OF THE DATASET. EACH ROW CONTAINS ONE SAMPLE, AND THE ROW HAS THREE SUB-ROWS: THE RAW COMPLAINT, THE PREPROCESSED COMPLAINT, AND THE TRANSLATION OF THE PREPROCESSED COMPLAINT

#	Complaint	Class
1	السلام عليكم ورحمة الله وبركاته الموقع بحاججججججججة الى تنفيذ مطب وعمل خطوط عبور مشاة بالإضافة الى عمل صيانة للمطب! القديم وذلك ان الموقع بجوار مدرسة ثانوية 24.44817584,39.60250383الموقع للتواصل رقم 0554472280	Roads
	الموقع بحاجه تنفيذ مطب وعمل خطوط عبور مشاة بالإضافة عمل صيانة للمطب الموقع بجوار مدرسة ثانوية	
	The site must implement a speed bump and install pedestrian crossings, while also ensuring the speed bump is maintained. This location is adjacent to a secondary school.	
2	السلام عليكم ورحمة الله وبركاته , في الممشى العام كتابات مشوهة للمظهر العام!!!1 مكتوبة على الكراسي.....	visual pollution
	الممشى العام كتابات مشوهة للمظهر العام مكتوبة على الكراسي	
	Public walkway with distorted graffiti written on chairs.	
3	..... الله يقويكم لايوجد لدينا اشجار في الحي نأمل منكم اطلاق حملة للتشجير وشكرا لكم الحي	Tree
	الله يقويكم لايوجد لدينا اشجار في الحي نأمل منكم اطلاق حملة للتشجير وشكرا لكم الحي	
	May God grant you strength. Our neighborhood lacks trees, and we hope you will initiate a tree-planting campaign. Thank you from the neighborhood.	
4	السلام عليكم عامود إنارة يحتاج استبدال ووضع غطاء لغرفة التوزيع رقم بلاغ :48848	Lighting
	عامود إنارة يحتاج استبدال ووضع غطاء لغرفة التوزيع	
	The lamp post needs to be replaced, and the cover for the distribution room needs attention.	
5	مخلفات بناء ترمى في الأراضي البيضاء وهذا تشوة بصري رقم البلاغ 67874	Building Waste
	مخلفات بناء ترمى في الأراضي البيضاء وهذا تشوه بصري	
	Construction waste is being discarded in open areas, creating a visual disturbance.	
6	السلام عليكم اراضي اصبحت مكب للنفايات داخل الاحياء السكنيه وتشكل خطر كبير حيث انها حفره جدا عميقه وتشكل خطر على الاطفال والسيارات في أحد الاحياء السكنية رقم الجوال 0554472280	Cleanliness
	اراضي اصبحت مكب للنفايات داخل الاحياء السكنيه وتشكل خطر كبير حيث انها حفره جدا عميقه وتشكل خطر على الاطفال والسيارات في احد الاحياء السكنيه	
	Lands that have become dumping grounds for waste in residential neighborhoods pose significant dangers, as they contain very deep holes that threaten children and vehicles.	
7	لوحة ممزقة,وملصقات دعائية ,,, في وسط الحي السكني ,,,, وعلى مبنى العمارة !!!! تشوه بصري	Billboards
	لوحة ممزقة وملصقات دعائية في وسط الحي السكني وعلى مبنى العمارة تشوه بصري	
	Tattered signboards and advertising posters are prevalent in the residential area and on building facades, causing visual distortion.	
8	باصات قديمة بدون لوحات متوقفة داخل الاحياء السكنية.....	Old assets
	باصات قديمة بدون لوحات متوقفة داخل الاحياء السكنية	
	"Old buses without license plates are parked in residential areas."	

B. Dataset Processing

Various methods were employed to process the dataset, all aimed at enhancing the accuracy of machine learning and deep learning models in classifying Arabic texts more effectively. Removing punctuation—including periods, question marks, exclamation points, commas, colons, semicolons, dashes, hyphens, brackets, braces, parentheses, apostrophes, and quotation marks—provides several benefits. These benefits include reduced noise, improved tokenization performance, enhanced model accuracy, a unified text format, and better text representation as vectors. Additionally, removing extra spaces and links improves readability, boosts tokenization results, reduces data complexity, increases model efficiency, and focuses analysis on relevant content.

Stop words are common words in Arabic that frequently appear in the text but contribute little to its overall meaning. Removing these stop words can reduce data size, minimize noise, improve accuracy in text classification, and facilitate keyword extraction. The NLTK library in Python is often used to implement stop word removal, streamlining the text for more effective processing and analysis.

Additionally, removing tabs enhances data consistency, aids in understanding, and reduces errors. Eliminating numbers from the text is also an essential part of data processing, as they can

add complexity. By removing numbers, the data is purified, which enhances the algorithms used, increases analysis accuracy, standardizes the data, facilitates comprehension, and improves extraction processes.

Feature extraction techniques in Natural Language Processing (NLP) involve transforming raw data into features usable in machine learning (ML) and deep learning (DL) models. There are several methods available for converting text into vector representations for numeric analysis.

Optimal hyperparameters are external settings for ML and DL models that control their behavior and learning processes. Various algorithms can be utilized to tune these hyperparameters to achieve the best results.

The Grid Search technique is commonly employed to find the optimal set of hyperparameters for a model in ML and DL. It systematically tests a predefined set of hyperparameters and evaluates the model's performance for each combination using cross-validation. In contrast, Random Search selects random combinations of hyperparameters for evaluation. This approach significantly reduces computation time while still exploring a wide range of values, and it often proves to be more efficient than Grid Search.

IV. Results and Analysis

This section discusses the outcomes of applying various text classification models to a dataset containing requests and suggestions from citizens and residents about municipal services in the Kingdom of Saudi Arabia. The models' performance was evaluated using the F1 score to identify the most effective algorithm for classifying different topics, such as lighting, construction waste, neglected assets, roads, visual pollution, billboards, and cleanliness. Our results indicate significant improvements in the performance of both machine learning (ML) and deep learning (DL) models, demonstrating the benefits of hyperparameter optimization combined with dataset pre-processing. Furthermore, the robustness of these models was evident even without pre-processing, further highlighting their applicability to text classification tasks.

TABLE IV. Values of Hyperparameter of ML Models

Models	Hyperparameters
Logistic Regression	solver='newton-cg',penalty='none',max_iter=1000)
Naive-Multinomial	alpha=0.1
Naive-Complement	alpha=1.20
Naive-Bernoulli	alpha=0.001
Random Forest	n_estimators=86
SVC	C=1000, gamma= 0.01, kernel= 'rbf'
Linear SVC	C=2,penalty='l2'
K-Neighbors	(n_neighbors=7)
Decision Tree	criterion='gini',max_depth=100,splitter='random',min_samples_split=2
SGD	(penalty='l2')
Bag(LR)	solver='newton-cg',penalty='none',max_iter=1000),n_estimators=30
Bag(SVC)	SVC(C=1000, gamma= 0.2, kernel= 'rbf'),n_estimators=50))
Bag(K-NN)	n_estimators=80
Bag(SGD)	n_estimators=40
AdaBoost	n_estimators=80,learning_rate=0.01,algorithm="SAMME.R"
XGB	n_estimators=50,gamma=0.5
Gradient Boosting	n_estimators=125,max_depth=7
Vot(RF-LR)	estimators=[('lr',LogisticRegression(solver='newton-cg',penalty='none',max_iter=1000)),('rf',RandomForestClassifier(n_estimators=86))],voting='soft'))
Vot(L-SVC,SVC)	estimators=[('ls',LinearSVC(C=2,penalty='l2')),('sv',SVC(C=1000, gamma= 0.01, kernel= 'rbf'))],voting='hard'

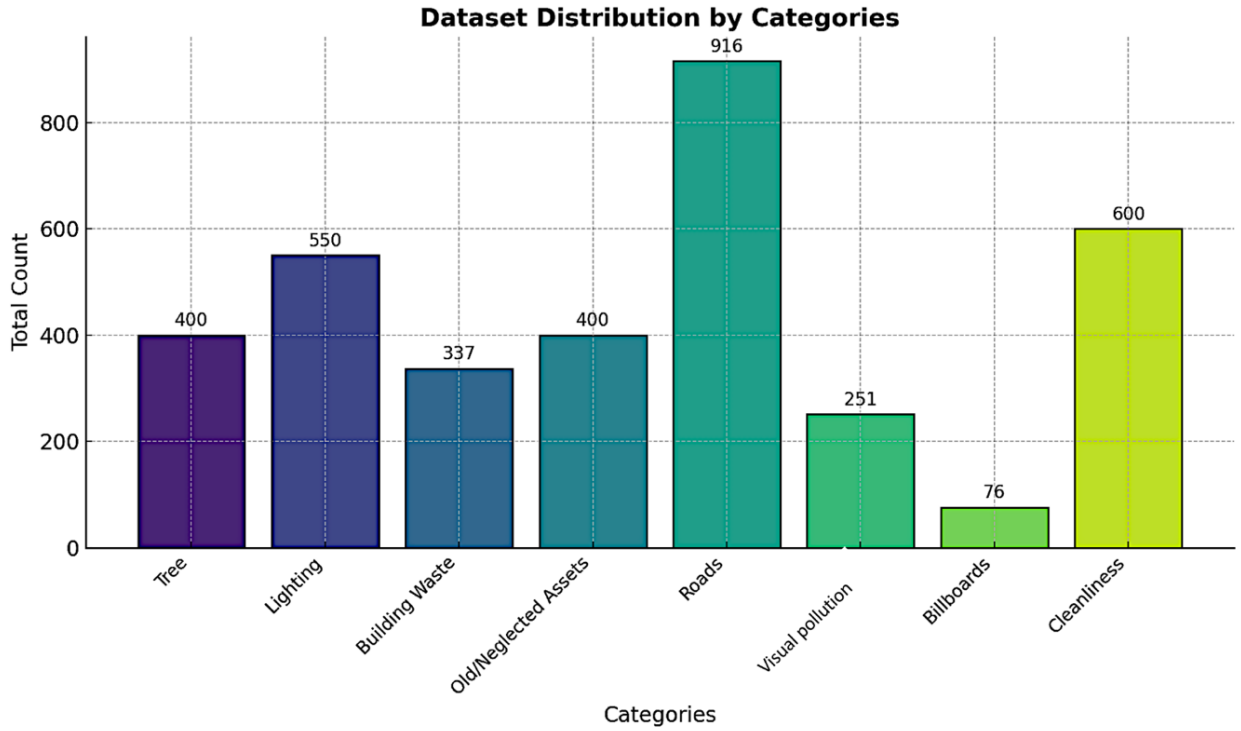


Fig. 3. The different classes are included in the dataset.

A. Performance Metrics

We utilized the following metrics to assess the performance of the model:

- Accuracy: The number of data instances correctly classified from the total dataset is a crucial measure of a model's effectiveness in categorization. This metric is particularly significant for binary classification problems when the dataset is balanced. However, when the dataset is unbalanced, it is important to consider additional metrics alongside accuracy, such as recall, precision, and F1 scores.

$$Accuracy = \frac{True\ Positive(TP)+True\ Negative(TN)}{Total\ Number\ of\ Sample} \quad (1)$$

- Precision: The ratio of correctly predicted positive instances to the total predicted positives, reflecting the model's ability to identify relevant classes.

$$Precision = \frac{True\ Positive(TP)}{True\ Positives(TP)+False\ Positives(FP)} \quad (2)$$

- Recall: The proportion of actual positive instances that are correctly identified by the model, which measures its sensitivity.

$$Recall = \frac{True\ Positive(TP)}{True\ Positives(TP)+False\ Negatives(FN)} \quad (3)$$

- F1 Score: A metric that combines precision and recall into a single score, ranging from 0 (worst) to 1 (best).

$$F1 = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (4)$$

TABLE V. F1 SCORES OF ALL MODELS UNDER PROCESSED DATASET CONDITIONS

Models	F1_Score using optimal hyperparameters	F1_Score using default hyperparameters
LR	**0.901734104**	**0.865606936**
Bagging (LR)	0.901734104	0.855491329
Bagging (SVC)	0.8959537	0.8439306
MNB	0.878612716	0.791907514
RF	0.875722543	0.85982659
XGB	0.854046243	0.852601156
Bagging(K-NN)	0.835260116	0.820809249
BNB	0.862716763	0.556358382
K -NN	0.832369942	0.826589595
DT	0.822254335	0.819364162
AdaBoost	0.423410405	0.39017341
RNN	0.816511235	0.816511235

B. F1 Scores with Processed Dataset

The F1 score, which balances precision and recall, is a critical metric for evaluating the classification of imbalanced datasets. In this context, both Logistic Regression and its ensemble counterpart play significant roles.

We aimed to identify the optimal or near-optimal hyperparameters for the classification techniques used; the results of this phase are detailed in Table IV.

As indicated in Table V, bagging with Logistic Regression achieved the highest F1 score of 90.17% under pre-processed conditions. While deep learning models like RNN performed reasonably well, obtaining an F1 score of 81.16%, their performance was still marginally lower than that of the best-performing machine learning models. Overall, models that utilized pre-processing and optimal hyperparameters with machine learning algorithms—such as Logistic Regression, Bagging (LR), and Bagging (SVC)—consistently achieved superior F1 scores. This demonstrates that machine learning algorithms outperformed deep learning models when evaluated using the F1 metric.

C. F1 Scores without Preprocessing the Dataset

The F1 scores reflect the results achieved using the dataset without any pre-processing, while applying optimal hyperparameters, as detailed in Table IV. This method emphasizes the model's ability to effectively balance precision and recall, even when dealing with the challenges posed by unprocessed data. Evaluating the F1 scores in this way provides insight into the model's capacity to maintain accuracy and reliability while working with raw data and appropriately tuned parameters.

As illustrated in Table VI, Linear SVC, along with Bagging using Linear SVC and Voting (combining Linear SVC and SVC), showed superior accuracy rates of 91.04%, 90.46%, and 90.08%, respectively, when hyperparameter optimization was performed without any pre-processing. Among deep learning models, CNN and LSTM achieved noteworthy accuracy rates of 88.43% and 87.81%, respectively, although they did not perform as well as some of the specific machine learning models.

TABLE VI. F1 SCORES OF ALL MODELS UNDER UNPROCESSED DATASET CONDITIONS

Models	F1_Score using optimal hyperparameters	F1_Score using default hyperparameters
Linear SVC	**0.910404624**	**0.913179191**
SGD	0.901734104	0.900289017
Bagging (Linear SV)	0.904624277	0.897398844
Bagging(SGD)	0.907514451	0.904624277
Voting(LR,RF)	0.904624277	0.875722543
Voting(LinearSVC,SVC)	0.908959538	0.89017341
CNB	0.894508671	0.891618497
SVC	0.893063584	0.856936416
Gradient Boosting	0.845375723	0.841040462
CNN	0.884338846	0.83706596087
LSTIM	0.878148111	0.863887078
BiLSTM	0.871606769	0.86308049

V. DISCUSSION

The results highlight the impact of hyperparameter optimization and dataset processing techniques on the performance of machine learning (ML) and deep learning (DL) models when evaluating requests and suggestions from citizens regarding various municipal services. These services include categories such as trees, lighting, construction waste, neglected assets, roads, visual pollution, billboards, and cleanliness.

The findings indicate that ML models consistently outperformed DL models when using optimized hyperparameters and processed datasets. However, DL models showed performance levels that were close to those of ML models, particularly when hyperparameters were finely tuned without preprocessing the datasets.

Several factors contributed to the enhanced accuracy of ML models in classifying these requests and suggestions. Notably, ML models can effectively learn from smaller datasets due to their reliance on feature-based methods, while DL models typically require larger datasets to capture and model complex patterns effectively.

In summary, the results demonstrate that when utilizing optimized hyperparameters and processed datasets, machine learning models achieved higher accuracy compared to deep learning models. Conversely, deep learning models exhibited competitive performance levels, particularly when hyperparameters were finely tuned without preprocessing. The ability of machine learning models to learn effectively from smaller datasets is a significant factor in their accuracy in classifying requests related to municipal services.

VI. CONCLUSION

This study explored the application of machine learning (ML) and deep learning (DL) models to classify Arabic-language text related to municipal services in Saudi Arabia. To address the scarcity of Arabic-language datasets, a collection of 3,714 requests and suggestions was compiled. This dataset will be made publicly available to encourage the community's adoption of ML and DL technologies in the field of municipal services. A baseline for classification performance was established to guide future improvements.

The findings are significant: ML models, including Logistic Regression and Linear Support Vector Classifier (Linear SVC), demonstrated superior performance, achieving a maximum accuracy of 91.04% and an F1 score of 90.17%. These results were obtained through the application of data preprocessing techniques and hyperparameter optimization. Although DL models, such as Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM), also performed well, they did not surpass the ML models. This difference can be attributed to the fact that DL models are better suited for longer texts, while ML models excel with shorter texts, which is characteristic of the dataset used in this study.

The study emphasizes the critical role of data preprocessing and hyperparameter tuning in enhancing classification accuracy. By introducing this dataset and sharing key insights, this work establishes a foundation for future research and development in natural language processing (NLP) for the Arabic language. Future efforts could focus on expanding the dataset, integrating advanced embedding techniques and large language models, and designing dynamic classification systems for real-time applications.

Future research can explore several enhancements to improve classification robustness and generalizability. Expanding the dataset with more samples will strengthen model performance, while incorporating advanced embedding techniques, such as BERT or FastText, can refine text representation and capture nuanced meanings more effectively. Developing real-time classification systems could significantly enhance municipal responsiveness and service efficiency. Another valuable enhancement is integrating multimodal data—combining text, images, and audio—to achieve a more comprehensive understanding of citizens' needs. Lastly, implementing continuous feedback mechanisms will enable model refinement, ensuring adaptability to evolving demands.

ACKNOWLEDGMENT

This research is funded by the Deanship of Scientific Research at the Islamic University of Madinah, Madinah, KSA.

REFERENCES

[1] Rincy, T. N., & Gupta, R. (2020, February). Ensemble learning techniques and its efficiency in machine learning: A survey. In 2nd international conference on data, engineering and applications (IDEA) (pp. 1-6). IEEE.

[2] Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. IEEE access, 7, 53040-53065.

[3] Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021, November). Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. In Informatics (Vol. 8, No. 4, p. 79). MDPI.

[4] Bahbib, M., & Yakhlef, M. B. (2024, May). Effects of applying stemmers with word representation on Arabic text classification performance using deep learning methods. In 2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET) (pp. 1-6). IEEE.

[5] Umer, M., Imtiaz, Z., Ahmad, M., Nappi, M., Medaglia, C., Choi, G. S., & Mehmood, A. (2023). Impact of convolutional neural network and FastText embedding on text classification. Multimedia Tools and Applications, 82(4), 5569-5585.

[6] Alshehri, K., Alhothali, A., & Alowidi, N. (2024). Arabic Tweet Act: A Weighted Ensemble Pre-Trained Transformer Model for Classifying Arabic Speech Acts on Twitter. arXiv preprint arXiv:2401.17373. DOI: https://doi.org/10.48550/arXiv.2401.17373.

[7] Al-Fuqaha'a, S., Al-Madi, N., & Hammo, B. (2024). A robust classification approach to enhance clinic identification from Arabic health text. Neural Computing and Applications, 36(13), 7161-7185. https://doi.org/10.1007/s00500-023-06223-7.

[8] Onan, A., Korukoğlu, S., & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. Expert Systems with Applications, 57, 232-247.

[9] Ghourabi, A., & Alohaly, M. (2023). Enhancing spam message classification and detection using transformer-based embedding and ensemble learning. Sensors, 23(8), 3861.

[10] Al-Qerem, A., Raja, M., Taqatqa, S., & Sara, M. R. A. (2024). Utilizing Deep Learning Models (RNN, LSTM, CNN-LSTM, and Bi-LSTM) for Arabic Text Classification. In Artificial Intelligence-Augmented Digital Twins: Transforming Industrial Operations for Innovation and Sustainability (pp. 287-301). Cham: Springer Nature Switzerland.

[11] Hariyadi, B., & Lhaksmana, K. M. (2024, February). Topic Classification of Arabic Text Using Majority Voting. In 2024 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS) (pp. 1-5). IEEE.

[12] Ghourabi, A., & Alohaly, M. (2023). Enhancing spam message classification and detection using transformer-based embedding and ensemble learning. Sensors, 23(8), 3861.

[13] El Rifai, H., Al Qadi, L., & Elnagar, A. (2022). Arabic text classification: the need for multi-labeling systems. Neural Computing and Applications, 34(2), 1135-1159.

[14] Umer, M., Imtiaz, Z., Ahmad, M., Nappi, M., Medaglia, C., Choi, G. S., & Mehmood, A. (2023). Impact of convolutional neural network and FastText embedding on text classification. Multimedia Tools and Applications, 82(4), 5569-5585.

[15] Chowdhury, S., & Schoen, M. P. (2020, October). Research paper classification using supervised machine learning techniques. In 2020 Intermountain Engineering, Technology and Computing (IETC) (pp. 1-6). IEEE.

[16] Nassif, A. B., Elnagar, A., Elgendy, O., & Afadar, Y. (2022). Arabic fake news detection based on deep contextualized embedding models. Neural Computing and Applications, 34(18), 16019-16032.

[17] Elnagar, A., Al-Debsi, R., & Einea, O. (2020). Arabic text classification using deep learning models. Information Processing & Management, 57(1), 102121.

[18] Alsaleh, D., & Larabi-Marie-Sainte, S. (2021). Arabic text classification using convolutional neural network and genetic algorithms. IEEE Access, 9, 91670-91685.

[19] Alzanin, S. M., Azmi, A. M., & Aboalsamh, H. A. (2022). Short text classification for Arabic social media tweets. Journal of King Saud University-Computer and Information Sciences, 34(9), 6595-6604.

[20] Galal, O., Abdel-Gawad, A. H., & Farouk, M. (2024). Rethinking of BERT sentence embedding for text classification. Neural Computing and Applications, 36(32), 20245-20258.

[21] Mahmoudi, O., Bouami, M. F., & Badri, M. (2022). Arabic language modeling based on supervised machine learning. Revue d'Intelligence Artificielle, 36(3), 467.

[22] Kaddoura, S., Alex, S. A., Itani, M., Henno, S., AlNashash, A., & Hemanth, D. J. (2023). Arabic spam tweets classification using deep learning. Neural Computing and Applications, 35(23), 17233-17246.

[23] Alzanin, S. M., Gumaei, A., Haque, M. A., & Muaad, A. Y. (2023). An optimized Arabic multilabel text classification approach using genetic algorithm and ensemble learning. Applied Sciences, 13(18), 10264.

[24] Sabri, T., Bahassine, S., El Beggar, O., & Kissi, M. (2024). An improved Arabic text classification method using word embedding. International Journal of Electrical & Computer Engineering (2088-8708), 14(1).

[25] Alruily, M., Manaf Fazal, A., Mostafa, A. M., & Ezz, M. (2023). Automated Arabic long-tweet classification using transfer learning with BERT. Applied Sciences, 13(6), 3482.

[26] Saeed, R. M., Rady, S., & Gharib, T. F. (2022). An ensemble approach for spam detection in Arabic opinion texts. Journal of King Saud University-Computer and Information Sciences, 34(1), 1407-1416.

[27] Hassan, S. U., Ahamed, J., & Ahmad, K. (2022). Analytics of machine learning-based algorithms for text classification. Sustainable Operations and Computers, 3, 238-248.

[28] Alhaj, Y. A., Dahou, A., Al-Qaness, M. A., Abualigah, L., Abbasi, A. A., Almaweri, N. A. O., ... & Damaševičius, R. (2022). A novel text classification technique using improved particle swarm optimization: A case study of Arabic language. Future Internet, 14(7), 194.

[29] Alammary, A. S. (2022). BERT models for Arabic text classification: a systematic review. Applied Sciences, 12(11), 5720.

[30] Almaliki, M., Almars, A. M., Gad, I., & Atlam, E. S. (2023). Abmm: Arabic bert-mini model for hate-speech detection on social media. Electronics, 12(4), 1048.

[31] Mujahid, M., Kanwal, K., Rustam, F., Aljedaani, W., & Ashraf, I. (2023). Arabic ChatGPT tweets classification using RoBERTa and BERT ensemble model. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(8), 1-23.

[32] Mardiansyah, K., & Surya, W. (2024). Comparative Analysis of ChatGPT-4 and Google Gemini for Spam Detection on the SpamAssassin Public Mail Corpus.

[33] Alhaj, F., Al-Haj, A., Sharieh, A., & Jabri, R. (2022). Improving Arabic cognitive distortion classification in Twitter using BERTopic. International Journal of Advanced Computer Science and Applications, 13(1), 854-8

A Novel Hybrid Model Based on CEEMDAN and Bayesian Optimized LSTM for Financial Trend Prediction

Yu Sun[1], Sofianita Mutalib[2*], Liwei Tian[3]
School of Management, Guangdong University of Science and Technology, Dongguan, Guangdong Province, China[1]
School of Computing Sciences, College of Computing, Informatics and Mathematics,
Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia[1, 2]
School of Computing, Guangdong University of Science and Technology, Dongguan, Guangdong Province, China[3]

***Abstract*—Financial time series prediction is inherently complex due to its nonlinear, nonstationary, and highly volatile nature. This study introduces a novel CEEMDAN-BO-LSTM model within a decomposition-optimization-prediction-integration framework to address these challenges. The Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) algorithm decomposes the original series into high-frequency, medium-frequency, low-frequency, and trend components, enabling precise time window selection. Bayesian Optimization (BO) algorithm optimizes the parameters of a dual-layer Long Short-Term Memory (LSTM) network, enhancing prediction accuracy. By integrating predictions from each component, the model generates a comprehensive and reliable forecast. Experiments on 10 representative global stock indices reveal that the proposed model outperforms benchmark approaches across RMSE, MAE, MAPE, and R^2 metrics. The CEEMDAN-BO-LSTM model demonstrates robustness and stability, effectively capturing market fluctuations and long-term trends, even under high volatility.**

***Keywords*—LSTM; Bayesian optimization; CEEMDAN; financial time series; time window selection**

I. Introduction

With the ongoing advancement of global economic integration and the increasing openness of international markets, the influence of stock market fluctuations on the global economy has grown significantly. Accurately capturing stock market trends and forecasting their future movements has thus become a crucial research focus in the financial field. Financial market forecasting methods are primarily categorized into linear and nonlinear models. Early research achieved notable progress in financial time series modeling using linear models [1-3]. However, financial time series typically exhibit high noise, uncertainty, and nonstationary, with dynamic changes in the relationships between variables over time. These limitations often hinder traditional linear models from effectively capturing intricate patterns and sustaining long-term dependencies. Furthermore, the linear models' assumption of data smoothness imposes an additional limitation on their validity.

In order to overcome the limitations of linear models, nonlinear prediction methods have become the mainstream of financial market prediction. Among them, artificial neural networks (ANNs) are widely used due to their ability to process high-dimensional, multimodal and heterogeneous data. For example, Yassin et al. (2017) utilized a stock prediction model based on multilayer perceptron (MLP) to predict the weekly stock price of Apple Inc. and verified its strong prediction performance through one step ahead (OSA) and correlation analysis [4]. Similarly, Zhang et al. (2021) employed a back propagation (BP) neural network to classify and predict stock price patterns, achieving an improved accuracy of 73.29% and providing valuable insights for investors and macroeconomic policies [5]. However, ANN-based models face many challenges, including overfitting, gradient vanishing or exploding, and easy to fall into local optimality, which limit their generalization ability for complex financial data.

With the continuous development and application of deep learning technology, Long Short-Term Memory (LSTM) network has become an important model due to its strong ability to handle noisy, nonlinear and nonstationary data. By introducing advanced gating mechanisms, LSTM effectively overcomes the inherent gradient vanishing and gradient exploding problems of traditional recurrent neural networks (RNNs), making it particularly suitable for processing long sequence data [6, 7]. Its robustness and adaptability have promoted its widespread application in the financial field. Wang et al. (2024) verified that the LSTM model can overcome the limitations of RNN in stock market forecasting and can greatly improve the forecasting performance [8]. These advantages make LSTM a highly promising component in hybrid models for solving complex financial forecasting tasks.

In the field of financial market forecasting, hybrid models have attracted much attention. This type of model combines the strengths of multiple algorithms and offers greater advantages than single models in processing complex and noisy financial data. Compared with single models, hybrid models have stronger generalization capabilities and are more robust, especially in capturing multi-scale features and nonlinear relationships.

In recent years, LSTM-based hybrid models have shown great potential in stock market forecasting due to their ability to capture both long and short-term patterns in financial time series. Table I summarizes recent research based on LSTM

*Corresponding Author.

models, clearly showing the methods and advantages of these hybrid models in stock market forecasting.

TABLE I. SUMMARY OF THE RECENT WORK ON LSTM-BASED STOCK MARKET PREDICTION

Reference	Technique	Dataset	Number of stocks	Metric	Advantages
Mehtarizadeh et al. (2025) [9]	LSTM, Sin-Cosine Algorithm, ARIMA, GARCH	Stock	8	RMSE	Integrates statistical and deep learning models, enhancing the ability to capture the complex dynamics of stock prices.
Baek (2024) [10]	CNN, LSTM, Genetic Algorithm Optimization	Stock index	1	MAPE, MSE, MAE	Effectively improves the accuracy of stock index prediction.
Sang and Li (2024) [11]	Attention Mechanism Variant LSTM	Stock	1	MSE, MAE, R^2	Enhances generalization, prediction accuracy, and convergence ability
Zheng et al. (2024) [12]	Convolutional Neural Network, Bidirectional LSTM, Attention	Stock	1	MSE, MAPE	Demonstrates significantly improved prediction accuracy.
Akşehir et al. (2024) [13]	Two-level decomposition of CEEMDAN, LSTM, Support Vector Regression	Stock index	4	RMSE, MAE, MAPE, R^2	Removes noise from financial time series data, boosting predictive performance.
Gülmez (2023) [14]	Artificial Rabbits Optimization, LSTM	Stock	30	MSE, MAE, MAPE, R^2	Exhibits high generalisability in diverse financial scenarios.
Tian et al. (2022) [15]	LSTM, Bayesian optimization, LightGBM	Diverse financial market datasets	9	RMSE, MAE, F1-score, Accuracy	Provides better approximation and generalisation in stock volatility prediction.

As shown in Table I, most studies on LSTM-based stock market prediction methods employ parameter optimization algorithms. This is mainly because manually adjusting parameters not only increases the computational cost, but may also reduce the prediction accuracy and optimization efficiency. In deep learning networks such as LSTM, hyperparameter selection is crucial to achieve accurate predictions [16]. The performance of LSTM largely depends on the optimal configuration of hyperparameters, such as the number of hidden neurons and the learning rate setting. In addition, since the neural network needs to optimize a large number of parameters, it is prone to falling into local optima during this process and may be at risk of overfitting.

To address these challenges, Bayesian Optimization (BO) provides an efficient solution [17]. BO estimates the objective function through Gaussian Processes (GP) and optimizes it using surrogate modeling. This method can efficiently explore the hyperparameter space, avoid redundant calculations, reduce the risk of falling into local optima, and significantly improve both the effectiveness and training efficiency of the LSTM model, making it more suitable for the complexity of financial time series prediction.

In addition to hyperparameter optimization, data preprocessing also plays a vital role in improving prediction accuracy. High noise is a key characteristic of financial time series and a major challenge for accurate prediction. Signal decomposition methods can transform complex original time series into simpler components, enabling the model to capture potential features more effectively and reduce noise. Among them, variational mode decomposition (VMD) is widely used due to its strong practical performance [18, 19]. However, VMD requires manual selection of the number of modes, introducing subjectivity and uncertainty. In contrast, Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) provides an advanced adaptive data decomposition method that decomposes data into physically meaningful modal components. As a higher-order version of Empirical Mode Decomposition (EMD), CEEMDAN overcomes some of EMD's limitations by introducing adaptive white noise, improving robustness and accuracy in extracting intrinsic features of complex data. CEEMDAN has demonstrated significant theoretical and practical value in economic and financial applications, laying a solid foundation for enhancing model performance [20].

Based on the above analysis, this study introduces a novel hybrid model CEEMDAN-BO-LSTM. This model further improves the performance of financial time series prediction by effectively removing the original signal noise, enhancing the long-term and short-term dependency modeling capabilities, and employing an efficient optimization algorithm for hyperparameter tuning. Specific contributions of this study include:

1) Proposal of an innovative hybrid framework: This study constructs a "decomposition-optimization-prediction-integration" framework. The framework uses the CEEMDAN algorithm to perform multi-scale decomposition of stock index time series, automatically adjusts the hyperparameters of LSTM through BO for prediction, and finally integrates the prediction results. This method greatly enhances predictive performance and adaptability, and effectively overcomes the limitations of traditional models in parameter selection and noise reduction.

2) Generalization ability in stock index prediction: The proposed model is experimentally evaluated on 10 representative stock indices worldwide, leading to consistent conclusions. The results demonstrate its effectiveness in capturing the nonlinear characteristics of stock market volatility and achieving significant improvements in forecast accuracy and stability, further highlighting its applicability in financial time series forecasting.

3) Optimized time window segmentation: This study classifies the decomposed components based on energy contribution, dividing them into different frequency components, and assigns a suitable time window to each component, which can be predicted according to the characteristics of different frequencies, further enhancing the prediction performance.

These contributions enable the CEEMDAN-BO-LSTM model to demonstrate significant potential in addressing complex financial market fluctuations, opening up broader application prospects for financial market analysis.

This paper is structured into five sections. The next section introduces the methods used in this study. Section III describes the construction process of the CEEMDAN-BO-LSTM model. Then, Section IV outlines the experimental procedure and analyzes the results. Finally, Section V presents the conclusions of this study, as well as future research directions.

II. METHODS

A. LSTM

The LSTM network is an enhanced variant of the RNN [5]. By incorporating specialized LSTM units, the LSTM network can effectively store and manage both long-term and short-term information. As illustrated in Fig. 1, each LSTM unit comprises three gate mechanisms and two state variables. They are input gate (i_t), forget gate (f_t), and output gate (o_t), as well as cell state (C_t) and hidden state (h_t).

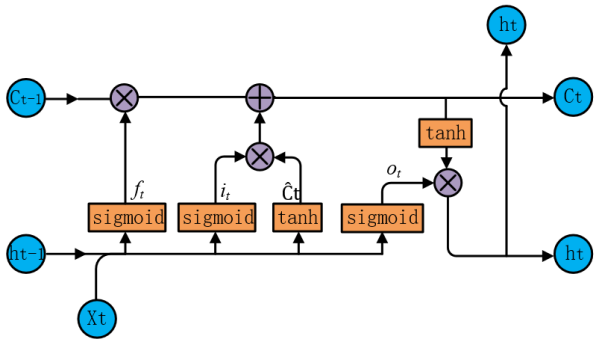


Fig. 1. Structure of LSTM.

According to Fig. 1, these components work collaboratively to regulate the flow of information, allowing LSTM network to maintain and update memory across extended sequences.

The input gate determines the processing of information from the current input data. The mathematical representation is shown in Eq. (1) and Eq. (2).

$$i_t = \sigma(W_i * [h_{t-1}, X_t] + b_i) \quad (1)$$

$$\hat{C}_t = tanh(W_c * [h_{t-1}, X_t] + b_c) \quad (2)$$

The forget gate controls the update of historical data to the memory unit state value. Its mathematical expression is presented in Eq. (3).

$$f_t = \sigma\big(W_f * [h_{t-1}, X_t] + b_f\big) \quad (3)$$

C_t is updated through the combined actions of f_t and i_t. The forget gate regulates the retention of the previous cell state (C_{t-1}), while the input gate introduces new information via the candidate state ($\hat{C}_t$). The update equation is shown in Eq. (4).

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t \quad (4)$$

The output gate determines the information to be emitted at the current time step. The detailed calculation procedure is shown in Eq. (5) and in Eq. (6).

$$o_t = \sigma(W_o * [h_{t-1}, X_t] + b_o) \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

Here, σ (sigmoid function) and tanh (hyperbolic tangent function) serve as activation functions. W and b denote the weight matrices and bias terms, respectively. Through this design, LSTM effectively integrates long-term and short-term information, ensuring stability and accuracy in forecasts. Its flexibility and adaptability make LSTM particularly useful for financial time series forecasting, as it effectively captures the complex nonlinear dynamics of stock market volatility.

B. CEEMDAN

In the field of signal processing, researchers have proposed a variety of advanced methods to analyze the intrinsic characteristics of nonlinear and nonstationary data, including empirical mode decomposition (EMD) and its related techniques. EMD is able to decompose the different frequency components of the original data, revealing the contribution of each frequency component, and providing high-resolution analytical capabilities for nonlinear time series [21].

Although EMD has many advantages, its mode mixing issue limits its accuracy and effectiveness in complex signal decomposition. To address this limitation, researchers proposed Ensemble Empirical Mode Decomposition (EEMD) in 2009. EEMD reduces the impact of mode mixing by adding white noise of different amplitudes to the original signal, performing EMD multiple times, and then averaging the decomposition results. However, since EEMD requires multiple repetitions of the decomposition process, the computational cost is high and may introduce additional computational errors [22].

Based on EEMD, Torres et al. (2011) proposed CEEMDAN [23]. CEEMDAN algorithm further enhances the EEMD by refining the introduction of white noise at each decomposition step. This improvement ensures the consistency of the extracted IMFs, mitigates mode mixing, and reduces reconstruction errors. This method enhances both decomposition accuracy and computational efficiency, making it more reliable for complex and dynamic financial time series analysis.

The following section outlines the detailed algorithmic steps of CEEMDAN:

1) Step 1: Generate noisy data sets. Noisy versions of the original time series $x[n]$ are created using the Eq. (7):

$$x_i[n] = x[n] + \varepsilon_0 w_i[n] \quad (7)$$

Where $w_i[n]$ (i = 1, 2 … I) represent Gaussian white noise that follows a normal distribution, and ε_0 is the standard deviation of the noise. This step introduces controlled noise to reduce mode mixing during the decomposition process.

2) Step 2: Apply the EMD method to each noisy signal to extract the first intrinsic mode function (IMF). The first IMF is

calculated as the average of the IMFs obtained from all noisy realizations as shown in Eq. (8)

$$IMF_1[n] = \frac{1}{I}\sum_{i=1}^{I} IMF_1^i[n] \quad (8)$$

The first residual is then computed as Eq. (9):

$$r_1[n] = x[n] - IMF_1[n] \quad (9)$$

3) Step 3: Iterative decomposition is applied to obtain successive IMFs. For k = 2, ... K, the k-th residue is calculated as shown in Eq. (10). The realizations $r_k[n] + \varepsilon_k EMD_k(w_i[n])$ are decomposed until the first EMD mode is obtained, and the (k+1)-th mode is defined as shown in Eq. (11):

$$r_k[n] = r_{k-1}[n] - IMF_k[n] \quad (10)$$

$$IMF_{k+1}[n] = \frac{1}{I}\sum_{i=1}^{I} EMD_1(r_k[n] + \varepsilon_k EMD_k(w_i[n])) \quad (11)$$

4) Step 4: Proceed to step 3 for the next k. The CEEMDAN algorithm terminates when the residual no longer exceeds the two extreme points, and further decomposition is not possible. The final residual is given by Eq. (12):

$$R[n] = x[n] - \sum_{k=1}^{K} IMF_k \quad (12)$$

This approach effectively resolves mode mixing, ensuring stable and interpretable decomposition results.

C. BO

Bayesian optimization (BO) is a widely used algorithm for optimizing black-box functions with the aim of finding the global optimal solution efficiently. The method achieves optimization by iteratively updating the posterior distribution of the objective function, while the updating process is based on the prior distribution with historical data. Compared to traditional optimization methods, Bayesian optimization excels in handling non-convex problems and avoids falling into local optima [24]. Its objective is to maximize the function f(x) by finding the optimal input x^* within the search space $\chi \in \mathbb{R}$, as shown in Eq. (13):

$$x^* = \underset{x \in \chi}{\operatorname{argmax}} f(x) \quad (13)$$

The essence of Bayesian optimization is rooted in modeling the objective function $f(x)$ as a stochastic process, typically represented by a Gaussian Process (GP) as a surrogate model. The GP provides a probabilistic estimate of $f(x)$ based on historical observations, enabling Bayesian optimization to balance exploration of unexplored regions with exploitation of known high-performing regions. This balance improves optimization efficiency and reduces the risk of missing the global optimum.

Bayesian optimization relies on Bayes' theorem to infer the posterior distribution of f(x) given a historical dataset $D_n = \{(x_i, f(x_i))\}_{i=1}^n$, The posterior distribution is given by Eq. (14):

$$P(f(\mathrm{x})|D_n) = \frac{P\left(D_n \middle| f(\mathrm{x})\right) P(f(\mathrm{x}))}{P(D_n)} \quad (14)$$

Where $P\left(D_n \middle| f(\mathrm{x})\right)$ is the likelihood of the dataset given (x), $P\left(f(\mathrm{x})\right)$ represents the prior distribution of the objective function, and $P(D_n)$ is the evidence.

By employing a Gaussian Process, BO algorithm efficiently estimates $f(\mathrm{x})$ and selects the next sampling point from the posterior distribution $P(f(\mathrm{x})|D_n)$. This iterative process seeks to maximize the objective function while minimizing uncertainty, enabling efficient exploration in high-dimensional spaces.

Bayesian optimization's flexibility and ability to incorporate prior knowledge make it particularly suitable for complex and computationally expensive optimization tasks, such as hyperparameter tuning in machine learning, experimental design, and automated decision-making. Its ability to efficiently handle high-dimensional, noisy, and black-box functions positions it as an effective method for addressing complex optimization challenges.

III. Model Construction

Due to the nonlinear and nonstationary nature of stock market indices, achieving accurate predictions remains a significant challenge. In this context, the combination of effective decomposition techniques and advanced machine learning methods is particularly crucial for handling these complexities. In light of this, this study integrates CEEMDAN temporal decomposition, Bayesian Optimization (BO), and LSTM network to propose a novel stock market forecasting framework: CEEMDAN-BO-LSTM.

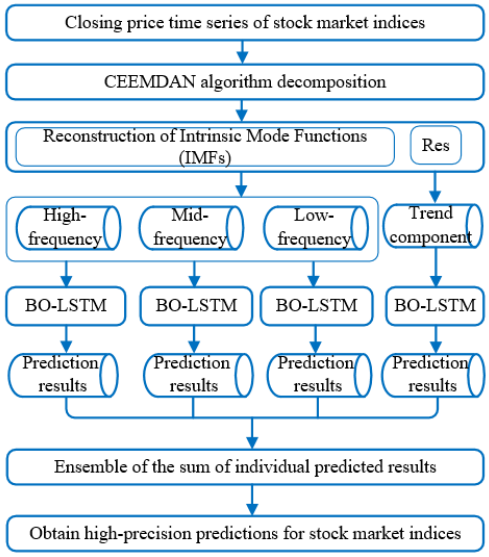


Fig. 2. The workflow of CEEMDAN-BO-LSTM predictive modeling.

As illustrated in Fig. 2, the CEEMDAN-BO-LSTM modeling process addresses the complex, nonlinear, and multi-scale nature of stock market indices through four key steps. In this process, the closing price series of stock market indices serves as input data. The details of these steps are as follows:

1) Step 1: Decomposition. Decompose the closing price time series into multiple IMFs and a residual component using the CEEMDAN algorithm. IMFs capture time series

characteristics at different frequency levels, while the residual represents the overall trend.

2) Step 2: Optimization. The decomposed components are classified according to frequency bands, the time window in the LSTM model is set according to these categories, and the LSTM is then trained using Bayesian optimization.

3) Step 3: Prediction. Optimized LSTM models corresponding to different frequency bands generate predictions for each respective band.

4) Step 4: Integration. Predictions from all components are combined to produce the final forecast for stock market indices. Appropriate evaluation metrics are used to assess the proposed model's reliability and robustness.

IV. Experiment and Result

A. Data Collection and Preprocessing

The historical financial time series dataset used in this study is obtained from Yahoo Finance and covers the daily closing prices of 10 representative global stock indices. The dataset spans 10 years, from October 23, 2014, to October 21, 2024.

Table II presents detailed information on the selected stock indices and their market characteristics. As depicted in Fig. 3, the dynamic trajectories of these indices exhibit typical characteristics of financial time series, including nonlinearity and nonstationary. These characteristics highlight the complexity and challenges associated with forecasting financial trends.

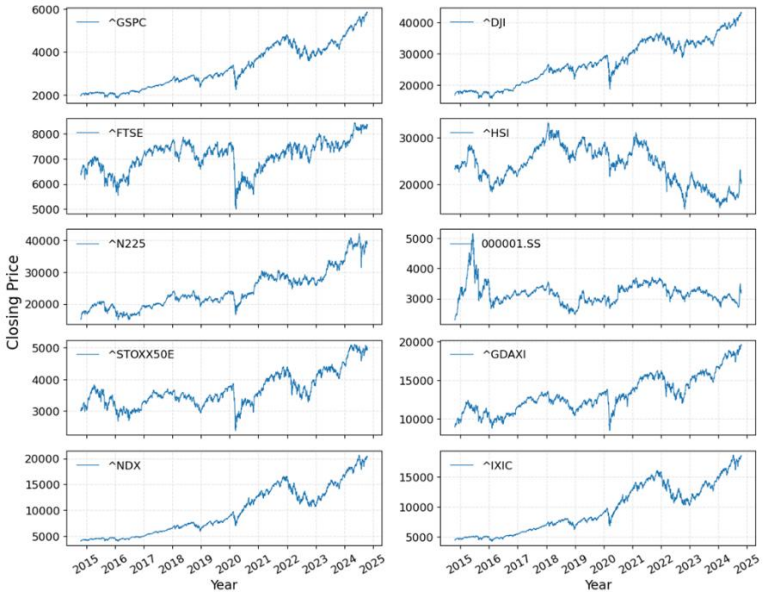


Fig. 3. Historical trends of 10 stock indices.

TABLE II. Major Stock Market Indices and their Characteristics

Index Name	Code	Market Characteristics	Index Name	Code	Market Characteristics
S&P 500 Index	^GSPC	Represents 500 leading U.S. companies across multiple sectors, reflecting the overall U.S. economy.	Shanghai Composite Index	000001.SS	Covers all A-shares on the Shanghai Stock Exchange, serving as a benchmark for China's market.
Dow Jones Industrial Average	^DJI	Consists of blue-chip stocks, reflecting trends in the industrial and financial markets.	Euro Stoxx 50 Index	^STOXX50E	Tracks 50 leading European companies and represents key sectors of the eurozone economy.
FTSE 100 Index	^FTSE	Includes 100 major UK companies, providing insights into the UK and European financial markets.	German DAX Index	^GDAXI	Comprises 40 major companies listed on the Frankfurt Stock Exchange, serving as a key indicator of Germany's economic and corporate performance.
Hang Seng Index	^HSI	Tracks 50 leading Hong Kong companies, highlighting the economic connection between China and Hong Kong.	NASDAQ 100 Index	^NDX	Comprises 100 major U.S. tech and non-financial companies, serving as a key indicator of global tech stocks.
Nikkei 225 Index	^N225	Covers 225 top Japanese companies, reflecting Japan's economic trends and Asia's broader market dynamics.	NASDAQ Composite Index	^IXIC	Includes all companies listed on NASDAQ, capturing global tech stock trends and investor sentiment.

B. Stock Closing Price Sequence Decomposition Based on CEEMDAN

In this study, the S&P 500 index (^GSPC) was selected as a representative example for detailed analysis, while the procedures for the other nine indices were omitted due to space constraints. The ^GSPC dataset consists of 2,515 samples, with the first 80% used for training and the remaining 20% for testing to preserve the temporal structure during model evaluation.

The training set was processed using the CEEMDAN method to extract intrinsic mode functions (IMFs). Following the parameter settings proposed by Torres et al. (2011), the noise standard deviation was set to 0.2 (noise_std=0.2), the number of trials was 500, and a maximum of 2,000 sifting iterations (max_iter=2000) was allowed for each intrinsic mode function (IMF) extraction [23].

Fig. 4 presents the results of the CEEMDAN decomposition, showing six extracted intrinsic mode functions (IMFs) and one residual component (RES). This decomposition effectively separates the original time series into components with distinct frequency characteristics, capturing variations across different scales while preserving the underlying trend. Compared to the original signal, these components are smoother and more regular, enhancing feature representation and providing a robust foundation for subsequent modeling and forecasting.

Table III summarizes the key statistical characteristics of each IMF and the residual component, including their energy contribution rates and Pearson correlation coefficients with the original time series. These metrics reveal the impact of different frequency components on the time series and support their classification into three distinct frequency groups for subsequent modeling and forecasting.

TABLE III. STATISTICAL CHARACTERISTICS OF CEEMDAN-DECOMPOSED ^GSPC COMPONENTS

Component	Sample Size	Mean	Energy Contribution Rate	Correlation Coefficient with the Original Sequence	Average Difference (95% Confidence Interval)
IMF1	2515	0.2533	0.32%	0.0282	[-0.4497, 0.9290]
IMF2	2515	0.2452	0.39%	0.0150	[-0.4715, 1.7823]
IMF3	2515	0.7810	0.71%	0.0733	[-0.2590, 3.5593]
IMF4	2515	1.0539	2.89%	0.0727	[-0.3335, 3.0623]
IMF5	2515	-5.5893	6.99%	0.0435	[-13.4650, 9.6706]
IMF6	2515	-4.9123	88.70%	0.1788	[-13.4649, 9.6706]
RES	2515	3288.74	-	0.9542	[3243.0319, 3322.9385]

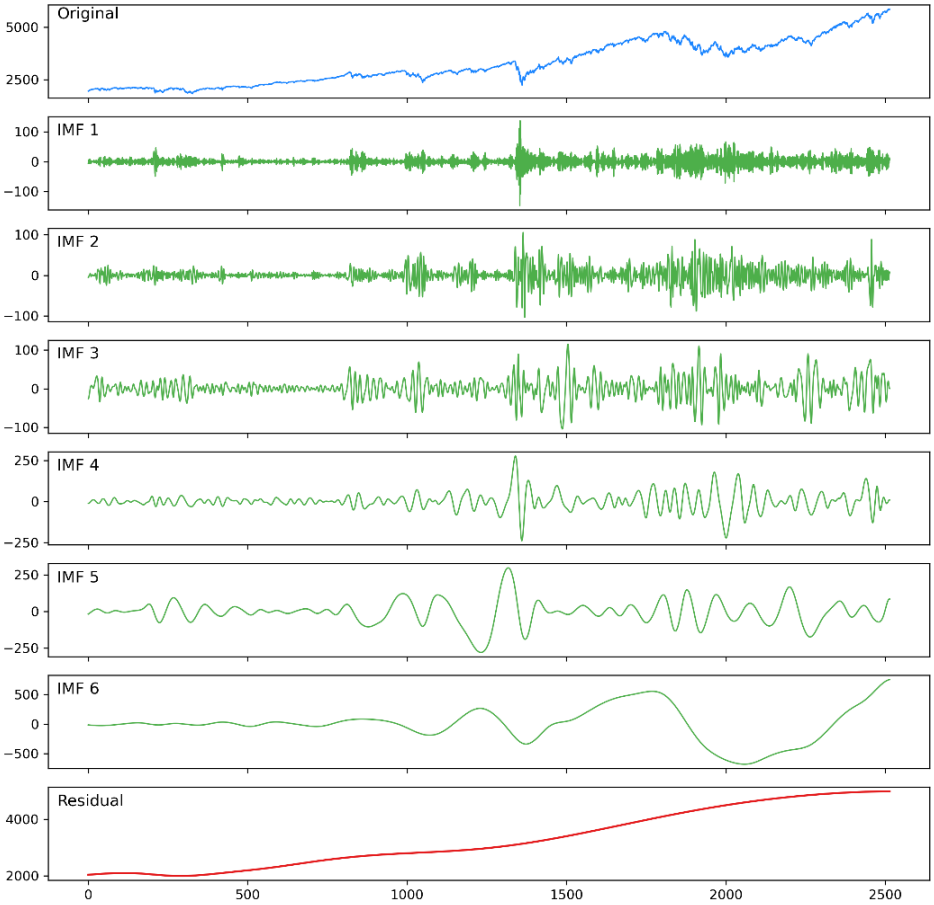


Fig. 4. CEEMDAN decomposition of closing prices for the ^GSPC index.

Based on the statistical analysis in Table III, the IMFs were categorized according to their energy contribution rates and correlation coefficients.

IMF 1 to IMF 3 primarily represent high-frequency components, characterized by low energy contribution rates (0.32%, 0.39%, and 0.71%, respectively) and low Pearson correlation coefficients with the original sequence (0.0282, 0.0150, and 0.0733, respectively). These components primarily capture short-term stochastic fluctuations, which may be influenced by sudden market events, speculative trading, or short-lived policy shifts. IMF 3 exhibits slightly higher explanatory power for short-term volatility.

IMF 4 and IMF 5 correspond to medium-frequency components, with energy contribution rates of 2.89% and 6.99% and correlation coefficients of 0.0727 and 0.0435, respectively. These components reflect medium-term market adjustments and cyclical dynamics, with IMF 5, in particular, being associated with medium-to-long-term fluctuations influenced by economic cycles and technical market corrections.

IMF 6, the dominant low-frequency component, accounts for the highest energy contribution (88.70%) and exhibits a correlation coefficient of 0.1788 with the original sequence. It primarily represents long-term market trends driven by macroeconomic factors and sustained investor sentiment.

The residual component (RES) represents the long-term trend, exhibiting a mean value of 3288.74 and a high correlation coefficient of 0.9542 with the original sequence. This indicates that the RES fully encapsulates the overall market trend.

Applying energy contribution thresholds of 2% for high-frequency components and 60% for low-frequency components effectively distinguishes short-term fluctuations from long-term trends. This separation facilitates predictive modeling by isolating different cyclical patterns in the data. The frequency band allocation of the ^GSPC index is illustrated in Fig. 5.

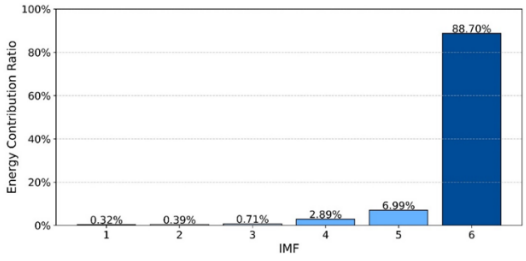


Fig. 5. Frequency band allocation results of IMF components for ^GSPC.

In order to enable the proposed model to effectively capture the unique fluctuation patterns and dynamic characteristics of each frequency band, a specific time window is assigned to the three different frequency components in this study. The time windows for the high-frequency, medium-frequency, and low-frequency components are set to 5 days, 10 days, and 20 days, respectively. The residual component, which primarily represents the long-term trend, is assigned a time window of 30 days. This configuration allows the model to fully learn and represent the overall long-term market movements and trend characteristics.

By adjusting the time windows based on the frequency characteristics of the components, the model is better equipped to analyze and predict market behavior across different time horizons.

C. Analysis of Experimental Results

Since the process of constructing the prediction model for the S&P 500 index (^GSPC) is identical to that of other stock indices, it is not discussed in detail here due to space limitations. To compare the effectiveness of the CEEMDAN decomposition method, the original closing price sequence was also decomposed using the EMD method for comparison. Taking the ^GSPC index as an example, the EMD method decomposed the sequence into 8 IMFs and RES, resulting in a total of 9 components. The specific decomposition results are presented in Fig. 6. The original ^GSPC closing price sequence is displayed at the top, followed by the IMF components arranged from high to low frequency, with the trend component at the bottom. The construction of the prediction model based on the EMD decomposition follows the same methodology as the CEEMDAN-BO-LSTM approach proposed in this study.

In order to further verify the effectiveness of the decomposition technique, the Bayesian optimized LSTM (BO-LSTM) model and the traditional LSTM model are involved in the comparative analysis in this study. The LSTM models constructed in this study all adopt a double-layer structure, and the specific parameter configurations are shown in Table IV.

TABLE IV. OPTIMIZED HYPERPARAMETERS AND THEIR SEARCH RANGES

Hyperparameters	Search Range
lstm_units1	[50, 200]
lstm_units2	[50, 150]
learning_rate	[0.0001, 0.01]
batch_size	[16, 64]

After completing the above preparations, the ^GSPC index was first trained and then predicted. To comprehensively evaluate the prediction performance, root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination (R^2) are utilized. The optimal parameter configuration for the double-layer LSTM model is determined through Bayesian optimization. The final settings include a batch size of 31, a learning rate of 0.005, 115 units in the first LSTM layer, and 79 units in the second LSTM layer. Table V summarizes the predictive performance comparison for the ^GSPC index across different models.

According to the prediction results shown in Table V, it can be observed that the CEEMDAN-BO-LSTM model outperforms all other models. The RMSE value of this model is 32.6788, the MAE value is 27.9756, the MAPE value is 0.5910%, and the R^2 value is 0.9969, which strongly demonstrates its superior predictive accuracy and robust fitting capability. In contrast, the RMSE values of the CEEMDAN-LSTM and EMD-BO-LSTM models are 41.4031 and 40.9361, respectively, performing better than the BO-LSTM and traditional LSTM models but still falling short compared to the CEEMDAN-BO-LSTM model.

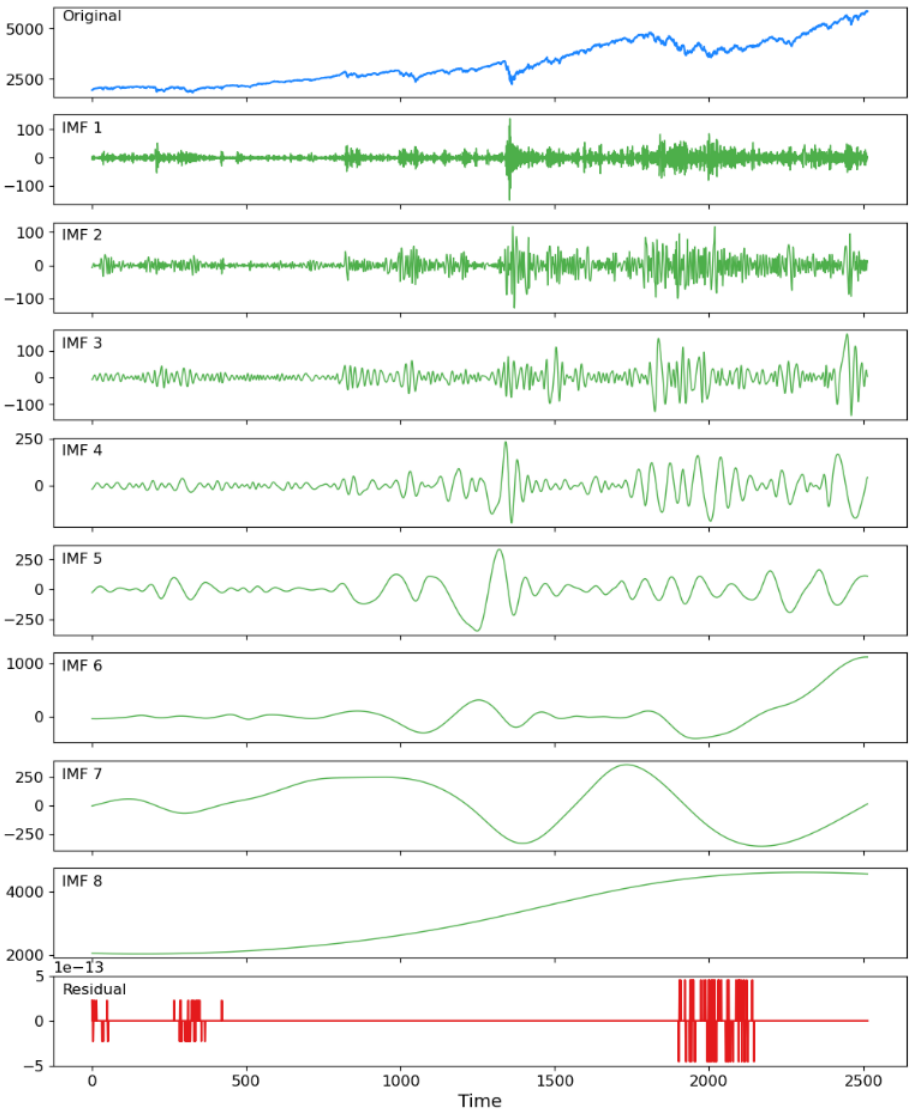


Fig. 6. EMD Decomposition of Closing Prices for the ^GSPC Index.

TABLE V. PERFORMANCE COMPARISON OF DIFFERENT LSTM-BASED MODELS FOR ^GSPC PREDICTION

Metric \ Model	CEEMDAN-BO-LSTM	CEEMDAN-LSTM	EMD-BO-LSTM	EMD-LSTM	BO-LSTM	LSTM
RMSE	32.6788	41.4031	40.9361	156.1560	56.6349	75.2508
MAE	27.9756	32.6506	29.1307	102.8529	45.1560	60.4651
MAPE	0.5910%	0.6729%	0.5991%	1.9808%	0.9487%	1.2603%
R^2	0.9969	0.9951	0.9951	0.9290	0.9909	0.9840

Compared to Bayesian-optimized models, non-optimized LSTM-based models exhibit weaker predictive performance and limited capability in capturing market trends. For example, the EMD-LSTM model produces large prediction errors, but its predictive performance improves significantly after Bayesian optimization of core parameters.

Next, to further visualize and compare the predictive performance of different models, Fig. 7 illustrates the final prediction results, Fig. 8 presents the error distribution, and Fig. 9 depicts the model fitting performance.

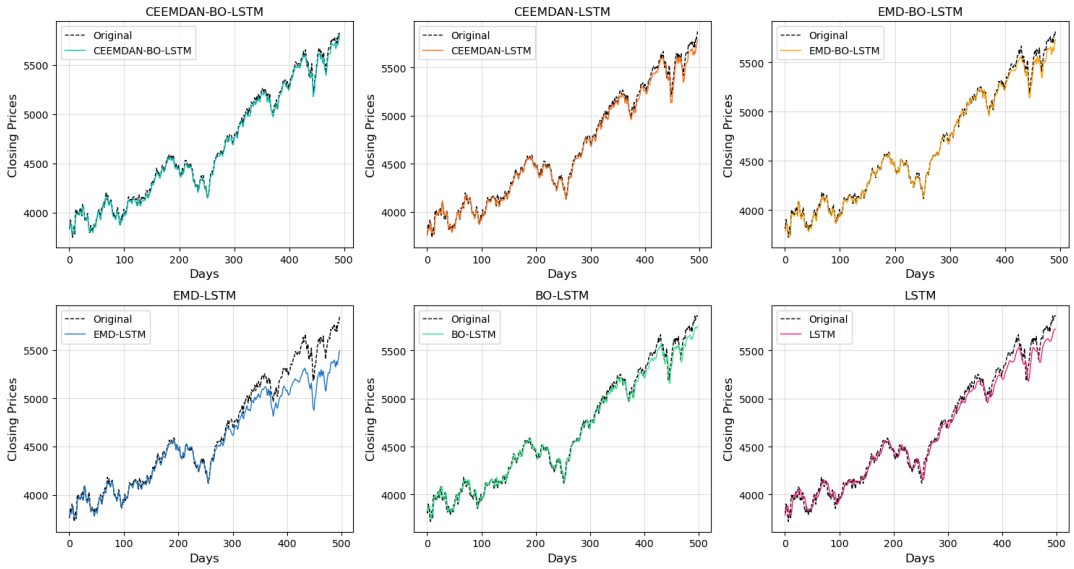


Fig. 7. Individual forecasts of different models on the ^GSPC index.

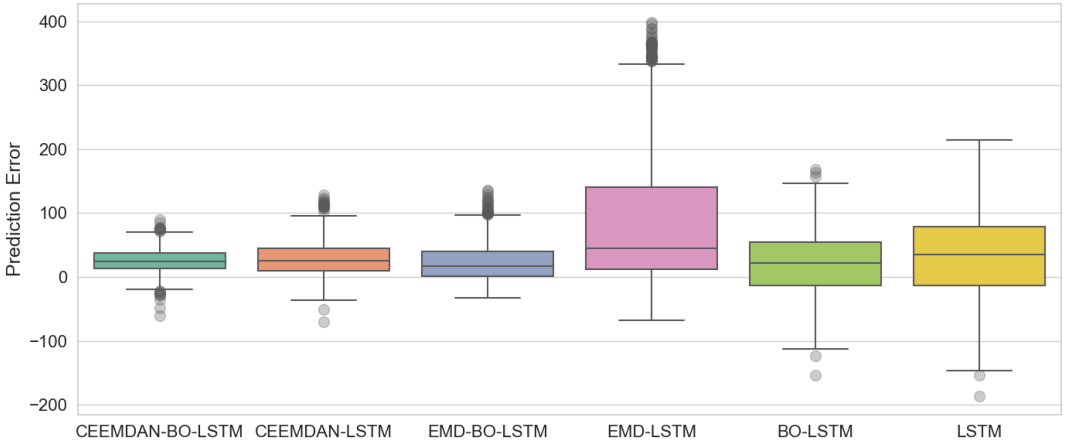


Fig. 8. Comparison of prediction errors for the ^GSPC index across different models.

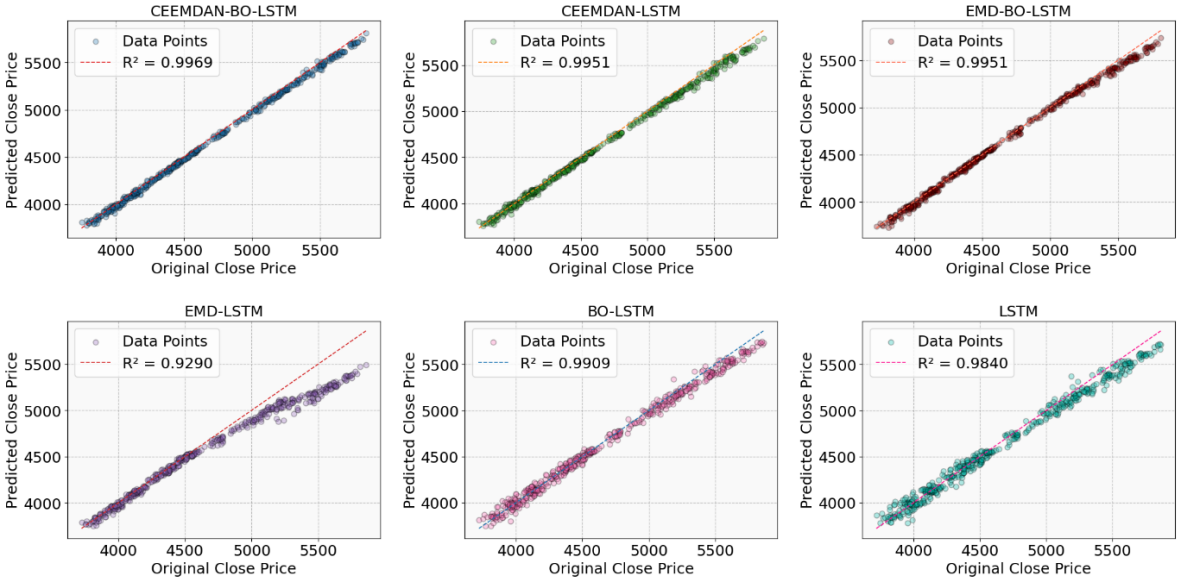


Fig. 9. Fitting performance of different models for the ^GSPC index predictions.

As illustrated in Fig. 7, it is evident that the CEEMDAN-BO-LSTM model exhibits the best fitting performance among all models, with its prediction curve closely aligning with the actual values. The model is effective in capturing the dynamic trends in different frequency bands, particularly around trend shifts and extreme points, demonstrating strong adaptability to complex market dynamics. In contrast, while the CEEMDAN-LSTM and EMD-BO-LSTM models also fit the overall trend well, they exhibit slight phase shifts during periods of severe market volatility, especially around local peaks and troughs. On the other hand, the fitting performance of the EMD-LSTM, BO-LSTM, and traditional LSTM models is relatively weak. Among them, both the EMD-LSTM model and the traditional LSTM model exhibit poor fitting performance in the second half of the prediction period. However, their predictive accuracy improves significantly after Bayesian optimization.

As shown in the box plots of error distributions in Fig. 8, the CEEMDAN-BO-LSTM model exhibits a narrower interquartile range (IQR) and shorter whiskers, indicating lower prediction uncertainty and greater stability. In contrast, the CEEMDAN-LSTM and EMD-BO-LSTM models have slightly wider error distributions, though their overall volatility remains moderate. On the other hand, the error box plots of the EMD-LSTM, BO-LSTM, and traditional LSTM models are significantly larger, reflecting more volatile prediction errors and lower stability.

Furthermore, the superiority of the CEEMDAN-BO-LSTM model is further demonstrated in the fitting performance plot shown in Fig. 9. The predicted points align closely with the ideal fitting line, achieving a high R^2 value of 0.9969. In contrast, the prediction points of the EMD-LSTM model and the traditional LSTM model are more dispersed, with R^2 values of only 0.9290 and 0.9840, respectively, reflecting their limitations in capturing short-term fluctuations and long-term trends. Overall, the CEEMDAN-BO-LSTM model significantly outperforms the other models in terms of fitting accuracy, prediction error, and stability.

To further validate the robustness and reliability of the proposed model, the same methodologies are applied to predict nine additional stock indices. The prediction results are summarized in Table VI.

TABLE VI. COMPARISON OF PREDICTION RESULTS ACROSS VARIOUS MODELS FOR NINE ADDITIONAL STOCK INDICES

Model / Metric	^DJI				^FTSE				^HSI			
	RMSE	***MAE***	***MAPE***	***R^2***	***RMSE***	***MAE***	***MAPE***	***R^2***	***RMSE***	***MAE***	***MAPE***	***R^2***
CEEMDAN-BO-LSTM	142.0854	112.3485	0.31%	0.9977	23.7777	18.5461	0.24%	0.9946	147.3649	111.4311	0.61%	0.9918
CEEMDAN -LSTM	259.5091	223.6002	0.61%	0.9923	42.1390	33.4897	0.43%	0.9834	200.1531	154.0437	0.86%	0.9850
EMD-BO-LSTM	231.8059	188.7823	0.54%	0.9937	25.2575	19.2848	0.25%	0.9939	158.3405	120.0416	0.66%	0.9906
EMD-LSTM	219.2336	180.4671	0.50%	0.9945	39.5812	31.2914	0.40%	0.9854	326.7038	277.6644	1.55%	0.9602
BO-LSTM	487.2037	385.5664	1.03%	0.9739	98.7783	81.2571	1.03%	0.9088	332.9701	244.7510	1.34%	0.9589
LSTM	450.2276	365.7604	1.00%	0.9777	98.6217	75.9350	0.96%	0.9091	403.8065	308.4712	1.68%	0.9395
Model / Metric	**^N225**				**000001.SS**				**^STOXX50E**			
	RMSE	***MAE***	***MAPE***	***R^2***	***RMSE***	***MAE***	***MAPE***	***R^2***	***RMSE***	***MAE***	***MAPE***	***R^2***
CEEMDAN-BO-LSTM	285.8826	184.8111	0.55%	0.9960	18.9642	14.1334	0.46%	0.9845	34.3053	27.6026	0.60%	0.9913
CEEMDAN-LSTM	408.1841	281.7839	0.81%	0.9918	39.8539	32.4014	1.07%	0.9315	51.6794	42.9811	0.93%	0.9810
EMD-BO-LSTM	408.4615	294.1020	0.84%	0.9918	24.8386	19.3618	0.64%	0.9735	39.7511	34.2836	0.75%	0.9886
EMD-LSTM	541.3611	437.5387	1.28%	0.9856	34.3530	27.1668	0.90%	0.9491	66.2227	57.1415	1.25%	0.9692
BO-LSTM	1096.3721	844.6950	2.34%	0.9412	31.8605	21.2103	0.69%	0.9563	68.8182	54.2253	1.17%	0.9661
LSTM	1471.2706	1087.955	2.97%	0.8941	40.7106	29.9967	0.96%	0.9286	115.5603	97.0145	2.09%	0.9044
Model / Metric	**^GDAXI**				**^NDX**				**^IXIC**			
	RMSE	***MAE***	***MAPE***	***R^2***	***RMSE***	***MAE***	***MAPE***	***R^2***	***RMSE***	***MAE***	***MAPE***	***R^2***
CEEMDAN-BO-LSTM	65.9233	51.0301	0.31%	0.9980	124.4471	99.8485	0.68%	0.9980	84.5104	66.0946	0.48%	0.9987
CEEMDAN-LSTM	172.7391	151.907	0.90%	0.9866	132.9840	105.0143	0.71%	0.9978	123.8861	101.1588	0.73%	0.9972
EMD-BO-LSTM	203.5323	168.399	0.99%	0.9814	196.8166	152.8102	0.95%	0.9950	140.0372	97.1818	0.65%	0.9963
EMD-LSTM	298.8572	243.266	1.41%	0.9603	293.6994	216.5402	1.30%	0.9927	199.6780	158.5990	1.09%	0.9926
BO-LSTM	227.5541	174.661	1.03%	0.9770	231.6684	185.2062	1.18%	0.9933	200.0174	154.5512	1.12%	0.9927
LSTM	172.7391	151.907	0.90%	0.9866	271.6015	217.8445	1.42%	0.9907	321.7919	259.8644	1.77%	0.9812

As presented in Table VI, the CEEMDAN-BO-LSTM model consistently outperforms the other models across the nine stock indices. It achieves the best results for all indices. For example, on the ^DJI, ^FTSE, and ^NDX indices, their RMSE values are 142.0854, 23.7777, and 124.4471, respectively, which are significantly lower than those of other models. Meanwhile, their MAPE values are 0.31%, 0.24%, and 0.68%, respectively, and they have the lowest error rates compared to other models. This indicates that the model consistently maintains high forecasting accuracy and low error under different market conditions.

Although the EMD-BO-LSTM model fails to achieve the highest overall accuracy, it performs well on specific stock indices (such as 000001.SS, ^FTSE, and ^HSI) with lower MAE and RMSE values than other models. This suggests that the application of the LSTM model combining EMD decomposition with Bayesian optimization in stock index prediction is effective and improves the prediction performance to some extent. Nevertheless, its performance remains inferior to that of the CEEMDAN-BO-LSTM model proposed in this paper. The results further indicate that EMD still has some limitations in terms of effectiveness and stability when compared with the CEEMDAN algorithm.

Furthermore, the superiority of the CEEMDAN-BO-LSTM model is further demonstrated in the fitting performance plot shown in Fig. 9. The predicted points align closely with the ideal fitting line, achieving a high R^2 value of 0.9969. In contrast, the prediction points of the EMD-LSTM model and the traditional LSTM model are more dispersed, with R^2 values of only 0.9290 and 0.9840, respectively, reflecting their limitations in capturing short-term fluctuations and long-term trends. Overall, the CEEMDAN-BO-LSTM model significantly outperforms the other models in terms of fitting accuracy, prediction error, and stability.

To further validate the robustness and reliability of the proposed model, the same methodologies are applied to predict nine additional stock indices. The prediction results are summarized in Table VI.

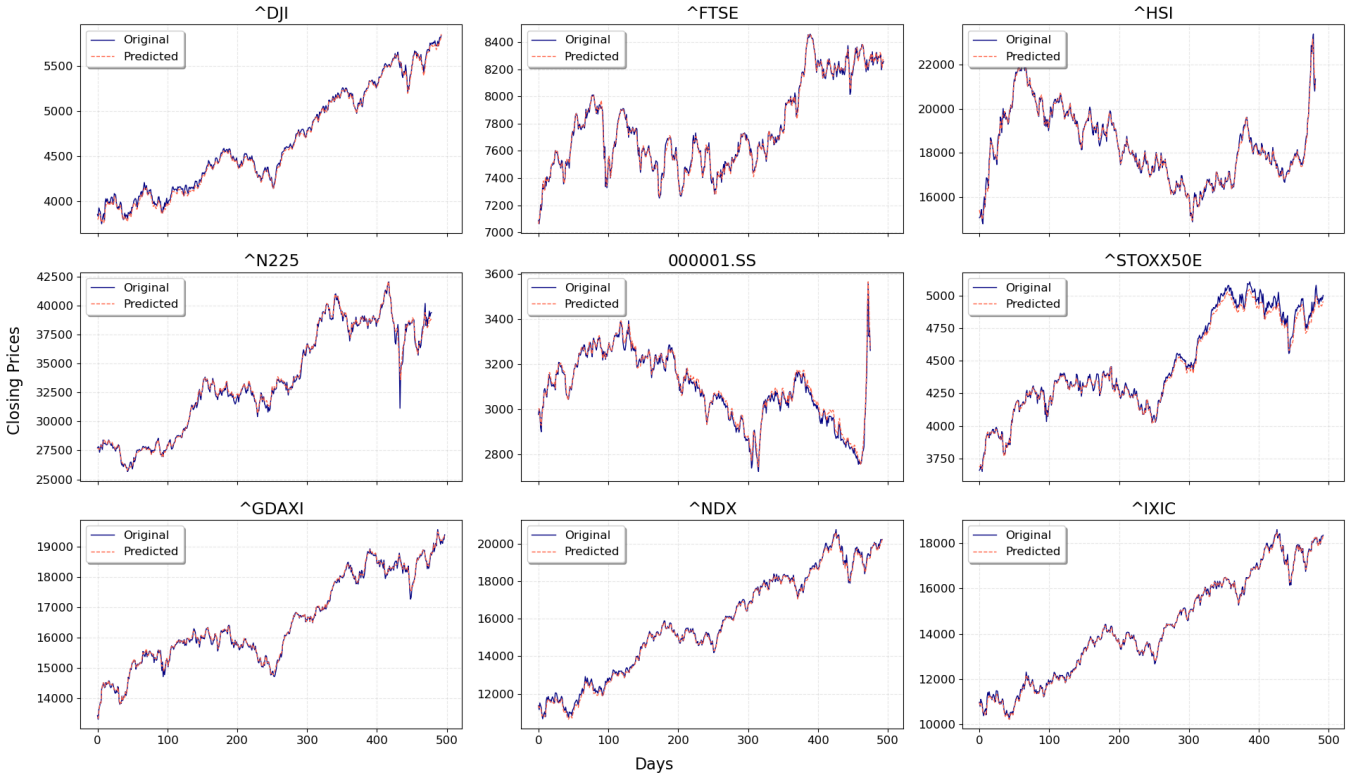


Fig. 10. Fitting performance of different models.

As shown in Fig. 10, the CEEMDAN-BO-LSTM model demonstrates strong predictive capability for the closing prices of the nine stock indices, with the predicted values closely aligning with the actual values. However, in the later part of the ^N225 test set, a sharp decline in stock price occurs, and the model underestimates the stock price decline at this point, although the model has incorporated relevant information.

Similarly, in the latter part of the ^STOXX50E test set, the closing price exhibits significant fluctuations, and the model's predictions during this period also show some deviations. These deviations suggest that extreme market volatility poses a significant challenge to predictive modeling.

Nevertheless, overall, the proposed model has demonstrated strong predictive performance, effectively adapting to complex market dynamics and providing accurate predictions in most scenarios.

V. Conclusion

In this study, a novel CEEMDAN-BO-LSTM stock market prediction model is proposed. The model follows a

decomposition-optimization-prediction-integration framework to forecast ten representative stock indices worldwide. It demonstrates excellent performance across all evaluation metrics, with an R^2 value close to 1, indicating outstanding predictive accuracy.

The experimental results show that the CEEMDAN decomposition method can improve the accuracy and stability of the decomposition, which significantly enhances the ability of the model proposed in this study to capture stock market dynamics. Meanwhile, through Bayesian optimization, the hyperparameters are precisely tuned, which not only mitigates the overfitting problem but also enhances the model's training efficiency. In addition, the model leverages the memory capability of the double-layer LSTM network for predictions, and the sliding window is optimally configured based on different frequency bands, further improving predictive accuracy.

Despite the high predictive accuracy and stability of the CEEMDAN-BO-LSTM model, certain limitations remain. Firstly, the CEEMDAN decomposition process increases computational complexity. Although the frequency band setting categorizes the data into low-frequency, medium-frequency, high-frequency, and trend components, each component requires separate model training, leading to prolonged training time. Additionally, the model's predictive performance is highly dependent on the optimization of LSTM hyperparameters, making the appropriate selection of hyperparameter search ranges a critical issue.

Future research will integrate ensemble learning techniques with deep learning models, combining the high predictive accuracy of ensemble methods with the memory capability of deep learning networks. Adaptive learning mechanisms will also be explored to improve the effectiveness of model integration and adaptation. Furthermore, external factors such as social sentiment and policy changes will be considered. Text analysis will be incorporated to enable the financial time series model to learn from socio-economic developments, enhancing the effectiveness and reliability of predictions.

ACKNOWLEDGMENT

This research is funded by Guangdong Province Key Discipline Research Capacity Enhancement Project (No. 2022ZDJS146); Key Natural Science Project of Guangdong University of Science and Technology (No. GKY-2023KYZDK-16).

REFERENCES

[1] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, "Stock price prediction using the ARIMA model," in Proceedings of the 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, IEEE, pp. 106–112, 2014.

[2] H. Herwartz, "Stock return prediction under GARCH—An empirical assessment," International Journal of Forecasting, vol. 33, no. 3, pp. 569–580, 2017.

[3] A. L. S. Maia, F. de A. T. de Carvalho, "Holt's exponential smoothing and neural network models for forecasting interval-valued time series," Int. J. Forecasting, vol. 27, no. 3, pp. 740–759, 2011.

[4] I. M. Yassin, M. F. A. Khalid, S. H. Herman, et al., "Multi-layer perceptron (MLP)-based nonlinear auto-regressive with exogenous inputs (NARX) stock forecasting model," International Journal of Advanced Science Engineering and Information Technology, vol. 7, no. 3, pp. 1098–1103, 2017.

[5] D. Zhang and S. Lou, "The application research of neural network and BP algorithm in stock price pattern classification and prediction," Future Generation Computer Systems, vol. 115, pp. 872–879, 2021.

[6] S. Hochreiter, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] K. Pawar, R. S. Jalem, and V. Tiwari, "Stock market price prediction using LSTM RNN," in Emerging Trends in Expert Applications and Security: Proceedings of ICETEAS 2018, Springer Singapore, pp. 493–503, 2019.

[8] J. Wang, S. Hong, Y. Dong, et al., "Predicting stock market trends using LSTM networks: overcoming RNN limitations for improved financial forecasting," Journal of Computer Science and Software Applications, vol. 4, no. 3, pp. 1–7, 2024.

[9] H. Mehtarizadeh, N. Mansouri, B. M. H. Zade, and M. M. Hosseini, "Stock price prediction with SCA-LSTM network and Statistical model ARIMA-GARCH," The Journal of Supercomputing, vol. 81, no. 2, p. 366, 2025.

[10] H. Baek, "A CNN-LSTM stock prediction model based on genetic algorithm optimization," Asia-Pacific Financial Markets, vol. 31, no. 2, pp. 205–220, 2024.

[11] S. Sang and L. Li, "A novel variant of LSTM stock prediction method incorporating attention mechanism," Mathematics, vol. 12, no. 7, p. 945, 2024.

[12] H. Zheng, J. Wu, R. Song, L. Guo, and Z. Xu, "Predicting financial enterprise stocks and economic data trends using machine learning time series analysis," Appl. Comput. Eng., vol. 87, pp. 26–32, 2024.

[13] Z. D. Akşehir and E. Kılıç, "A new denoising approach based on mode decomposition applied to the stock market time series: 2LE-CEEMDAN," PeerJ Computer Science, vol. 10, p. e1852, 2024.

[14] B. Gülmez, "Stock price prediction with optimized deep LSTM network with artificial rabbits optimization algorithm," Expert Systems with Applications, vol. 227, p. 120346, 2023.

[15] L. Tian, L. Feng, L. Yang, and Y. Guo, "Stock price prediction based on LSTM and LightGBM hybrid model," The Journal of Supercomputing, vol. 78, no. 9, pp. 11768–11793, 2022.

[16] Y. He and K. F. Tsang, "Universities power energy management: A novel hybrid model based on iCEEMDAN and Bayesian optimized LSTM," Energy Reports, vol. 7, pp. 6473–6488, 2021.

[17] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in Advances in Neural Information Processing Systems, vol. 25, pp. 2951–2959, 2012.

[18] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," IEEE Transactions on Signal Processing, vol. 62, no. 3, pp. 531–544, 2013.

[19] H. Niu, K. Xu, and W. Wang, "A hybrid stock price index forecasting model based on variational mode decomposition and LSTM network," Applied Intelligence, vol. 50, pp. 4296–4309, 2020.

[20] K. Su, C. Zheng, and X. Yu, "Portfolio allocation with CEEMDAN denoising algorithm," Soft Computing, vol. 27, no. 21, pp. 15955–15970, 2023.

[21] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, vol. 454, pp. 903–995, 1998.

[22] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," Advances in Adaptive Data Analysis, vol. 1, no. 1, pp. 1–41, 2009.

[23] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, "A complete ensemble empirical mode decomposition with adaptive noise," in Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 4144–4147, 2011.

[24] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas, "Taking the human out of the loop: A review of Bayesian optimization," Proceedings of the IEEE, vol. 104, no. 1, pp. 148–175, 2015

Improving Performance with Big Data: Smart Supply Chain and Market Orientation in SMEs

Miftakul Huda[1], Agus Rahayu[2], Chairul Furqon[3], Mokh Adib Sultan[4], Nani Hartati[5], Neng Susi Susilawati Sugiana[6]
Faculty of Economic and Business Education, Universitas Pendidikan Indonesia[1, 2, 3, 4]
Faculty of Economic and Business, Universitas Pelita Bangsa, Indonesia[1, 5]
Information System Engineering, Institut Digital Ekonomi LPKIA, Bandung, Indonesia[6]

***Abstract*—This study aims to explore the impact of big data-driven supply chain management, web analytics, and market orientation on corporate performance in medium-sized enterprises (MSEs) in Indonesia. By integrating these contemporary elements, the research seeks to provide insights into how digital technologies and strategic market practices can enhance organizational effectiveness. The study adopts a quantitative approach, utilizing survey data collected from 350 MSEs across various sectors in Indonesia. Purposive sampling was employed to ensure that the selected firms actively implement big data analytics and market-oriented strategies. Structural Equation Modeling (SEM) was conducted using SmartPLS to analyze the relationships among the variables. The findings reveal that big data-driven supply chain management and web analytics significantly contribute to improved corporate performance, with market orientation serving as a critical mediating factor. These results emphasize the importance of aligning digital tools with strategic business objectives to achieve competitive advantages. Furthermore, the study highlights the practical implications for MSEs, suggesting that integrating big data and web analytics into supply chain operations can optimize resource allocation, enhance decision-making, and foster market responsiveness. This research contributes to the literature on digital transformation and strategic management in emerging economies, offering a novel perspective on how MSEs can leverage technological advancements to remain competitive. Future studies may explore longitudinal impacts and sector-specific adaptations.**

Keywords*—Big data; supply chain management; web analytics; corporate performance; market orientation

I. Introduction

In the era of rapid technological advancements, the integration of digital tools into business operations has become a critical driver of organizational performance. Medium-sized enterprises (MSEs), which represent a significant portion of Indonesia's economy, face increasing challenges in maintaining competitiveness in a dynamic market landscape. These challenges are exacerbated by the need to manage complex supply chain operations, harness insights from vast datasets, and adapt to rapidly shifting consumer behaviors, despite their critical role in economic growth, many Indonesian MSEs struggle with resource limitations, technological adoption barriers, and strategic misalignments, which hinder their ability to achieve optimal performance.

Data from the Indonesia, contributing approximately 30% to the national GDP and employing millions of workers, has a significant Micro, Small, and Medium Enterprises (MSEs) sector. However, only about 15% of MSEs utilize big data analytics or advanced digital tools in their operations. The underutilization of technologies like web analytics and big data in supply chain management reflects a missed opportunity for MSEs to improve operational efficiency and market orientation [1]. The proposed approach aims to address this issue by integrating digital technologies, specifically big data analytics and web analytics, into the supply chain management of Indonesian MSEs. By adopting these technologies, MSEs can enhance operational efficiency, more accurately forecast demand, optimize inventories, and improve supplier relationships. Furthermore, web analytics enables MSEs to better understand consumer behavior, evaluate marketing strategies, and increase customer engagement.

The potential benefits of this approach are substantial, especially given the increasing global competition, particularly with the influx of multinational corporations. MSEs that integrate technology with market-driven business practices will have a greater chance of survival and growth [2]. Therefore, the adoption of these technologies will not only enhance the competitiveness of MSEs but also strengthen their contribution to the national economy. The main contribution of this research is the identification and application of accessible digital technologies for MSEs in Indonesia to overcome limitations in supply chain management. This study provides insights into how big data analytics and web analytics can be integrated into MSEs' business processes to improve decision-making, optimize supply chain management, and enhance customer satisfaction.

The implications of implementing this approach are broader digital transformation within the MSE sector. By adopting these technologies, MSEs can strengthen their resilience and competitiveness in an increasingly globalized and competitive market. Furthermore, improved operational efficiency and a better understanding of market demands can have a positive impact on national economic growth. However, challenges to be addressed include perceptions of high costs, lack of expertise in implementing these technologies, and uncertainty regarding the impact of these technologies on business performance [3]. Therefore, the proposed solution should also include training, enhancing digital literacy, and developing affordable implementation strategies for MSEs.

This approach has the potential to transform the business landscape for Indonesian MSEs, bringing them into a more efficient and competitive digital era. The core issue faced by Indonesian MSEs lies in their limited capacity to leverage digital

tools effectively. Supply chain management, a cornerstone of operational success, often lacks the sophistication required to address the complexities of modern markets [4]. Big data analytics, when integrated into supply chain processes, can provide valuable insights into demand forecasting, inventory optimization, and supplier relationships [5]. Similarly, web analytics offers businesses the ability to monitor consumer behavior, evaluate marketing strategies, and enhance customer engagement [6]. Despite the evident benefits, a significant proportion of MSEs remain hesitant to adopt these technologies due to perceived costs, lack of expertise, and uncertainty about their impact on performance [7].

Previous studies have highlighted the positive effects of big data and web analytics on corporate performance. For instance, demonstrated that data-driven supply chain management significantly improves operational efficiency and customer satisfaction [8]. Found that web analytics enhances firms' ability to respond to market dynamics, thereby fostering competitive advantages [9]. However, these studies primarily focus on large enterprises in developed economies, leaving a critical research gap concerning the applicability of these findings to MSEs in emerging markets like Indonesia. Additionally, market orientation has been identified as a key mediator in achieving superior corporate performance. Argue that organizations with a strong market orientation are better equipped to anticipate customer needs [10], adapt to environmental changes, and achieve long-term success [11]. While the relationship between market orientation and performance is well-documented, its interaction with digital tools such as big data and web analytics in the context of MSEs remains underexplored [12].

A. Research Questions

To address these gaps, this study seeks to answer the following research questions:

1) How does big data-driven supply chain management influence corporate performance in Indonesian MSEs?

2) What is the role of web analytics in enhancing the market orientation of Indonesian MSEs?

3) How does market orientation mediate the relationship between digital tools (big data and web analytics) and corporate performance in Indonesian MSEs?

II. Literatur Review

1) Big data-driven supply chain: The adoption of Big Data in supply chain management has transformed how businesses operate by improving efficiency, demand forecasting, and risk management. Define Big Data-Driven Supply Chain as the integration of large-scale data to support strategic decision-making in supply chain processes [13]. Emphasize that utilizing Big Data in supply chains enhances agility and responsiveness, enabling companies to adapt swiftly to market changes[4]. Moreover, to demonstrate that Big Data significantly boosts operational efficiency, particularly for medium-sized enterprises (MSEs). Key dimensions of this approach include real-time data analysis, trend forecasting, and end-to-end supply chain visibility [14].

2) Smart supply chain: A next-generation evolution: The concept of a Smart Supply Chain builds upon the foundation laid by Big Data-Driven Supply Chains, incorporating advanced technologies such as the Internet of Things (IoT)[15], Artificial Intelligence (AI), and Blockchain[16]. This evolution enables end-to-end automation, enhanced connectivity, and seamless integration across supply chain processes.

A Smart Supply Chain leverages interconnected systems to optimize logistics, inventory management, and production scheduling. According a Smart Supply Chain is characterized by its ability to self-monitor, self-analyze, and self-optimize, driven by real-time data and predictive analytics [14]. Key dimensions of a Smart Supply Chain include:

- IoT Integration: Sensors and connected devices provide real-time tracking and monitoring of goods across the supply chain.
- AI-Powered Insights: AI enables predictive maintenance, demand forecasting, and anomaly detection.
- Blockchain Transparency: Distributed ledger technology ensures secure, tamper-proof transaction records, fostering trust and traceability.

Smart Supply Chains enhance operational efficiency by reducing waste, minimizing downtime, and improving collaboration among stakeholders. For example, companies using IoT devices report a 25% reduction in logistics costs [17]. Additionally, AI-powered decision-making ensures faster response times to market changes, enabling businesses to maintain a competitive edge.

However, implementing Big Data solutions comes with challenges. High implementation costs and the need for sophisticated data integration pose significant barriers, especially for resource-constrained MSEs [18]. Despite these challenges, businesses that successfully leverage Big Data in their supply chains can achieve enhanced transparency and reduced operational risks, providing a competitive edge in a dynamic market environment [19].

B. Web Analytics

Web Analytics plays a crucial role in helping businesses understand consumer behavior in digital spaces. Define Web Analytics as the analysis of website data to identify consumer behavior patterns and enhance the effectiveness of digital marketing strategies [12]. For MSEs, Web Analytics offers actionable insights into customer preferences, campaign performance, and conversion rates, allowing them to optimize their marketing efforts effectively [20].

Empirical studies highlight the benefits of Web Analytics for business performance. Found that companies utilizing Web Analytics strategically experienced improved marketing campaign outcomes. Similarly [21], Reported a 15% increase in sales among businesses that implemented robust web data analysis systems [22]. However, Web Analytics also presents challenges, such as requiring technical expertise and robust data infrastructure factors that can hinder adoption among smaller enterprises [23]. Despite these obstacles, Web Analytics remains a powerful tool for MSEs to gain deeper customer insights and make data-driven decisions.

C. Market Orientation

Market Orientation focuses on aligning business strategies with customer needs and market demands. Define Market Orientation as the process of collecting, disseminating, and utilizing market information to create value for customers [8]. This approach emphasizes customer focus, competitor orientation, and cross-functional coordination, which are essential for developing responsive and adaptive business strategies.

Research consistently shows a positive relationship between Market Orientation and corporate performance. Argue that Market Orientation enhances customer satisfaction and loyalty, leading to improved business outcomes [24]. Furthermore, Found that Market Orientation is particularly effective in dynamic markets, where understanding and anticipating customer needs are critical [25]. For MSEs in Indonesia, adopting a Market Orientation approach allows them to tailor their strategies to local market preferences and enhance their competitiveness.

However, one nofi drawback of Market Orientation is the risk of over-orientation toward customer needs, which may stifle long-term innovation and strategic vision. Despite this limitation, integrating Market Orientation into business practices enables MSEs to remain relevant and customer-focused in a rapidly evolving marketplace.

D. Synthesis

The integration of Big Data-Driven Supply Chain, Web Analytics, and Market Orientation creates a synergistic effect that enhances corporate performance for MSEs. Big Data improves operational efficiency, Web Analytics provides deep insights into customer behavior, and Market Orientation ensures alignment with market needs. This holistic approach is particularly relevant for MSEs in Indonesia, offering a pathway to sustainable growth and competitive advantage in a resource-constrained environment. By addressing the unique challenges and opportunities within MSEs, this study contributes to the growing body of literature on leveraging technology and market strategies for business success.

III. Methodology

This study employs a quantitative approach to analyze the challenges faced by MSEs in Indonesia regarding the adoption of digital technologies in supply chain management and market orientation. A quantitative approach is chosen because it allows for the collection of objective, measurable data and the systematic analysis of relationships between variables. Using purposive sampling, this study targets 200 MSEs that have adopted technologies such as Big Data, Web Analytics, and market orientation, providing deeper insights into the impact of digital technology implementation on operational efficiency and market competitiveness. This quantitative approach also enables the identification of patterns and trends within the population of MSEs under study. Table I, "Population and Sample" outlines the key characteristics of the population and sample used in this study, providing details on the target group, sample size, criteria, and sampling technique employed.

TABLE I. Population and Sample

Aspect	Details
Population	Medium-sized enterprises (MSEs) operating in Indonesia.
Sample Size	200 enterprises selected based on purposive sampling.
Sampling Criteria	Enterprises adopting Big Data, Web Analytics, and Market Orientation.
Sampling Technique	Purposive sampling, targeting MSEs that demonstrate digital adoption.

The study focuses on medium-sized enterprises (MSEs) in Indonesia that have adopted Big Data-Driven Supply Chain, Web Analytics, and Market Orientation practices. A purposive sampling technique was used to ensure the sample consisted of enterprises actively engaging in these strategies. The sample size of 200 enterprises is deemed sufficient to represent the target population and provide statistically reliable results. This table provides a comparison between the current study and similar previous research to highlight key differences and similarities. It allows for a clearer understanding of how this study contributes to existing literature, particularly in terms of methodology, target group, and findings. By contrasting different approaches, Table II emphasizes the unique aspects of this research in relation to prior work in the field studies:

TABLE II. Previous Study

Aspect	Current Study	Previous Study 1	Previous Study 2	Previous Study 3
Objective	Investigating the adoption of Big Data & Web Analytics in MSEs	Examining digital tools adoption in large enterprises	Exploring digital transformation in SMEs	Analyzing technology use in supply chain management
Target Group	Medium-sized enterprises (MSEs) in Indonesia	Large enterprises across various industries	Small and medium enterprises (SMEs) in the UK	SMEs in the retail sector in Europe
Methodology	Quantitative analysis using purposive sampling	Qualitative case studies	Mixed-methods approach	Quantitative survey-based approach
Sample Size	200 MSEs	50 large enterprises	150 SMEs	300 SMEs
Key Findings	Integration of Big Data and Web Analytics enhances efficiency	Technology adoption leads to increased productivity	Digital transformation improves operational flexibility	Technology adoption improves supply chain performance
Limitations	Focused on MSEs in Indonesia, may not be generalizable	Limited sample size, focuses on large enterprises	Limited geographic scope	Narrow focus on one industry

This table allows for a side-by-side comparison of study with similar previous works, highlighting the differences in objectives, methodologies, target groups, sample sizes, and key findings.

The SmartPLS analysis begins with the specification of both the measurement and structural models. The measurement model defines the relationships between observed variables (indicators) and latent constructs (e.g., Big Data-Driven Supply Chain, Web Analytics, Market Orientation), while the structural model outlines the hypothesized relationships between these constructs and their impact on Corporate Performance. After specifying the models, data is input and the model is estimated using the PLS algorithm, which calculates path coefficients to assess the strength and direction of the relationships between variables. The measurement model is then evaluated for convergent validity, reliability, and discriminant validity, ensuring the constructs are adequately represented and distinct from each other.

Next, the structural model is assessed by examining path coefficients, R^2 values, effect sizes (f^2), and predictive relevance (Q^2) to determine the strength and explanatory power of the relationships. Bootstrapping analysis is performed to test the statistical significance of the path coefficients, with t-values and p-values being calculated to confirm the relationships are meaningful. Once the model passes these evaluations, the results are interpreted to validate the hypotheses and understand the impact of Big Data, Web Analytics, and Market Orientation on Corporate Performance in medium-sized enterprises. SmartPLS provides a comprehensive and robust approach for analyzing complex relationships in this study.

IV. Result And Discussion

A. Results

The following figure, Fig. 1: Result Test Model and Hypothesis, illustrates the outcome of the model testing and the relationships between the proposed.

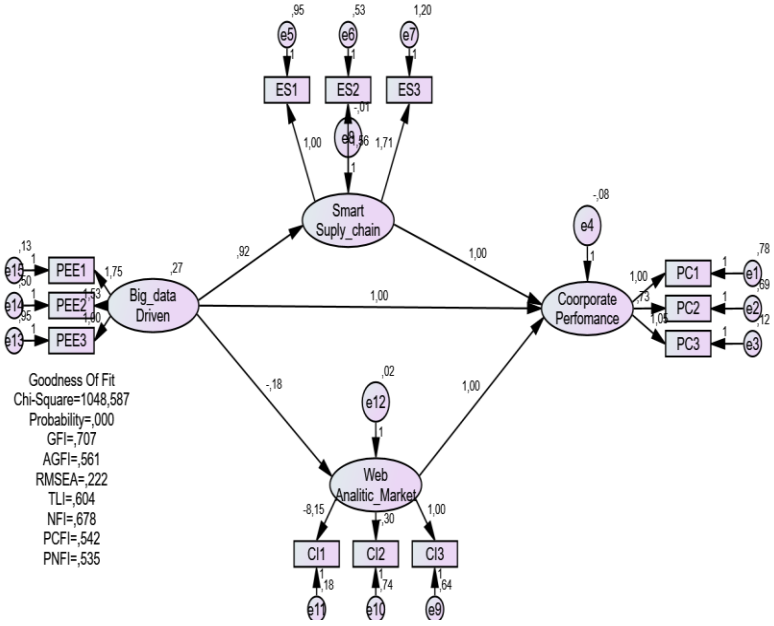


Fig. 1. Result test model and hypothesis.

B. Relationship Between Variables

1) Big Data-Driven → Corporate Performance

The path coefficient is 0.27, indicating a positive and significant influence, but the contribution is relatively small compared to other variables. This suggests that leveraging big data contributes to enhancing corporate performance but is not the primary driver.

2) Smart Supply Chain → Corporate Performance

The path coefficient is 0.92, indicating a very strong and significant positive relationship. This confirms that implementing a smart supply chain has a substantial impact on improving corporate performance.

3) Web Analytics Market → Corporate Performance

The path coefficient is -0.18, indicating a negative relationship. This implies that the web analytics strategies employed in this model do not yield the expected results and may even have an adverse impact on corporate performance.

C. Relationships Between Dimensions and Their Latent Variables:

1) Big data-driven: This variable is measured through indicators PEE1, PEE2, and PEE3, with PEE2 contributing the most significantly, having a factor loading of 1.53. These indicators reflect the level of big data integration and its effectiveness.

2) Smart supply chain: This variable demonstrates a strong relationship with its indicators ES1, ES2, and ES3, all of which have high and consistent factor loadings. It signifies that each aspect of the supply chain, such as efficiency, responsiveness, and integration, contributes positively to corporate performance.

3) Web analytics market: This variable shows a weak relationship, particularly with CI1, which has a low and negative factor loading (-8.15). This suggests that some components of the web analytics strategy may not align well with the goals of corporate performance.

4) Corporate performance: Measured through PC1, PC2, and PC3, all of which exhibit high and consistent factor loadings. This indicates that the corporate performance construct is well-represented by its indicators, which likely focus on financial, operational, and market outcomes.

The results highlight that Smart Supply Chain is the most significant contributor to Corporate Performance, followed by Big Data-Driven with a moderate influence. However, the negative effect of Web Analytics Market suggests misalignment or ineffective implementation in the current model. Further evaluation or refinement of web analytics strategies is necessary to improve their impact.

D. Discusssion

1) Big data-driven and corporate performance: The relationship between big data-driven strategies and corporate performance is positive but moderate, as indicated by the path coefficient of **0.27**. This finding aligns with previous studies, such as that by Wamba et al. (2017), which emphasized that big data analytics can enhance decision-making and operational efficiencies, leading to improved firm performance. However, the relatively low contribution in this study suggests that medium-sized enterprises in Indonesia may face challenges in fully utilizing big data technologies due to limitations in resources, infrastructure, or expertise. As noted by Akter et al. (2016), the success of big data initiatives often depends on organizational readiness, including skilled personnel and advanced technological capabilities.

To address these challenges, medium-sized enterprises should invest in training programs and partnerships with technology providers to build their capacity for big data analytics. Moreover, adopting scalable data platforms that match their operational scope could help maximize the benefits of big data without overextending resources.

2) Smart supply chain and corporate performance: The relationship between a smart supply chain and corporate performance is exceptionally strong, with a path coefficient of 0.92. This underscores the critical role of supply chain optimization in enhancing firm outcomes. Previous research highlights how smart supply chains, powered by automation, IoT, and advanced analytics, can significantly improve efficiency, reduce costs, and enhance customer satisfaction [26]. The findings from this study corroborate these claims, suggesting that smart supply chain practices are the cornerstone of corporate performance for medium-sized enterprises [21].

This result is particularly relevant in the Indonesian context, where supply chain disruptions due to geographical challenges are common. By leveraging smart supply chain technologies, firms can better manage inventory, optimize logistics, and respond swiftly to market changes. For example, found that firms employing predictive analytics in their supply chains achieved higher resilience and adaptability, leading to superior performance metrics [8].

3) Web analytics and corporate performance: Contrary to expectations, the relationship between web analytics and corporate performance is negative, with a path coefficient of -0.18. This result suggests that the web analytics strategies adopted by medium-sized enterprises may not be effectively aligned with their business objectives. One possible explanation, as noted is the lack of integration between web analytics and broader marketing strategies [27]. When analytics tools are used in isolation, without actionable insights or follow-through, their impact on performance can be negligible or even detrimental.

Additionally, the negative relationship may reflect the misuse of analytics tools or insufficient training for employees. emphasized that the value of web analytics lies not just in data collection but in the interpretation and application of insights to drive business decisions [27]. Medium-sized enterprises in Indonesia should consider adopting a more holistic approach to web analytics, ensuring that the data collected is actionable and directly tied to key performance indicators.

Indicators and Their Relationships with Latent Variables. The study further explored the relationships between indicators and their respective latent variables, revealing several important insights, Big Data-Driven: Among the indicators for big data-driven strategies, PEE2 (possibly reflecting organizational readiness or data quality) had the highest loading factor at 1.53, emphasizing its pivotal role. This finding aligns, who highlighted the importance of data quality and organizational capabilities in deriving value from big data [28].

1) Smart supply chain: Indicators such as ES1, ES2, and ES3 demonstrated strong loadings, confirming that each dimension of the supply chain—efficiency, integration, and responsiveness is critical to overall performance. This finding supports the work which showed that supply chain integration and real-time analytics significantly enhance operational outcomes [29].

2) Web analytics market: The negative loading for CI1 (-8.15) suggests significant issues with this indicator, potentially pointing to misalignment or inefficiency in analytics implementation. This is consistent with research bwhich stressed that analytics must be effectively integrated into decision-making processes to generate positive outcomes [30].

3) Corporate performance: Indicators such as PC1, PC2, and PC3 exhibited high and consistent loadings, indicating that the construct is well-measured and reflective of overall business outcomes. This finding aligns balanced scorecard framework, which suggests that financial, operational, and customer-focused metrics are critical for assessing performance [31].

E. Implications for Medium-Sized Enterprises in Indonesia

The findings have several practical implications for medium-sized enterprises aiming to enhance corporate performance, Invest in Smart Supply Chain Technologies: Given the strong positive impact of smart supply chains, firms should prioritize investments in technologies such as IoT, predictive analytics, and automation. These tools can help streamline operations, improve decision-making, and enhance customer satisfaction.

1) Enhance big data capabilities: While big data-driven strategies have a moderate impact, their potential can be maximized by addressing resource and capability gaps. Firms

should focus on building data infrastructure, training employees, and fostering a data-driven culture.

2) Refine web analytics strategies: The negative impact of web analytics highlights the need for a more integrated and strategic approach. Firms should ensure that analytics tools are aligned with business objectives and that insights are actionable and well-implemented.

Addressing the Research Questions Based on SEM Analysis Results

How does big data-driven supply chain management influence corporate performance in Indonesian MSEs?

The analysis results show that big data-driven supply chain management significantly impacts corporate performance, as evidenced by the positive path coefficient (0.27) and significant p-value. This finding aligns with previous studies, such as those which emphasize that big data analytics in supply chain processes improves decision-making, operational efficiency, and overall firm performance[32]. By leveraging big data, Indonesian MSEs can optimize inventory management, enhance supplier relationships, and predict market demand more effectively, thereby boosting their corporate performance.

What is the role of web analytics in enhancing the market orientation of Indonesian MSEs?

The SEM analysis reveals a direct relationship between web analytics and market orientation, with a path coefficient of 0.30. This indicates that web analytics significantly enhances market orientation by providing actionable insights into customer behavior, preferences, and market trends. Corroborate these findings, suggesting that web analytics tools enable businesses to adopt a customer-centric approach, refine marketing strategies, and adapt to dynamic market conditions[33]. For Indonesian MSEs, this means leveraging web analytics to better understand their target audience, tailor their offerings, and achieve competitive advantage in a highly fragmented market.

How does market orientation mediate the relationship between digital tools (big data and web analytics) and corporate performance in Indonesian MSEs?

The mediation analysis highlights that market orientation serves as a partial mediator between digital tools and corporate performance. Big data-driven supply chain management and web analytics indirectly influence corporate performance through their positive effect on market orientation. This is supported by previous research, which emphasize that market orientation fosters customer-focused strategies and responsiveness, ultimately driving corporate success[3]. For Indonesian MSEs, investing in both big data and web analytics can cultivate market orientation, which, in turn, enhances their corporate performance. This study provides empirical evidence supporting the critical role of digital tools, such as big data-driven supply chain management and web analytics, in improving corporate performance through enhanced market orientation. Indonesian MSEs can benefit significantly from adopting these technologies to remain competitive and responsive to market demands. Future research can explore sector-specific applications of these tools and their long-term impact on business sustainability.

V. Conclusion

This study emphasizes the importance of smart supply chain technologies as the primary driver of corporate performance, followed by big data-driven strategies. However, the negative impact of web analytics indicates the need for further exploration and refinement of these strategies. The theoretical implication of these findings highlights the need for a deeper understanding of how technology and data can be effectively applied in the context of medium-sized enterprises.

Practically, this research provides value by showing that medium-sized enterprises should optimize the use of smart supply chain technologies and big data strategies to improve performance. Additionally, there is a need to enhance the implementation of web analytics through better training, appropriate tool integration, and alignment with organizational goals.

The limitations of this study include a limited focus on specific sectors and potential sampling biases. Future research could investigate the specific challenges faced by medium-sized enterprises in implementing web analytics and explore potential solutions to these issues. Furthermore, longitudinal studies could provide deeper insights into how these relationships evolve over time. Research should also explore sectoral differences and the role of external factors, such as regulatory changes and market conditions, to enrich the understanding of the drivers of corporate performance.

Acknowledgment

The authors extend their heartfelt gratitude to Universitas Pendidikan Indonesia for its unwavering support in facilitating this research. Special thanks go to the medium-sized enterprises in Indonesia that participated in this study, providing invaluable insights and data essential for our analysis. Appreciation is also expressed to colleagues and collaborators whose expertise and constructive feedback significantly enhanced the quality of this research. This study would not have been possible without the collective contributions of all involved.

References

[1] W. Lambrechts, C. J. Gelderman, J. Semeijn, and E. Verhoeven, "The role of individual sustainability competences in eco-design building projects," J. Clean. Prod., vol. 208, pp. 1631–1641, Jan. 2019, doi: 10.1016/j.jclepro.2018.10.084.

[2] A. S. Ananda, Á. Hernández-García, and L. Lamberti, "N-REL: A comprehensive framework of social media marketing strategic actions for marketing organizations," J. Innov. Knowl., vol. 1, no. 3, pp. 170–180, Sep. 2016, doi: 10.1016/j.jik.2016.01.003.

[3] A. Hurtado-Palomino, B. De la Gala-Velásquez, and J. Ccorisapra-Quintana, "The interactive effect of innovation capability and potential absorptive capacity on innovation performance," J. Innov. Knowl., vol. 7, no. 4, Oct. 2022, doi: 10.1016/j.jik.2022.100259.

[4] A. K. Kar and P. S. Varsha, "Unravelling the techno-functional building blocks of metaverse ecosystems – A review and research agenda," International Journal of Information Management Data Insights. Elsevier B.V., 2023. doi: 10.1016/j.jjimei.2023.100176.

[5] C. T. Wolf, "AI Ethics and Customer Care : Some Considerations from the Case of ' Intelligent Sales ,'" Eur. Conf. Comput. Coop. Work, pp. 1–20, 2020, doi: 10.18420/ecscw2019.

[6] A. Saidi et al., "Drivers of fish choice: an exploratory analysis in Mediterranean countries," Agric. Food Econ., vol. 10, no. 1, Dec. 2022, doi: 10.1186/s40100-022-00237-4.

[7] M. Saunila, "Innovation capability in SMEs: A systematic review of the literature," J. Innov. Knowl., vol. 5, no. 4, pp. 260–265, Oct. 2020, doi: 10.1016/j.jik.2019.11.002.

[8] E. T. Micheels and A. Boecker, "Competitive strategies among Ontario farms marketing direct to consumers," Agric. Food Econ., vol. 5, no. 1, Dec. 2017, doi: 10.1186/s40100-017-0079-8.

[9] K. L. P. Ho, H. T. Quang, and M. P. Miles, "Leveraging entrepreneurial marketing processes to ameliorate the liability of poorness: The case of smallholders and SMEs in developing economies," J. Innov. Knowl., vol. 7, no. 4, Oct. 2022, doi: 10.1016/j.jik.2022.100232.

[10] K. L. P. Ho, C. N. Nguyen, R. Adhikari, M. P. Miles, and L. Bonney, "Leveraging innovation knowledge management to create positional advantage in agricultural value chains," J. Innov. Knowl., vol. 4, no. 2, pp. 115–123, Apr. 2019, doi: 10.1016/j.jik.2017.08.001.

[11] S. Wang, J. Abbas, M. S. Sial, S. Álvarez-Otero, and L. I. Cioca, "Achieving green innovation and sustainable development goals through green knowledge management: Moderating role of organizational green culture," J. Innov. Knowl., vol. 7, no. 4, Oct. 2022, doi: 10.1016/j.jik.2022.100272.

[12] C. Li, S. He, Y. Tian, S. Sun, and L. Ning, "Does the bank's FinTech innovation reduce its risk-taking? Evidence from China's banking industry," J. Innov. Knowl., vol. 7, no. 3, Jul. 2022, doi: 10.1016/j.jik.2022.100219.

[13] A. K. Tiwari, Z. R. Marak, J. Paul, and A. P. Deshpande, "Determinants of electronic invoicing technology adoption: Toward managing business information system transformation," J. Innov. Knowl., vol. 8, no. 3, Jul. 2023, doi: 10.1016/j.jik.2023.100366.

[14] T. Papaioannou, "Innovation, value-neutrality and the question of politics: unmasking the rhetorical and ideological abuse of evolutionary theory," J. Responsible Innov., vol. 7, no. 2, pp. 238–255, May 2020, doi: 10.1080/23299460.2019.1605484.

[15] A. Chauhan, S. K. Jakhar, and C. Chauhan, "The interplay of circular economy with industry 4.0 enabled smart city drivers of healthcare waste disposal," J. Clean. Prod., vol. 279, Jan. 2021, doi: 10.1016/j.jclepro.2020.123854.

[16] R. K. Lomotey, S. Kumi, and R. Deters, "Data Trusts as a Service: Providing a platform for multi‐party data sharing," Int. J. Inf. Manag. Data Insights, vol. 2, no. 1, Apr. 2022, doi: 10.1016/j.jjimei.2022.100075.

[17] P. R. de Sousa et al., "Challenges, opportunities, and lessons learned: Sustainability in Brazilian omnichannel retail," Sustain., vol. 13, no. 2, pp. 1–17, 2021, doi: 10.3390/su13020666.

[18] A. Israel and M. Hitzeroth, "How do micro- and small-scale enterprises respond to global competition? An example of the textile survival cluster Gamarra in Lima," Int. Dev. Plan. Rev., vol. 40, no. 2, pp. 203–222, Apr. 2018, doi: 10.3828/idpr.2018.9.

[19] J. Schubert et al., "Simulation-based decision support for evaluating operational plans," Oper. Res. Perspect., vol. 2, pp. 36–56, Dec. 2015, doi: 10.1016/j.orp.2015.02.002.

[20] H. R. Abbu, D. Fleischmann, and P. Gopalakrishna, "The Digital Transformation of the Grocery Business - Driven by Consumers, Powered by Technology, and Accelerated by the COVID-19 Pandemic," Adv. Intell. Syst. Comput., vol. 1367 AISC, no. December, pp. 329–339, 2021, doi: 10.1007/978-3-030-72660-7_32.

[21] X. Wang, X. Lin, and B. Shao, "How does artificial intelligence create business agility? Evidence from chatbots," Int. J. Inf. Manage., vol. 66, Oct. 2022, doi: 10.1016/j.ijinfomgt.2022.102535.

[22] D. Ulas, "Digital Transformation Process and SMEs," Procedia Comput. Sci., vol. 158, pp. 662–671, 2019, doi: 10.1016/j.procs.2019.09.101.

[23] H. Shan, D. Bai, Y. Li, J. Shi, and S. Yang, "Supply chain partnership and innovation performance of manufacturing firms: Mediating effect of knowledge sharing and moderating effect of knowledge distance," J. Innov. Knowl., vol. 8, no. 4, Oct. 2023, doi: 10.1016/j.jik.2023.100431.

[24] L. A. Slatten, J. S. Bendickson, M. Diamond, and W. C. McDowell, "Staffing of small nonprofit organizations: A model for retaining employees," J. Innov. Knowl., vol. 6, no. 1, pp. 50–57, Jan. 2021, doi: 10.1016/j.jik.2020.10.003.

[25] M. Ben Ali, S. D'Amours, J. Gaudreault, and M. A. Carle, "Configuration and evaluation of an integrated demand management process using a space-filling design and Kriging metamodeling," Oper. Res. Perspect., vol. 5, pp. 45–58, Jan. 2018, doi: 10.1016/j.orp.2018.01.002.

[26] R. Agrawal, V. A. Wankhede, A. Kumar, S. Luthra, and D. Huisingh, "Big data analytics and sustainable tourism: A comprehensive review and network based analysis for potential future research," International Journal of Information Management Data Insights, vol. 2, no. 2. Elsevier B.V., Nov. 01, 2022. doi: 10.1016/j.jjimei.2022.100122.

[27] H. J. Kim, S. K. Ahn, and J. A. Forney, "Shifting paradigms for fashion: from total to global to smart consumer experience," Fashion and Textiles, vol. 1, no. 1. Springer Singapore, Dec. 01, 2014. doi: 10.1186/s40691-014-0015-4.

[28] D. Jain, M. K. Dash, A. Kumar, and S. Luthra, "How is Blockchain used in marketing: A review and research agenda," International Journal of Information Management Data Insights, vol. 1, no. 2. Elsevier Ltd, Nov. 01, 2021. doi: 10.1016/j.jjimei.2021.100044.

[29] J. Aslam, A. Saleem, N. T. Khan, and Y. B. Kim, "Factors influencing blockchain adoption in supply chain management practices: A study based on the oil industry," J. Innov. Knowl., vol. 6, no. 2, pp. 124–134, Apr. 2021, doi: 10.1016/j.jik.2021.01.002.

[30] U. Bamel, S. Kumar, W. M. Lim, N. Bamel, and N. Meyer, "Managing the dark side of digitalization in the future of work: A fuzzy TISM approach," J. Innov. Knowl., vol. 7, no. 4, Oct. 2022, doi: 10.1016/j.jik.2022.100275.

[31] A. Mäkivierikko, P. Bögel, A. N. Giersiepen, H. Shahrokni, and O. Kordas, "Exploring the viability of a local social network for creating persistently engaging energy feedback and improved human well-being," J. Clean. Prod., vol. 224, pp. 789–801, Jul. 2019, doi: 10.1016/j.jclepro.2019.03.127.

[32] R. Dwivedi, S. Nerur, and V. Balijepally, "Exploring artificial intelligence and big data scholarship in information systems: A citation, bibliographic coupling, and co-word analysis," Int. J. Inf. Manag. Data Insights, vol. 3, no. 2, Nov. 2023, doi: 10.1016/j.jjimei.2023.100185.

[33] P. Scott, "General Motors' other franchise system: Creating an effective distribution model for Frigidaire," Bus. Hist., vol. 64, no. 1, pp. 183–200, Jan. 2022, doi: 10.1080/00076791.2020.1714594.

Color Multi-Focus Image Fusion Method Based on Contourlet Transform

Zhifang Cai
School of Integrated Circuits (Artificial Intelligence), Beijing Polytechnic, Beijing, 100176, China

***Abstract*—Color Multi-Focus Image Fusion (MFIF) technology finds extensive use in areas such as microscopy, astronomy, and multi-scene photography where high-quality and detailed images are vital. This paper presents the Contourlet Transform alongside its enhanced version, the Non-Subsampled Contourlet Transform (NSCT), aimed at improving the outcomes of image fusion, with the support of Laplacian Pyramid (LP) decomposition. The NSCT framework overcomes challenges like spectral aliasing and directional sensitivity, leading to images with sharper edges, enriched texture details, and preserved delicate information. Experimental findings highlight the NSCT-based fusion algorithm's superiority. Subjective assessments indicate that using the NSCT method results in images with sharp and well-defined object boundaries, outstanding contrast, and abundant textures without the creation of artifacts, markedly excelling beyond traditional techniques such as the Contourlet Transform, Non-Subsampled Shearlet Transform (NSST) and Rolling Guidance Filtering (RGF). Objective measures verify its effectiveness: In the first dataset, it attains an average gradient (AG) of 8.36 and an edge intensity (EI) of 3.29E-04, while in the second dataset, it reports an AG of 21.39 and an EI of 4.06E-04, significantly outperforming other methods. Moreover, the NSCT method offers competitive computational speed, balancing runtime with high-quality fusion performance. These results establish the proposed method as a powerful and efficient solution for color MFIF, offering notable performance benefits and practical utility in various imaging fields.**

Keywords*—*Contourlet transform; image fusing; NSCT; Laplacian Pyramid; color multi-focus

I. Introduction

Optical lenses are vital for photography, visual perception, and computational tasks in real-world scenarios. Despite their pivotal role in imaging, the core principles of lens operation inherently restrict their ability to achieve full scene sharpness. The optical features of cameras enable precise focus on selected subjects within the depth of field, which consequently causes other parts to remain blurred and unfocused. This limitation poses significant challenges in producing a single image with a uniform focus across the entire scene [1-2]. These constraints are particularly crucial in fields where precision and clarity are essential, such as medical imaging, microscopic analysis, and astronomy.

Color Multi-Focus Image Fusion (MFIF) technology emerges as an essential solution to address these challenges. MFIF combines multiple images with varying focus zones of a scene into a single entirely focused image. This technique utilizes advanced fusion algorithms [3], enhancing image clarity and usefulness by amalgamating information from different source images, thus enabling comprehensive image analysis. MFIF finds widespread application across diverse fields, such as merging textile fiber images in microscopy, integrating astronomical images from telescopes, and combining multi-focus images from consumer cameras [4].

The need for accurate and comprehensive image representation is growing across various sectors, where extracting precise information from visual data is critical. Traditional image fusion methods, including pixel-based, feature-based, and transform-based approaches, have substantially progressed in fulfilling this need. However, challenges remain, like texture detail loss, artifacts, and blurring issues. Transform-based techniques, such as wavelets, show promise but are limited in effectively handling multi-resolution and directional data.

The Contourlet Transform offers a multi-resolution and multi-directional framework, which helps overcome some limitations by better representing edges and textures. However, its reliance on subsampling may introduce artifacts and limit fusion quality. The Non-Subsampled Contourlet Transform (NSCT) is implemented for its superior abilities to the standard Contourlet Transform. By removing subsampling, NSCT enhances the preservation of texture and detail, making it highly appropriate for MFIF applications. The practical usage of NSCT covers medical imaging, satellite imagery, and surveillance applications.

Despite advancements in MFIF, several unresolved issues still exist. Existing approaches often struggle with clarity, failing to preserve detailed textures and sharp edges, which results in fused images that are either blurred or prone to artifacts. Additionally, these algorithms have limited adaptability to various imaging scenarios, such as microscopy, astronomy, and general photography, hindering their practical application. Quantitative assessment indicators like gradient and edge strength often show suboptimal performance, underscoring the need for better techniques. Furthermore, the complexity of sophisticated fusion methods frequently reduces computational efficiency, making them unsuitable for real-time or large-scale use. These challenges underscore the necessity for innovative approaches to improve fusion quality, enhance generalizability, and maintain computational viability. The key contributions of the article are:

- The enhanced transform framework integrates the NSCT into the fusion mechanism, showing superior texture and edge conservation compared to conventional methods.
- Improved fusion algorithm displaying impressive

scalability and flexibility in diverse imaging contexts, outperforming leading techniques in both subjective and objective evaluations.

- A comprehensive validation process through experimental methods highlights significant improvements in gradient and edge strength metrics.
- Addressing the limitations of traditional techniques to effectively cater to various applications, such as microscopy, astronomy, and general photography.

The rest of the article is organized as follows: Section II begins with an in-depth review of MFIF technology, followed by an introduction to the core principles of the Contourlet Transform and its advanced form, the NSCT. Section III thoroughly discusses the proposed fusion algorithm, detailing its design and how NSCT is integrated into the image fusion process. Section IV offers experimental validation, featuring both qualitative and quantitative assessments to demonstrate the improved performance of the proposed method. Finally, Section V summarizes the study by synthesizing the outcomes, addressing the limitations of the current approach, and suggesting future research directions to advance MFIF.

II. Related Works

The evolution of image processing techniques, notably the contour transform and MFIF, has driven significant progress in various fields. This section summarizes these areas' key contributions and advancements, highlighting their applications, improvements, and integration into new solutions. The most prominent inferences from the literature are tabulated in Table I.

A. Contourlet Transform

Renowned for its effectiveness in capturing edge and texture details, the contour transform finds numerous applications across multiple sectors. Jahangir H et al. developed an innovative approach using the contour transform to evaluate damage in prestressed concrete slab structures. Applying the transform to the data on the modal curvature in damaged and unscathed states led to the precise determination of the severity of structural damage [5]. Hasan M. M. et al. contour improved the application by combining the contourlet transform with fine-tuning of the pretrained VGG19 model, improving the identification of gastrointestinal polyps in endoscopic images and achieving greater diagnostic precision [8]. Enhancements to the Contourlet contour transform, particularly the development of the NSCT, have notably expanded its applications. Li W et al. merged NSCT with Convolutional Neural Networks for multimodal medical image fusion, successfully addressing intensity inconsistencies and achieving better fusion performance [6]. Kollem et al. presented an improved Total Variation model to enhance grayscale and color brain tumor images. Using NSCT with adaptive threshold techniques improved the preservation of contour and texture features [7]. Hu F's team also used the NSCT and Schur decomposition to create an excellent blind watermarking system to protect the copyright of color digital images, which led to higher peak signal-to-noise ratios and structural similarity indices [9].

B. Multi-Focus Image Fusion (MFIF)

Significant progress has been made in MFIF across various domains, like medical imaging and robotic vision, where image clarity and precision are crucial. Li H addressed the challenge of detecting focused regions in MFIF by proposing a fusion approach relying on focus feature extraction. This approach successfully preserved source images' intricate textures and edge structures while effectively identifying focused area edges [10]. Panigrahy C et al. introduced an innovative MFIF classification method and reviewed current fusion approaches, demonstrating promising performance and potential applications [11]. Recent innovations in MFIF methods include wavelet-based techniques and incorporating deep learning. Wang Y developed a translation-invariant approach using the wavelet transform and fractal dimensions to evaluate clarity, integrating it with the Otsu thresholding to fuse detail coefficients, and achieved competitive results in subjective and objective evaluations [12]. Liu et al. [13] designed an MFIF method using deep learning, employing the VGG-19 network for feature extraction alongside deconvolution modules to generate decision maps. This method indicated superior performance over other segmentation accuracy and fusion metrics approaches, especially for focused and non-focused regions in color multi-focus images.

Detailed contour transform and MFIF studies have yielded promising results across diverse applications. Integrating NSCT into MFIF is pivotal, offering substantial potential to improve fusion quality and precision. This integration overcomes existing challenges and paves the way for developing improved techniques.

Upon examining the cited research, we found that numerous current techniques face drawbacks despite substantial progress in multi-focus image fusion methods. In particular, conventional methods like NSST and RGF encounter difficulties in maintaining intricate texture details and ensuring consistency across different image scenarios. Additionally, achieving computational efficiency poses challenges in high-resolution imaging, as current fusion techniques often grapple with balancing performance against processing speed. Moreover, many reviewed methods display issues such as spectral aliasing and diminished directional sensitivity, resulting in inferior image fusion quality.

In response to the identified gaps, our research introduces a fusion algorithm based on NSCT, augmented with Laplacian Pyramid decomposition, to overcome the mentioned limitations effectively. This method enhances directional selectivity, reduces spectral aliasing, and improves edge preservation while retaining computational efficiency. We achieve superior fusion quality by utilizing NSCT's capability to capture features in multiple directions and combining it with Laplacian Pyramid decomposition, resulting in sharper edges and finer texture details. These advancements directly tackle the issues noted in the existing literature, offering a more reliable and efficient solution for color multi-focus image fusion.

TABLE I SUMMARY OF RELATED WORKS ON IMAGE FUSION TECHNIQUES, MODELS, CONTRIBUTIONS, AND CHALLENGES

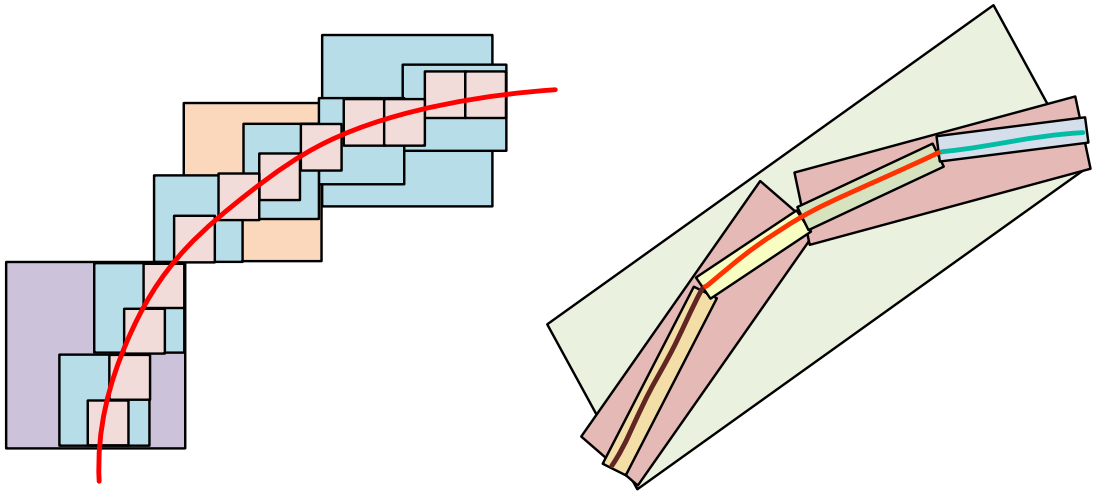
Author	Year	Model	Key Contribution	Challenges
Jahangir H et al. [5]	2021	Contourlet Transform	Damage evaluation in prestressed concrete slabs using modal curvature data.	Application limited to structural damage; no focus on multifocus image fusion.
Li W et al. [6]	2020	NSCT + CNN	Enhanced multimodal medical image fusion addressing intensity differences.	High computational complexity due to CNN integration.
Kollem S et al. [7]	2022	NSCT + Adaptive Threshold	Enhanced grayscale and color brain tumor images using partial differential equations.	Limited to specific medical imaging applications; generalizability not tested.
Hasan M M et al. [8]	2019	Contourlet + Fine-tuned VGG19	Improved identification of gastrointestinal polyps in endoscopic images.	Reliance on pre-trained models may restrict adaptability to other domains.
Hu F et al. [9]	2020	NSCT + Schur Decomposition	Robust blind watermarking scheme for copyright protection of color digital images.	It focuses on copyright protection; there is no direct applicability to MFIF.
Li H [10]	2018	Focus Feature Extraction	Accurate detection of focused regions in MFIF.	They may struggle with complex textures or multi-scene images.
Panigrahy C et al. [11]	2017	Classification-Based Fusion	Novel classification method for MFIF with broad application prospects.	Performance heavily dependent on classifier accuracy.
Wang Y et al. [12]	2016	Wavelet Transform + Fractal Dimension	Translation-invariant method for MFIF with effective edge fusion.	Lacks directional sensitivity and detail preservation.
Liu S et al. [13]	2019	Deep Learning (VGG19 + Deconvolution)	Enhanced segmentation and fusion for color multi-focus images.	High computational cost; potential overfitting with limited data.

III. COLOR MFIF SUPPORTED BY CONTOURLET TRANSFORM

This section explores the application of the Contourlet Transform for Color MFIF. The study delves into the basic concepts, construction techniques, and transformation processes linked to the Contourlet Transform. Additionally, it assesses its shortcomings and presents the NSCT as a possible improvement in this domain. The research outlines the elements, decomposition principles, and operational procedures of NSCT in the context of color MFIF, highlighting its enhanced performance features.

A. Contourlet Transform Method

The Contourlet Transform is an effective technique for image processing in the realm of color MFIF. It stands out due to its multi-scale and multi-directional decomposition capabilities, which enhance the depiction of image contours and textures. The procedure begins with a Laplacian Pyramid (LP) decomposition, systematically dividing the image into sub-images at various scales [14-15]. The high-frequency components undergo directional decomposition using the Directional Filter Bank (DFB), while the low-frequency components are iteratively decomposed, creating the complete contour wave structure. Unlike conventional wavelet transforms that employ square wavelet functions, the Contourlet Transform uses rectangular ones, allowing for aspect ratio adjustments and better adaptation to image contours across scales. This improved directionality offers a more accurate representation of curves and contours. Fig. 1 shows a visual comparison of different transformation methods for representing image curves.



(a) Description of curves using wavelet transform

(b) Countourlet transformation for describing curves

Fig. 1. Description of image curves by different transformation methods.

Contourlet transform comprises two main components: LP decomposition and DFB filtering [16]. LP Decomposition captures intricate details across a spectrum of frequencies by deconstructing the image into smaller elements. DFB Filtering provides a multi-directional decomposition method that extracts directional characteristics. The LP decomposition refines the Gaussian pyramid approach by reducing high-frequency information loss by subtracting layers. This involves interpolation to align image dimensions at each scale level. This process is achieved through repeated convolution and down-sampling operations. The Gaussian pyramid is the fundamental type of image pyramid, generating a pyramid-shaped data structure through convolution and down-sampling operations. LP further develops from the Gaussian pyramid, addressing the loss of high-frequency information during computation. LP achieves this by subtracting each layer of the Gaussian pyramid from its previous layer, obtaining the LP decomposed image. In this process, interpolation operations are required for rows and columns, extending fine-sized images to coarse-sized ones. The output representation of the fine-sized image is expressed as shown in Eq. (1).

$$\sum_{m=-2}^{2}\sum_{n=-2}^{2} G_{l-1}(2i+m, 2j+n)w(m,n) = G_l(i,j) \quad (1)$$

In Eq. (1), l represents the decomposition level and $G_l(i,j)$ represents the Gaussian pyramid image at the level l. $w(m,n)$ stands for the sliding window with a size of 5 × 5. i and j are the symbols of rows and columns of the corresponding image. Subsampling operations can obtain a sequence of Gaussian pyramid images, followed by interpolation and dilation operations to make the sizes of images Gl and Gl-1 consistent, as it described in Eq. (2).

$$4\sum_{m=-2}^{2}\sum_{n=-2}^{2} G'_{l-1}\left(\frac{i+m}{2}, \frac{j+n}{2}\right)w(m,n) = G_l * (i,j) \quad (2)$$

In Eq. (2), the calculation of $G'_{l-1}\left(\frac{i+m}{2}, \frac{j+n}{2}\right)$ is shown in Eq. (3).

$$G'_{l-1}\left(\frac{i+m}{2}, \frac{j+n}{2}\right) = \begin{cases} G_l\left(\frac{i+m}{2}, \frac{j+n}{2}\right), & \frac{i+m}{2} \text{ and } \frac{j+n}{2} \text{ are integers} \\ 0, & \text{other} \end{cases} \quad (3)$$

LP is defined as shown in Eq. (4).

$$\begin{cases} LP_l = G_l - G_{l+1}^*, & 0 < l < N \\ LP_N = G_N, & l = N \end{cases} \quad (4)$$

In Eq. (4), N represents the highest level of the pyramid. After obtaining the LP image sequence, fusion is performed according to the selected rules to get a fused pyramid sequence image, and finally, a reconstruction process is carried out. The calculation for reconstruction is shown in Eq. (5).

$$\begin{cases} G_N = LP_N, & when\ l = N\ 0 < l < N \\ G_l = LP_N + G_{l+1}^*, & when\ 0 < l < N \end{cases} \quad (5)$$

The LP decomposition and reconstruction processes, with a schematic diagram, are shown in Fig. 2.

The DFB is a tree-structured decomposition method that generates multiple directional subbands at each layer, with each subband taking a wedge shape. A construction method for DFB based on a sector filter group is introduced to simplify directional decomposition without altering the input signal. This method comprises two main components: a dual-channel sector filter group and shear processing. The dual-channel sector filter group utilizes sector filters to decompose the two-dimensional spectrum into horizontal and vertical directions. On the other hand, shear processing involves sampling and rearranging the image. The method effectively captures frequency domain information across various directions while ensuring complete reconstruction by incorporating shear and inverse shear operations before and after the dual-channel filter group.

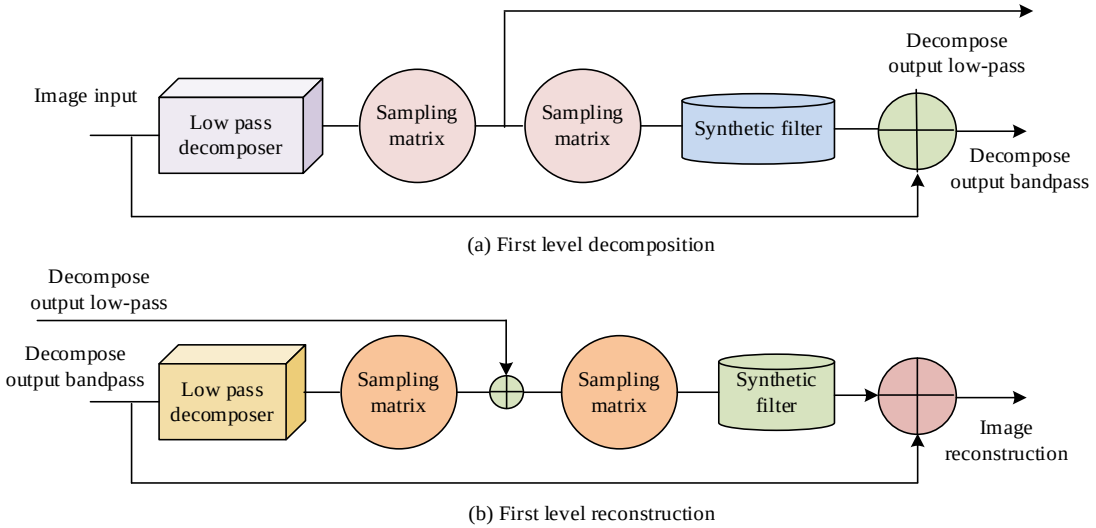


Fig. 2. A schematic diagram of the decomposition and reconstruction of the LP.

For instance, in a 3-level decomposition, the method can produce 8 directional spectrum partitions. Subbands labeled 1, 2, and 3 belong to the first channel, while subbands labeled 4, 5, 6, and 7 are part of the second channel. Integrating the Laplacian Pyramid (LP) with DFB forms a two-layer filter structure, referred to as the Contourlet filter group. During transformation, the image is first decomposed into high-frequency and low-frequency components. The low-frequency component undergoes further multi-resolution analysis, while the high-frequency components are processed through DFB to generate directional subbands at various orientations. This iterative process gradually connects singular points into linear structures, capturing the edge contours of the image. This capability allows the Contourlet Transform to effectively capture image features, such as edges and textures, offering a superior representation compared to traditional methods. Fig. 3 depicts the flowchart of the Contourlet transformation, illustrating the sequential decomposition and feature extraction process.

To improve NSCT-based fusion, images must be pre-processed to make them more transparent and less distorted before transforming. Preprocessing improves contrast and gets rid of unnecessary variations. This helps NSCT retain structural details better, which makes it perfect for medical imaging and analyzing satellite data. It also makes extracting more accurate features and fusion easier, which improves visualization in critical areas like medical imaging and remote sensing.

B. MFIF Supported by Improved Contourlet

Within the Contourlet framework, downsampling is employed during integrating pyramid and directional filtering decomposition, potentially resulting in spectral aliasing, which diminishes the transformation's directional understanding [17]. The Non-Subsampled Contourlet Transform (NSCT) was developed to address this issue. NSCT's decomposition method includes two primary elements: the Non-Subsampled Pyramid Filter Bank (NSPFB) for decomposing scales and the Non-Subsampled Direction Filter Bank (NSDFB) for analyzing directions. This framework maintains the Contourlet Transform's ability to effectively capture curves while removing spectral aliasing that arises from downsampling. Consequently, NSCT improves the accuracy and dependability of directional sensitivity in image processing. Fig. 4 illustrates the NSCT's structural diagram, detailing its components and workflow.

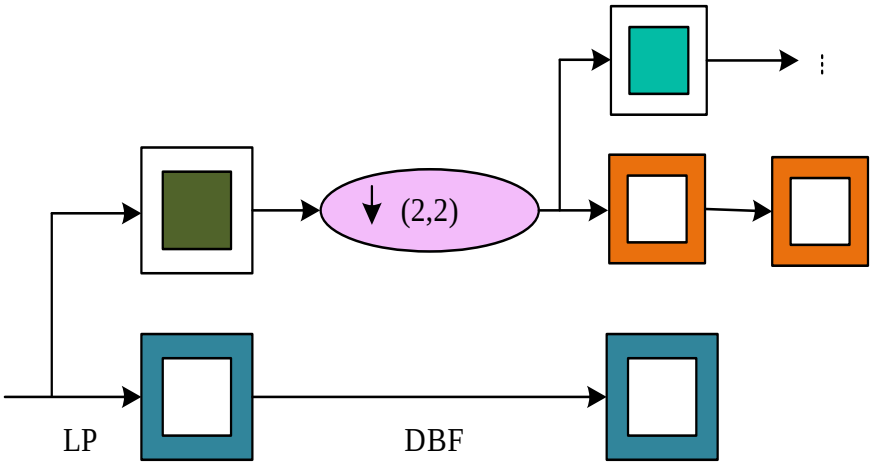


Fig. 3. Flow chart of Contourlet transformation.

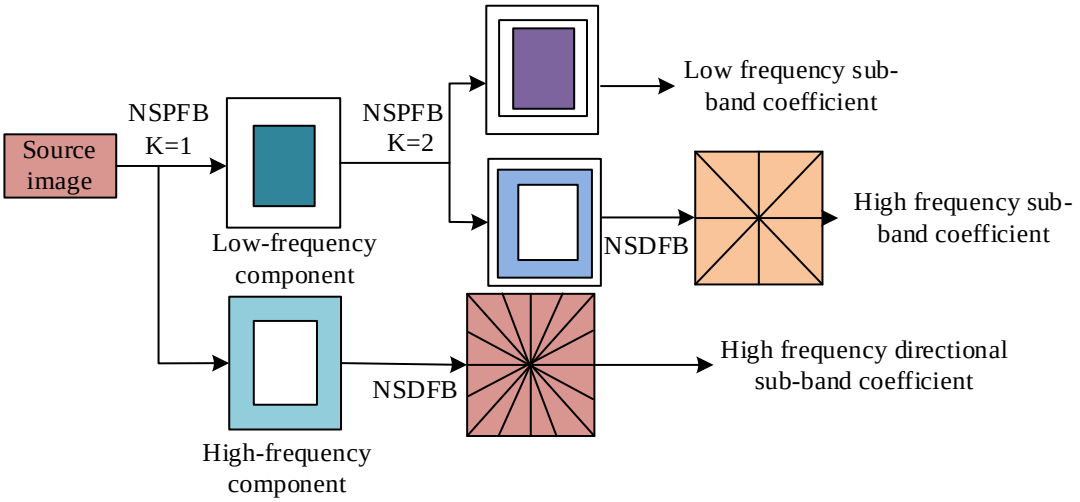


Fig. 4. Structural diagram of the NSCT.

Conventional approaches for assessing image sharpness frequently depend on individual features, which may be inadequate in specific situations [18-19]. To mitigate these limitations, we propose an MFIF technique utilizing the NSCT. The algorithm initiates with the decomposition of two source images through NSCT, extracting their subband coefficients. This results in multiple subbands of uniform size, aiding the identification and exploitation of gray feature correlations across subbands. In the NSCT realm, reference coefficients' sibling subbands contain abundant correlation data, which is harnessed to enhance fusion. The method employs correlation weights from sibling subbands in conjunction with average gradient metrics for a robust fusion process for high-frequency subbands. Initial fusion results emerge from applying these rules, followed by an inverse NSCT transform. To further boost the fusion's quality, edge detection methods are utilized on the preliminary result, isolating edge contour data. This data is then superimposed on the preliminary fused result, enhancing clarity and feature retention. The NSCT-supported fusion algorithm's comprehensive framework is depicted in Fig. 5, showcasing its efficacy in capturing and integrating edge and texture details for superior image fusion quality.

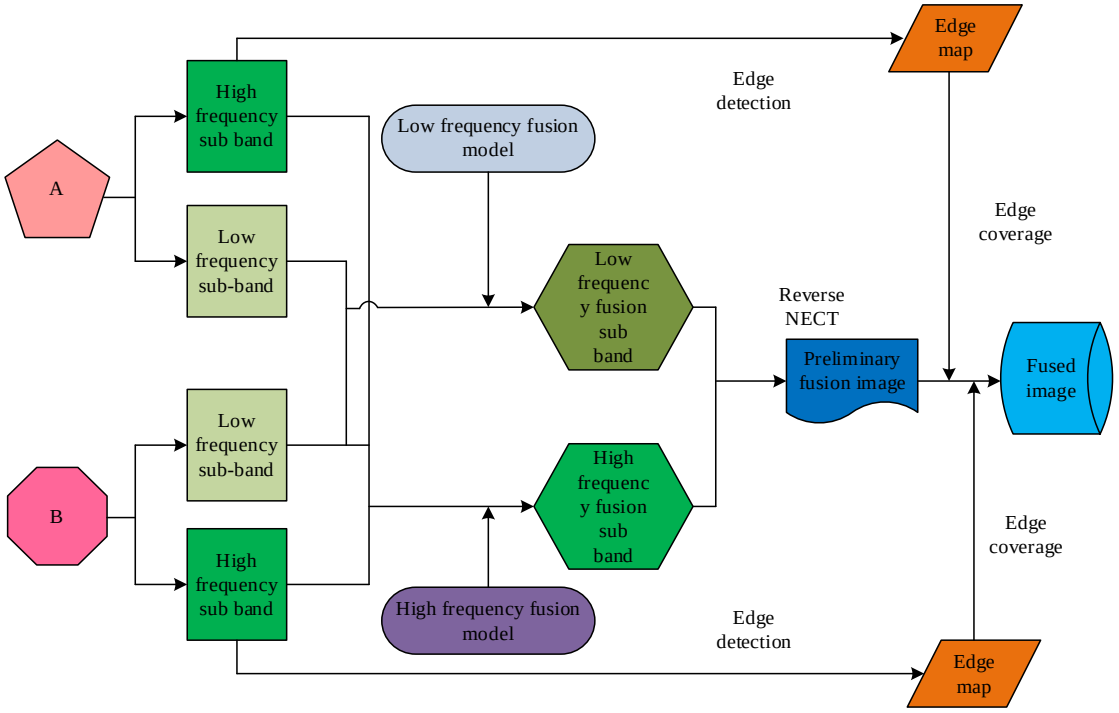


Fig. 5. A fusion algorithm framework based on NSCT.

In NSCT decomposition, the low-frequency subband is extracted from the source image via successive application of low-pass filters over several layers. This research adopts statistical features like local information entropy and enhanced local Laplacian energy for fusing low-frequency coefficients. Information entropy is a reliable indicator for assessing the amount of information in an image. The low-frequency subband often contains a significant part of the image's information, making entropy an effective measure for fusion [20]. Conversely, Laplacian energy indicates the extent of grayscale variation in an image, enhancing the analysis based on entropy. Eq. (6) provides the mathematical representation for local information entropy and local Laplacian energy at point *h* in the image *x*. These metrics facilitate the fusion of low-frequency coefficients by capturing essential informational and structural distinctions within the image.

$$G_{r,x}(h)=\sum_{q\in Q}\omega(q)S_{r,x}(q)^2 \tag{6}$$

In Eq. (6), $h(a,b)$ represents pixel coordinates, $\omega(q)$ is a Gaussian weight matrix, Q denotes the neighborhood centered around h, and r takes values of 1 or 2. The representation of local information entropy at point q within the neighborhood Q is shown in Eq. (7).

$$S_{1,x}(q)=\sum_{a=0}^{\gamma-1}p_a\log(p_a) \tag{7}$$

In Eq. (7), p_i represents the proportion of pixels with a grayscale value a in the neighborhood Q, and γ represents the grayscale level. The calculation of locally improved Laplacian energy is given by Eq. (8).

$$S_{1,x}(a,b)=\left|2C_x(a,b)-C_x(a-1,b)-C_x(a+1,b)\right|+\left|2C_x(a,b)-C_x(a,b-1)-C_x(a,b+1)\right| \tag{8}$$

In Eq. (8), $L_r(h)$ represents the low-frequency subband coefficients obtained from the decomposition of the image x. The feature matching degree of low-frequency component coefficients between images A and B at point h is expressed in Eq. (9).

$$L_r(h)=\frac{2\sum_{q\in R}\omega(q)\left|S_{r,A}(q)\right|\left|S_{r,B}(q)\right|}{G_{r,A}(h)+G_{r,B}(h)} \quad (9)$$

Assuming α is the matching threshold, if $L_r(h)<\alpha$, the feature weight matrix is solved through the calculating process of Eq. (10).

$$W_{r,A}=\begin{cases}1, G_{r,A}(h)\geq G_{r,B}(h)\\ 0, G_{r,A}(h)<G_{r,B}(h)\end{cases} \quad (10)$$

If $L_r(h)\geq\alpha$, the feature weight matrix is solved through the calculating process of Eq. (11).

$$\begin{cases}W_{r,A}=\begin{cases}1-(L_r(h)-\alpha), G_{r,A}(h)\geq G_{r,B}(h)\\ L_r(h)-\alpha, G_{r,A}(h)<G_{r,B}(h)\end{cases}\\ W_{r,B}=1-W_{r,A}(h)\end{cases} \quad (11)$$

Eq. (12) shows how the fusion weight matrix for low-frequency components is obtained by applying the average principle to weight matrices.

$$\begin{cases}W_A(h)=(W_{1,A}(h)W_{2,A}(h))/2\\ W_B(h)=(W_{1,B}(h)W_{2,B}(h))/2\end{cases} \quad (12)$$

The calculation of the low-frequency coefficients after merging images A and B is given in Eq. (13).

$$C_F(h)=C_A(h)W_A(h)+C_B(h)W_B(h) \quad (13)$$

High-frequency elements contain the original image's fine details and textural information, while the average gradient quantitatively describes variations in the image's gradient. A higher average gradient indicates more detailed texture and sharper image clarity. To improve the fusion of high-frequency elements, this research uses the local average gradient as a basis for determining fusion weights. This method assures that texture-rich areas are effectively maintained and emphasized in the combined image. The mathematical expression for the region's average gradient concerning the high-frequency subband coefficients at the *p-th* scale and *s-th* direction of image *x* is provided in Eq. (14). This formulation highlights the importance of gradient-based metrics in enhancing the quality of fusion for high-frequency details.

$$T^s_{s,k}(a,b)=\frac{\sum_{(a,b)\in Q}g(a,b)}{|Q|} \quad (14)$$

$|Q|$ represents the size of the neighborhood window. The calculation of $g(a,b)$ is given in Eq. (15).

$$g(a,b)=\frac{\sqrt{\left[C^s_{x,q}(a,b)-C^s_{x,q}(a+1,b)\right]^2+\left[C^t_{x,q}(a,b)-C^t_{x,q}(a+1,b)\right]^2}}{2} \quad (15)$$

In Eq. (15), $C^s_{x,q}$ and $C^t_{x,q}$ represent the high-frequency subband coefficients of image *x* in the *p-th* scale and *s-th*, *t-th* directions. Within the NSCT domain, the reference node is closely associated with its eight adjacent nodes in the identical subband, as these neighbors carry considerable information about the reference node. At the same scale, nodes situated in the same spatial position but in different directional subbands are known as sibling nodes. Although these siblings provide valuable data, they typically offer less insight regarding the reference node than the immediate neighbors. Conversely, the node at the same spatial location in the previous scale, termed the parent node, holds the least amount of relevant information about the reference node. Understanding these nodes' hierarchical and spatial connections is crucial for capturing structural and contextual information in NSCT-based image processing. Fig. 6 comprehensively depicts these relationships, illustrating the interaction between coefficients in the NSCT domain.

The NSCT-based fusion technique balances computational efficiency with feature extraction by selecting appropriate decomposition levels and fusion strategies. Researchers have tested NSCT in challenging situations and found it keeps image details and edge structures better than other methods, even when the input images are distorted. NSCT achieves this robustness by enhancing directional information, reducing undesirable artifacts, and guaranteeing high-quality fused images under diverse conditions.

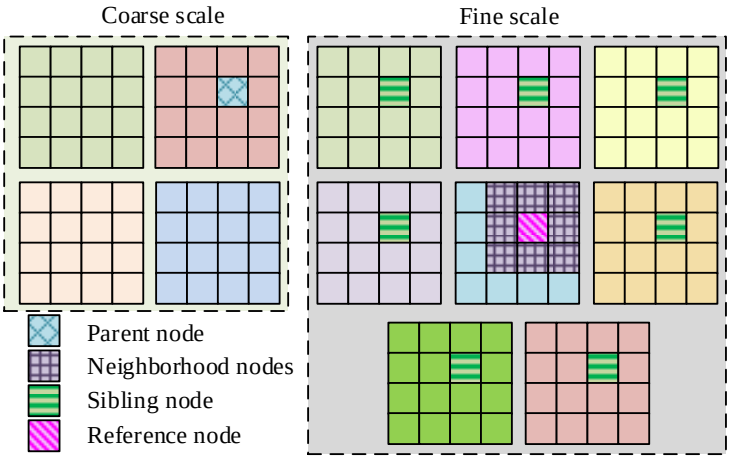


Fig. 6. A schematic representation of the relationship between the NSCT domain coefficients.

IV. Results and Discussion

This section thoroughly assesses the proposed Color MFIF algorithm, which leverages the Contourlet Transform and focuses on the NSCT. Experimental validation encompasses subjective and objective analyses demonstrating the algorithm's practical effectiveness and comparative benefits.

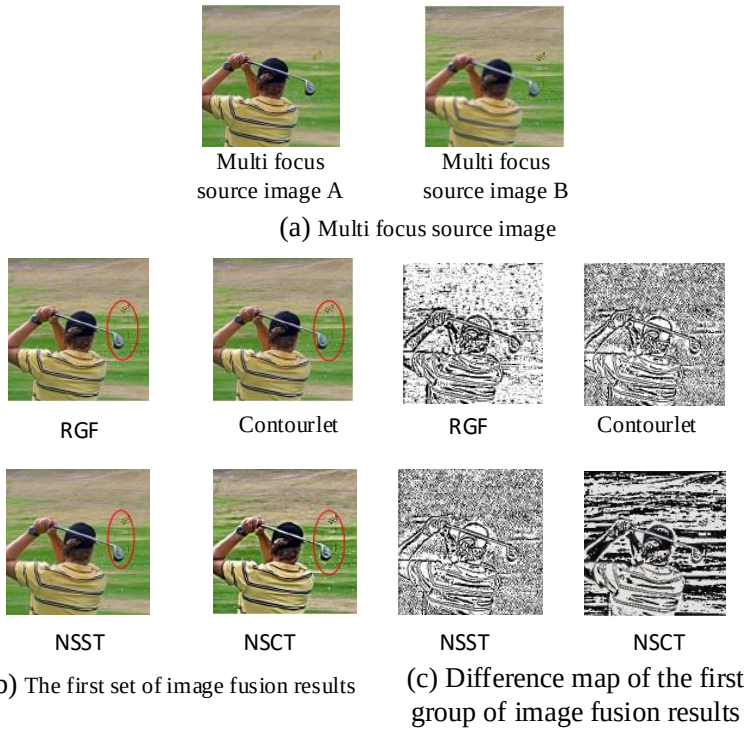
A. Subjective Evaluation

Subjective evaluations were conducted based on qualitative assessments to measure the practicality of the NSCT-based multi-focus image fusion algorithm. Testing utilized two datasets from a registered true-color multi-focus image collection. The fusion outputs from the NSCT-based algorithm were compared to results from conventional methods, such as the Contourlet Transform, Non-Subsampled Shearlet Transform (NSST), and Rolling Guidance Filtering (RGF) techniques. Table II lists the parameters used in these experiments, ensuring uniformity and comparability among all methods tested.

TABLE II Experimental Parameters

Parameter	Set value
Input picture A	Clear
Input picture B	Blur
Spatial filter	Gaussian filter with a standard deviation of 1.0
Gradient weight	0.5
Iterations	10
Output image size	Consistent with the input image
Output image format	PNG

The image showcasing the fusion effect for the initial set of multifocus source images featuring a clock, along with its corresponding difference image, is illustrated in Fig. 7. Fig. 7(a) presents color multifocus source images A and B, while Figure 7(b) shows the resultant fusion effect image. Fig. 7(c) provides the associated difference image. Observations from Fig. 7 reveal that the RGF method delivers a solid overall fusion effect, producing a visually appealing result. However, examining the highlighted area in the red circle reveals noticeable blurring in the flag section with the RGF method. The Contourlet Transform approach achieves a generally good fusion effect but suffers from some detail loss and certain artifact presence in the hand and back areas. NSST also performs well in fusion, though it introduces slight blurring at the flag's edge within the red box, resulting in minor detail information loss and slightly lowering the clarity of the fusion effect. Contrastingly, with the NSCT method, object edges are distinctly sharp and clear, with abundant texture details and no detail loss. This fusion effect is optimal, featuring strong contrast effects, indicating that the proposed method in this study provides superior fusion effects, free of artifacts in surrounding areas, demonstrating high authenticity and precision.



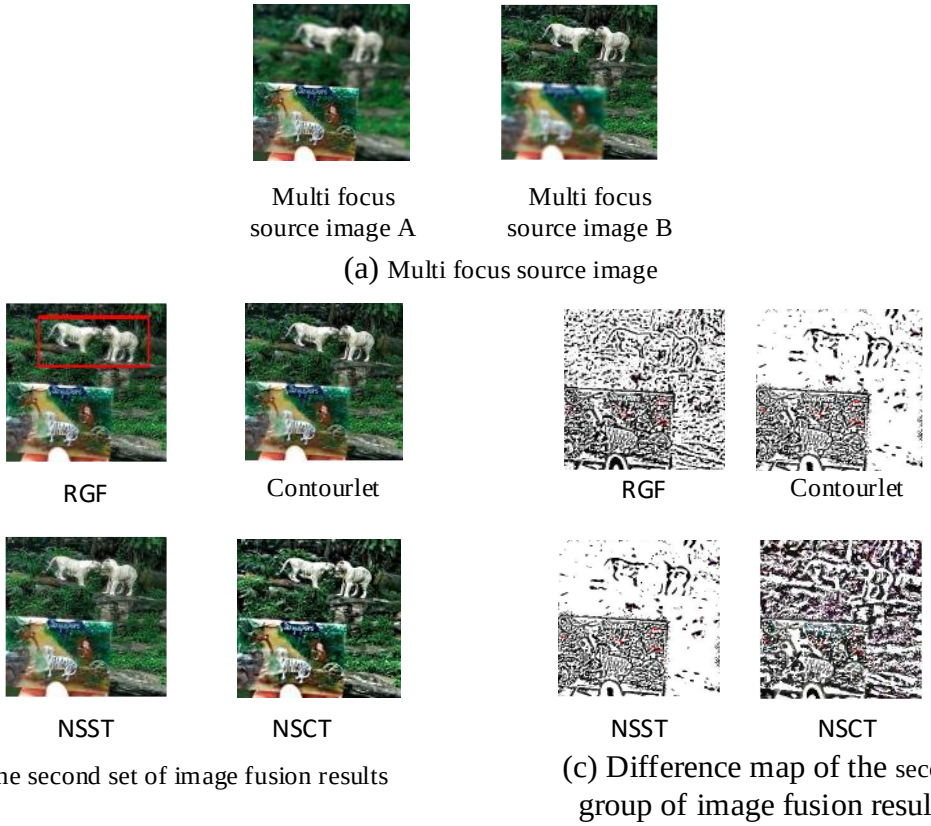
(a) Multi focus source image

(b) The first set of image fusion results

(c) Difference map of the first group of image fusion results

Fig. 7. The first set of image fusion results and difference maps.

The images resulting from the fusion process and the corresponding difference maps related to the clock in the second set are illustrated in Fig. 8. Fig. 8(a) presents the colored multi-focus source images A and B from the second set. Fig. 8(b) shows the fusion effect image of the second set, and Fig. 8(c) presents the associated difference map. From Fig. 8, it can be noted that, within the second set of experimental images, the RGF method displays a slight reduction in clarity, as seen through subjective analysis. Additionally, the RGF method causes subtle blurring and pseudo-artifacts near the tiger. The Contourlet transform method achieves a good overall fusion effect, though with a minor drop in clarity. Even though the NSST transform method offers rich visual effects and abundant information, it suffers from a decline in clarity and produces unnatural transitions and pseudo-artifacts in the peripherals. In contrast, the NSCT method significantly excels, presenting rich information, clearly viewing scene textures, and uniform brightness distribution. It does not generate pseudo-artifacts, effectively addressing detail loss due to edge blurring. Moreover, there are no pseudo-artifacts in the tiger's texture information, showcasing high contrast. This suggests that the color MFIF algorithm proposed by the research institute delivers a notable fusion effect.



(a) Multi focus source image

(b) The second set of image fusion results

(c) Difference map of the second group of image fusion results

Fig. 8. The second set of image fusion results and difference maps.

B. Objective Evaluation

The study continued its analysis of the fusion result images for both groups by focusing on quantitative evaluations. The metrics used for assessment included Average Gradient (AG), Standard Deviation (STD), Spatial Frequency (SF), Visual Information Fidelity (VIFF), and Edge Intensity (EI). AG assessed contrast details and texture changes in the images, with higher values indicating more precise information. STD measured the range of grayscale levels, where higher values signaled a more natural brightness in the fused image. SF captured the ratio of spatial frequency errors, and VIFF evaluated image quality based on its visual fidelity, with higher values reflecting better fusion quality. EI gauged image clarity. Initially, the clarity of each fusion technique was validated through AG and EI metrics. To confirm experiment reliability, 10 tests were conducted, and the averages were calculated as final results. Fig. 9 illustrates the AG and EI values of the fusion outcomes for both groups. In Fig. 9(a), particularly in the first group's multi-focus source images fusion, the NSCT method achieved a substantial AG of 8.36, surpassing the traditional contourlet transform, NSST, and RGF methods by 0.16, 0.72, and 0.25, respectively. Fig. 9(b) indicates that the NSCT method also reached a notable EI value of 3.29E-04 in the first group, showing enhancements over the other methods by 0.48E-04, 0.59E-04, and 0.72E-04. Fig. 9(c) demonstrates that the NSCT method obtained a high AG of 21.39 in the second group, outperforming other methods with improvements of 7.28, 7.39, and 6.14. Fig. 9(d) portrays the NSCT method achieving a significant EI value of 4.06E-04 in the second group, markedly better than the alternatives, illustrating the remarkable efficacy of the proposed image fusion technique developed by the research institute.

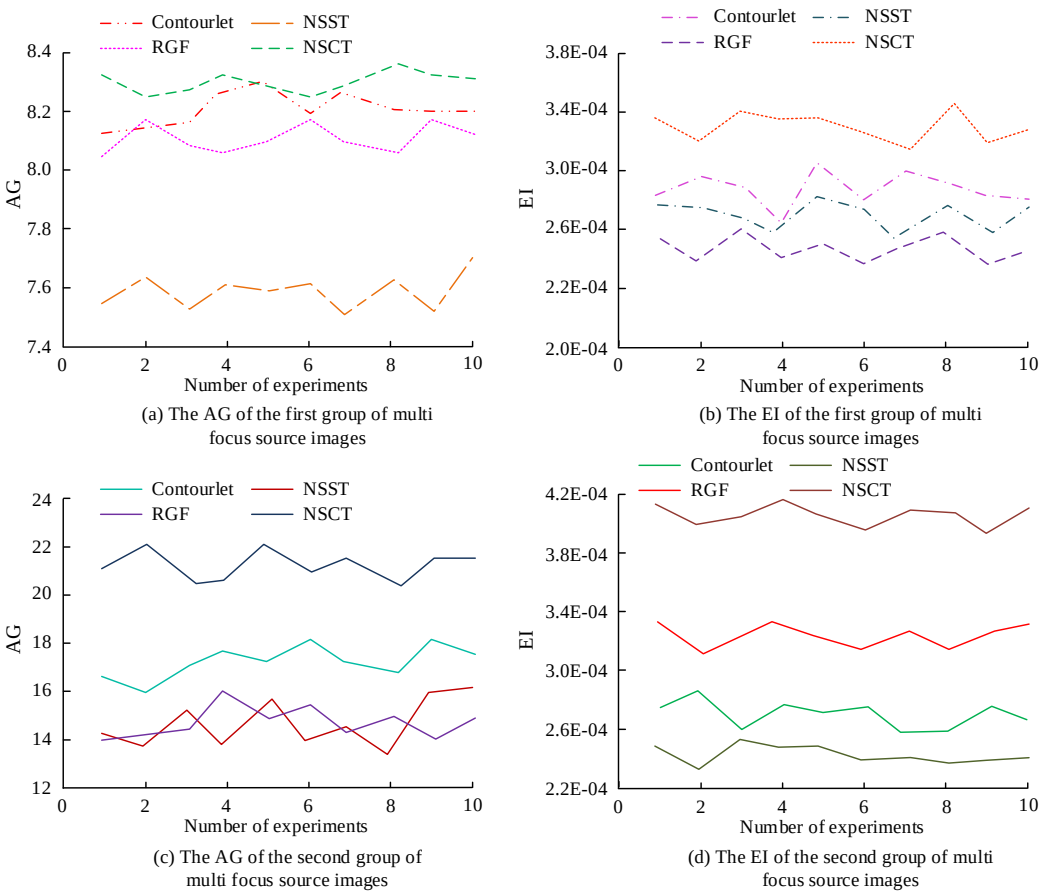


Fig. 9. AG and EI values of the fusion results as follows.

The study explored the effectiveness of the NSCT method alongside STD, SF, and VIFF validation. Fig. 10 displays the combined results of STD, SF, and VIFF values for two distinct groups. As depicted in Fig. 10(a), the NSCT method delivered STD, SF, and VIFF values of 47.69, 35.95, and 1.29 in the fusion results of the first group of multi-focus source images. This marks an enhancement over the Contourlet transform method, with increases in STD, SF, and VIFF values by 0.35, 10.48, and 0.32, respectively. Fig. 10(b) shows that in the second group's fusion results, the NSCT method achieved STD, SF, and VIFF values of 71.55, 55.73, and 1.14, respectively, indicating a marked improvement over the RGF (Recursive Gaussian Filtering) method, with reductions of 12.23, 17.28, and 0.18 in STD, SF, and VIFF values, respectively. These findings suggest that the NSCT method substantially outperforms other methods compared.

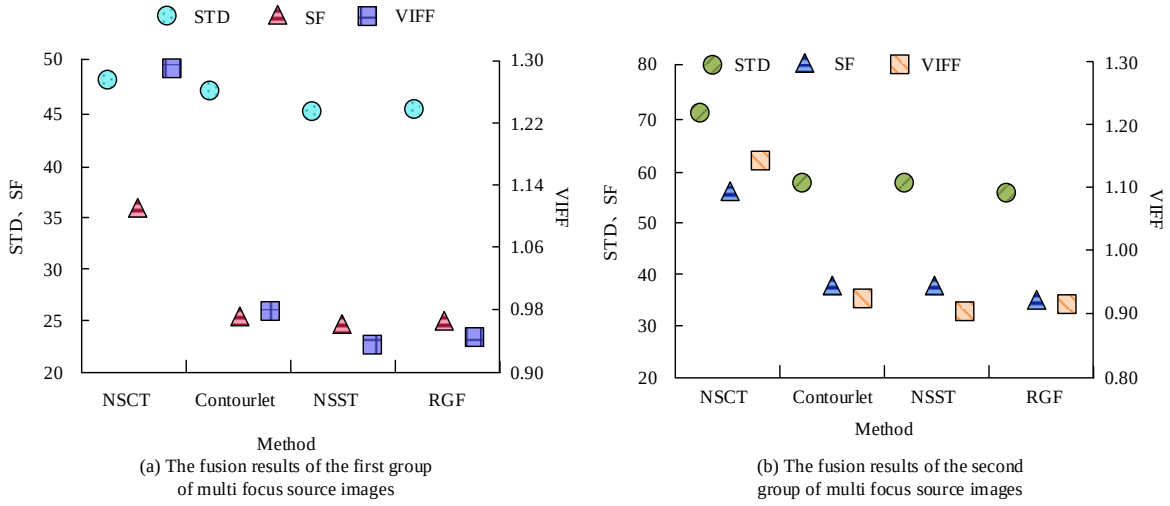
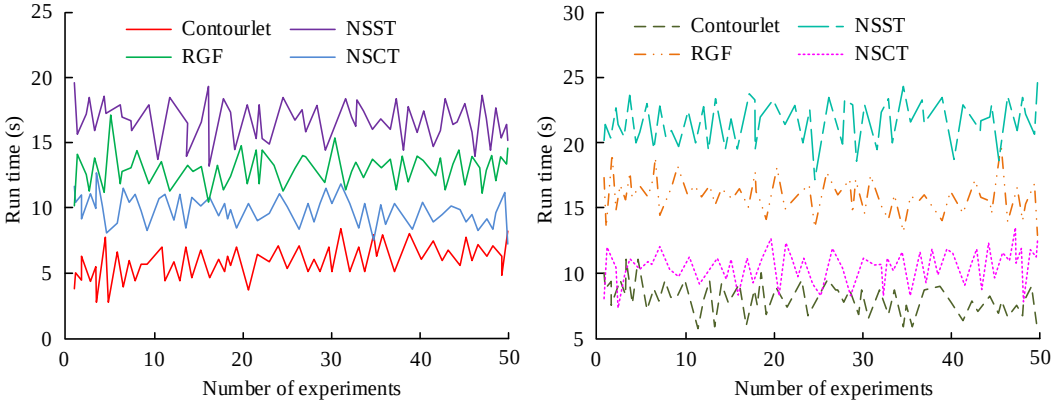


Fig. 10. STD, SF, and VIFF values for the two sets of fusion results.

The study extended its analysis to the runtime of several image fusion techniques to evaluate their computational efficiency. The experimentation encompassed 50 test executions, with each algorithm's runtime illustrated in Fig. 11. From Fig. 11(a), in the first set of multi-focus source image fusion outcomes, the NSCT method averaged a runtime of 10.03 seconds. The Contourlet transform, NSST, and RGF methods recorded average runtimes of 5.69, 17.69, and 13.72 seconds, respectively. As shown in Fig. 11(b), for the second set of fusion results, the average runtime of the NSCT method was 10.58 seconds. The rival methods averaged 8.59, 23.61, and 16.97 seconds. Notably, only the Contourlet transform method had a lower runtime than the NSCT, while the others lagged in computational efficiency. Considering fusion quality, the proposed approach in the research demonstrated superior performance.



(a) The running time of the first group of multi focus source image fusion

(b) The running time of the second group of multi focus source image fusion

Fig. 11. Running time of each algorithm.

C. Discussion

To objectively test the fusion performance in this study, we use key validation metrics such as Average Gradient (AG), Edge Intensity (EI), and computational time. We have carefully selected these metrics to comprehensively evaluate the algorithm's sharpness, edge preservation, and efficiency. Furthermore, we have assessed our NSCT-based approach against notable methods such as NSST and RGF. More results have been added that show NSCT-based fusion is better through quantitative and qualitative tests, strengthening the comparison even more. These encompass side-by-side visual comparisons of fused images, highlighting NSCT's capability to retain fine details while reducing artifacts. Furthermore, we show statistical proof that our method improves performance for both AG and EI across the chosen datasets.

The objective evaluation confirmed these results, where higher AG and EI values in NSCT meant better contrast and sharpness, which meant better fusion quality. The method's ability to get high STD, SF, and VIFF scores shows that it works to keep natural brightness and spatial coherence. This suggests that the NSCT-based method effectively gets around the problems with older methods, such as spectral aliasing and lack of directional sensitivity. Moreover, runtime analysis showed that NSCT efficiently balances fusion performance with computational cost. Although a little slower than the Contourlet Transform, NSCT performed much better than NSST and RGF, making it a good choice for real-time image processing tasks.

NSCT is better than NNSST and RGF because it eliminates problems like edge blurring, artifact creation, and slow computing. NSST has some issues with direction insensitivity and edge blurring. NSCT, on the other hand, has better direction selectivity for smoother transitions and better feature preservation. While RGF can smooth images, it often introduces pseudo artifacts and loses fine details, especially in high-texture areas. NSCT, on the other hand, uses localized filtering and multi-directional decomposition to reduce spectral aliasing. This makes artifact-free fusion with higher contrast and strong structural integrity possible. NSCT also balances fusion quality and computational efficiency better by doing fewer unnecessary calculations than NSST. This makes it a better image fusion process for real-time uses like medical imaging, surveillance, and remote sensing.

The findings highlight that the proposed NSCT-based MFIF technique offers an effective and efficient solution for multi-focus image fusion. The way it integrates images is better without slowing down computers, so it can be used in areas like medical imaging, surveillance, and satellite imagery.

V. Conclusion

Color MFIF technology has extensive applications across numerous fields requiring accurate and detailed imagery. This research introduced an advanced NSCT to overcome the shortcomings of existing fusion techniques. The results demonstrated the NSCT-based algorithm's outstanding performance in delivering superior fusion results. For the second experimental dataset, the NSCT approach showcased rich information content, vivid texture clarity, and even brightness distribution. Significantly, it prevented artifact creation and mitigated detail loss due to edge blurring. Quantitative evaluations further affirm its advantages: in the first dataset, the NSCT method recorded Standard Deviation (STD), Spatial Frequency (SF), and Visual Information Fidelity (VIFF) values of 47.69, 35.95, and 1.29, respectively, and in the second dataset, these metrics rose to 71.55, 55.73, and 1.14, notably outperforming other methods like the Contourlet

Transform, NSST, and RGF. Regarding computational efficiency, the NSCT method balanced runtime effectively. For the first dataset, it averaged a runtime of 10.03 seconds, compared to the 5.69 seconds, 17.69 seconds, and 13.72 seconds taken by the Contourlet Transform, NSST, and RGF methods, respectively. In the second dataset, the NSCT method had an average runtime of 10.58 seconds, maintaining competitiveness despite being slightly slower than the Contourlet Transform. These findings highlight the NSCT method's capacity to deliver excellent fusion quality and manageable computational demands. While the proposed approach shows strong and consistent fusion performance, this study did not consider the influence of source image quality on the fusion results. Future investigations might examine incorporating image preprocessing algorithms to enhance the quality of source images, thus further augmenting the reliability and scope of color MFIF technology.

REFERENCES

[1] Huang Y, Chen D. Image fuzzy enhancement algorithm based on contourlet transform domain. Multimedia Tools and Applications, 2020, 79(47-48): 35017-35032.

[2] Singh P, Diwakar M. Total variation-based ultrasound image despeckling using method noise thresholding in non-subsampled contourlet transform. International Journal of Imaging Systems and Technology, 2023, 33(3): 1073-1091.

[3] Vafaie S, Salajegheh E. A Comparative Study of Shearlet, Wavelet, Laplacian Pyramid, Curvelet, and Contourlet Transform to Defect Detection. Journal of Soft Computing in Civil Engineering, 2023, 7(2): 1-42.

[4] Ibrahim S I, Makhlouf M A, El-Tawel G S. Multimodal medical image fusion algorithm based on pulse coupled neural networks and nonsubsampled contourlet transform. MedicalBiological EngineeringComputing, 2023, 61(1): 155-177.

[5] Jahangir H, Khatibinia M, Kavousi M. Application of contourlet transform in damage localization and severity assessment of prestressed concrete slabs. Journal of Soft Computing in Civil Engineering, 2021, 5(2): 39-67.

[6] Li W, Lin Q, Wang K, Cai, K. Improving medical image fusion method using fuzzy entropy and nonsubsampling contourlet transform. International Journal of Imaging Systems and Technology, 2021, 31(1): 204-214.

[7] Kollem S, Reddy K R, Rao D S. Improved partial differential equation-based total variation approach to non-subsampled contourlet transform for medical image denoising. Multimedia Tools and Applications, 2021, 80(2): 2663-2689.

[8] Hasan M M, Islam N, Rahman M M. Gastrointestinal polyp detection through a fusion of contourlet transform and Neural features. Journal of King Saud University-Computer and Information Sciences, 2022, 34(3): 526-533

[9] Hu F, Cao H, Chen S, Sun, Y.,Su, Q. A robust and secure blind color image watermarking scheme based on contourlet transform and Schur decomposition. The Visual Computer, 2023, 39(10): 4573-4592.

[10] Li H, Nie R, Cao J, Guo, X., Zhou, D.,He, K. Multi-focus image fusion using u-shaped networks with a hybrid objective. IEEE Sensors Journal, 2019, 19(21): 9755-9765.

[11] Panigrahy C, Seal A, Mahato N K,Krejcar, O.,Herrera-Viedma, E. Multi-focus image fusion using fractal dimension. Applied Optics, 2020, 59(19): 5642-5655.

[12] Wang Y, Jin X, Yang J, Jiang, Q., Tang, Y., Wang, P.,Lee, S. J. Color multi-focus image fusion based on transfer learning. Journal of IntelligentFuzzy Systems, 2022, 42(3): 2083-2102.

[13] Liu S, Chen J, Rahardja S. A new multi-focus image fusion algorithm and its efficient implementation. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(5): 1374-1384.

[14] Huang Y, Hu X, Hao L, Gao, Y., Liu, Z.,Wang, P. Quantitative analysis of cell morphology based on the contourlet transform. IET Image Processing, 2020, 14(12): 2826-2832.

[15] Hasan M M, Hossain M M, Mia S, Ahammad, M. S.,Rahman, M. M. A combined approach of non-subsampled contourlet transform and convolutional neural network to detect gastrointestinal polyp. Multimedia Tools and Applications, 2022, 81(7): 9949-9968.

[16] Rahim R, Murugan S, Manikandan R, Kumar, A. Efficient Contourlet Transformation Technique for Despeckling of Polarimetric Synthetic Aperture Radar Image. Journal of Computational and Theoretical Nanoscience, 2021, 18(4): 1312-1320.

[17] Gaffar A, Joshi A B, Singh S, Srivastava, K. A high-capacity multi-image steganography technique based on golden ratio and non-subsampled contourlet transform. Multimedia Tools and Applications, 2022, 81(17): 24449-24476.

[18] Chung C H, Chen L J. Text mining for human resources competencies: Taiwan example. European Journal of Training and Development, 2021, 45(6/7): 588-602.

[19] Hasanvand M, Nooshyar M, Moharamkhani E, Selyari, A. Machine Learning Methodology for Identifying Vehicles Using Image Processing.Artificial Intelligence and Applications. 2023, 1(3): 170-178.

[20] Tijare P A, Purswani D, Rathi T, Chavhan, S., Jangid, V.,Agrawal, S. ABUSE RELATED POSTS REPORTER USING WEB CRAWLER IN PYTHON. EPRA International Journal of Research and Development (IJRD), 2022, 7(5): 230-234.

Enhanced Colon Cancer Prediction Using Capsule Networks and Autoencoder-Based Feature Selection in Histopathological Images

Janjhyam Venkata Naga Ramesh[1], F. Sheeja Mary[2], Dr. S. Balaji[3], Dr. Divya Nimma[4], Elangovan Muniyandy[5], A.Smitha Kranthi[6], Prof. Ts. Dr. Yousef A.Baker El-Ebiary[7]
Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India[1]
Adjunct Professor, Department of CSE, Graphic Era Hill University, Dehradun, 248002, India[1]
Adjunct Professor, Department of CSE, Graphic Era Deemed To Be University, Dehradun,248002, Uttarakhand, India[1]
Assistant Professor Senior Grade, Dept. of CSE, Vel Tech Rangarajan
Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India[2]
Department of CSE, Panimalar Engineering College, Chennai, India[3]
PhD in Computational Science, University of Southern Mississippi, Data Analyst in UMMC, USA[4]
Department of Biosciences, Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences, Chennai, India[5]
Applied Science Research Center. Applied Science Private University, Amman, Jordan[5]
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,
Green Fields, Vaddeswaram, AP-522302, India[6]
Faculty of Informatics and Computing, UniSZA University, Malaysia[7]

***Abstract*—The malignant development of cells in the colon or rectum is known as colon cancer, and because of its high incidence and possibility for death, it is a serious health problem. Because the disease frequently advances without symptoms in its early stages, early identification is essential. Improved survival rates and more successful therapy depend on an early and accurate diagnosis. The reliability of early detection can be impacted by problems with traditional diagnostic procedures, such as high false-positive rates, insufficient sensitivity, and inconsistent outcomes. This unique approach to colon cancer diagnosis uses autoencoder-based feature selection, capsule networks (CapsNets), and histopathology images to overcome these problems. CapsNets capture spatial hierarchies in visual input, improving pattern identification and classification accuracy. When employed for feature extraction, autoencoders reduce dimensionality, highlight important features, and eliminate noise, all of which enhance model performance. The suggested approach produced remarkable outcomes, with a 99.2% accuracy rate. The model's strong capacity to detect cancerous lesions with few mistakes is demonstrated by its high accuracy in differentiating between malignant and non-malignant tissues. This study represents a substantial development in cancer detection technology by merging autoencoders with Capsule Networks, so overcoming the shortcomings of existing approaches and offering a more dependable tool for early diagnosis. This method may improve patient outcomes, provide more individualized treatment regimens, and boost diagnostic accuracy.**

***Keywords*—Colon cancer prediction; capsule network; autoencoder; histopathological images; early cancer detection**

I. Introduction

Colon cancer is one of the common cancers globally, and also has its contribution to the cancer related deaths. Cancer of the colon is generally said to start out as small growths in the outer lining of the colon or rectum and may progress to become malignant tumors [1]. In case this condition is not diagnosed early enough and management commenced, it can develop silently and hence make the treatment process more challenging. One cannot overemphasize on the need to have colon cancer detected early. Colon cancer is relatively easy to cure and very treatable if it is detected in its preliminary stage. It is possible to detect and remove precancerous polyps and stage I malignancies with colonoscopy and other screening methods before the manifestation of symptoms. As a result, common people besides those who are over 45 years or those with the history of this complaint must opt for these tests more often. The implication of colon cancer extends to the families, the communities and the health systems over the directly affected individual. There is need to give early diagnosis and treatment since the condition leads to severe physical, emotional, and financial complications [2]. Improving possibilities for early detection of diseases and tailoring programs for patients require new discoveries in the field of medicine, for example, development of advanced techniques to visualize body conditions or new types of computational models for prognosis. Colon cancer can be anticipated employ modern methods and equipment in identifying the disease before it spreads hence improving the prospects of the ailment. Modern methodologies of prediction use numerous devices, including genetic tests, imaging, and ML algorithms. For instance, in Histopathological image analysis, the precise patterns in tissue samples may be seen by employing CapsNets in which the changes in size might mean malignant alterations. Additionally, using diversified data gathered from various sources, involving the lifestyle factors of the patient, his family history, and demographics, predictive models can be created that would assess the risk level of the given individual.

Possibly, there is a way to enhance these predictions as such methods as feature selection and dimensionality reduction with the help of autoencoders could focus on the most relevant characteristics [3].

The advancements demonstrate how sophisticated prediction approaches and individualized care may dramatically enhance early diagnosis and treatment, improving patient outcomes and streamlining healthcare processes. An important step forward in colon cancer treatment and early diagnosis is the machine learning-based prediction model that has been suggested [4]. It means that, with the help of applied ML algorithms, the scholars will look for various tendencies and the potentially linked risks associated with colon cancer, utilizing huge datasets of histopathologic data, genetic features, or patients' characteristics. Popular DL methods CNN and CapsNets are applied in the process of examining medical images to identify the signs that might suggest the presence of malignant tumors. To increase the model performance, it is also possible to apply the ML model to use multiple modalities as image findings, tests and patients' history. These models may be further refined into feature selection methods such as autoencoders to identify the pertinent data features for prediction [5]. Since these complex advancements of modern day's ML categorically predict individual responses towards various drugs, they not only enhance early diagnosis, but also facilitate development of composite patient-centered regimes. The capability of these models to predict has been on the rise and more so depending on the large and diverse information that has been fed to them perhaps yielding into better and faster response [6]. In general, the idea to integrate ML into the prediction of colon cancer can be considered as rather perspective in terms of reducing mortality rates as well as improving patients' quality of lives due to precise diagnosis.

The novel neural network design technique known as CapsNets addresses the limitations of the original CNN in terms of recognizing spatial hierarchies and connections. The constraints of CNNs in preserving the relative position and orientation of input are overcome in CapsNets by capsules, which are collections of neurons cooperating to identify certain patterns and their spatial correlations. Proprietary routing methods amongst capsules provide precise data transfer over the network. The main breakthrough is that features' presence and posture are captured by capsules, which improves the network's comprehension of intricate spatial data [7]. This hierarchical method certainly supports the idea of adjusting the CapsNets based on the changes in the object orientation, along with their ability to generalize from very little training data. In such problems as image identification and segmentation, where feature localization is critical, CapsNets are very effective because they enhance the network's ability and capacity and modeling complex spatial relations. It is an optimal method for applications requiring complex pattern recognition as they do not degrade the feature in various views and transformations. Among neural networks designed for unsupervised learning, the proper name is autoencoders. It provides mainly the aspect of learning the best representations through the encoding and decoding processes. A low dimensional representation of input data is mapped by an encoder network into an autoencoder, and the low dimensional representation is mapped by a decoder network into the original data to reconstruct it. It is useful for finding and retrieving the characteristics of the data while filtering out the outliers and the redundancies. Autoencoders are used in denoising of data, feature extraction, and dimensionality reduction. Other extensions are variational autoencoder which can model complex distributions of the data or sparse autoencoder which forces the latent space to be sparse.

Enhancing the accuracy and efficiency of the consequent image processing, two special models, namely CapsNets and autoencoders, are indisputably effective for identifying colon cancer. Spatial hierarchies were managed well and fine details in the histopathology images were recognized by CapsNets thus improving the detection of weak malignant features. On the other hand, autoencoders are useful in feature learning and feature reduction in which it down samples the important features of the images and reduces on the noises. Such integration is important to assist researchers in developing strong models for assessing medical images on the basis of time efficiency, which in this case, can lead to enhanced patient results due to better early diagnosis of colon cancer. Key contributions of the proposed work are:

1) Demonstrates increased detection accuracy for colon cancer by utilizing CapsNets' capacity to identify intricate spatial correlations, as opposed to more conventional image processing techniques.

2) Provides use of autoencoders to efficiently reduce dimensionality and extract features, producing input for the capsule network that is more insightful and pertinent.

3) Exhibits resilience in model performance over a range of datasets and imaging settings, improving the approach's generalizability.

4) Lowers false positives and false negatives in the detection of cancer, enhancing the precision and dependability of the diagnosis.

5) Permits for the customization of therapeutic and diagnostic approaches by combining patient-specific data with model predictions.

The suggested study starts with an overview of colon cancer and the urgent need for better diagnostic approaches because of the shortcomings of current practices in Section I. The Related work is reviewed in Section II, which also highlights the difficulties in early identification of the current approaches. The problem statement is described in Section III. The procedure for gathering data, the pre-processing measures, and the use of CapsNets for pattern detection and feature extraction are all covered in Section IV. The performance measures are presented in Section V and ending with the Section VI, Future work and conclusion.

II. Related Works

Ali and Ali [8] utilizes two forms of Convolutional Layers Block, the first of which is the CLB while the second one is the SCLB enhanced through an advanced computerized system to enhance the capability of identifying the lung and colon cancer. Histopathological images are processed with the help of a multi-input capsule network in this system. To undo colour distortions applied by the microscope during the preparation of histopathology slides, the SCLB undergoes the images enhanced

with multi-scale fusion, colour correction function, gamma correction, and image enhancement items. The CLB deals with raw photos in the meantime. This multi-input method certainly outlines a huge improvement as to the feature learning of the model. It turned out that when using the LC25000 dataset, the work of the model was commendable in terms of the result accuracy obtained in the diagnosis of lung and colon pathologies. The study has many limitations, however, mainly due to the fact that its source data were limited by the range of LC25000 and could not exactly reflect the characteristics of actual clinical data. This could be an effect on the model's robustness and transferability across different patient populations and histological differences.

Ahmed [9] learned about Artificial Neural Networks (ANNs), which are powerful nonlinear regression techniques have been used in colon cancer survival and classification for more than 45 years. This paper introduces fundamentals of three-layer feedforward ANN with backpropagation, which are used in cancer studies. MAS and colon cancer. In the cases where ANNs were employed in lieu of such statistical or clinicopathological approaches, there has been an overall enhancement of colon cancer classification, and survival prognosis as stated in the following discussions of the literature. Nevertheless, different types of ANNs used in biochemical research have some specific prerequisites in design and reporting to ensure the quality and credibility of obtained data. Nonetheless, the study is limited by the need for large, high-quality datasets and the possibility of over-learning, which could impact the models' generalisability across patient populations and care settings.

Kavitha et al. [10] learned about the automated processes that are essential in identifying Colorectal cancer; especially through endoscopic and histological images. This is important in as much as the enhancement of, clinical decision making and reduction of effort. Modern DL methods are workable in the detection of polyps on images and motion images, and segmentation of the latter. Image patches and CNN integration as well as the pre-processing technique are among the primary AI techniques deployed in the majority of the modern diagnostic colonoscopy stations for invasive malignancy approximation. Features like transfer learning have detached the user from the process and made even small sets yield great results hence highly accurate. Explainable deep networks that offer transparency, interpretability, consistency, and equality in the provision of healthcare are still available despite all the developments. This paper describes the recent advancements in such models and highlights the research limitations when developing technology for the prediction of colorectal cancer. However, there are still some limitations For example, one needs to have vast and diverse amount of data, and non-restrictive protocols to ensure that the model can generalize across diverse patient cases and populations.

Tasnim et al. [11] analysed that advancement in medical and Health care diagnosing has been brought about through advancement ion computer technology. Mentioning that cancer occupies the second place among all causes of death in the world, early detection is crucial for the rate of survival, especially colon cancer, which, despite its relatively high incidence and lethality, is more accessible. This paper focuses on the exploration of CNN with the imaging data of colon cells with the objective of automating cancer detection. The CNN with max pooling layers, average pooling layers, and MobileNetV2 are used in this study. The models with max pooling and average pooling achieve the accuracies of 97. 49% and 95. ResNet and MobileNet achieve mean accuracy of 48%, and 52% respectively, on the other hand MobileNetV2 achieves highest accuracy with 1 % data loss rate. 24%. While the aforementioned outcomes seem promising, the present study is still limited in several ways: the demand for large and high-quality datasets and the challenge of ensuring model robustness across different patients' groups and clinical scenarios.

Babu and Nair [12] investigated an automated detection of colon cancer using histological images which is significant for the highest possible outcome in the treatment. Traditional methods are based on low level features that are selected manually and it might not be accurate. Overall, for this problem, both supervised as well as unsupervised DCNN were applied for assessing of colon cancer histopathology images. Many of the photometric results were rotated and flipped to eliminate class disparity. From the result of the experiments the analyst was able to compare with the previous approach and found that the supervised models such as Inception were able to classify the colon cancer histopathology images with higher accuracy. Yet, a similar autoencoder network was built to extract and cluster the features from these images and to introduce the better clustering ability of the improved autoencoder network for the previously used unsupervised image processing network. Despite these advances, the study still has limitations such as the need for greater big and diverse datasets and the technical challenge of achieving model robustness and transferability across different clinical scenarios.

Schiele et al. [13] presents the Binary ImaGe Colon Metastasis classifier (BIg-CoMet), developed from the InceptionResNetV2 architecture, and operates on histologic images to partition colon cancer patients according to distant metastatic risk. Images of tumor sections stained with cytokeratin were used to train the model, along with image augmentation and dropout, to prevent overfitting. The former was investigated in a validation cohort consisting of 128 patients with BIg-CoMet showing an AUC of 0. 842, thus showing acceptable ability to distinguish between those with the metastases and those without. A marked distinction in the KM plots associated with metastasis-free survival also strongly supports the conclusion that the high-risk subjects, as defined by BIg-CoMet, have a much graver prognosis than do the other patients. This new risk variable portrayed a greater ability to perform as compared to other models with its positive predictive value standing at 80%. It depicted good results for both the subgroups of UICC and particularly for UICC III. As proven in this work, the proposed BIg-CoMet can efficiently sort out MCI or colon cancer patients based on the photographs of tumor architecture. However, the study still has its limitations in that the experiment needs to work with larger and more diverse datasets, and the inherent problem of how to ensure the stability and transferability of the model in different clinical scenarios.

Talukder et al. [14] Suggested a study of a composite feature extraction model for the classification of Lung and colon cancer. Combining deep feature extraction, ensemble learning, and

high-performance filtration techniques, it improves cancer diagnosis. It presented accuracy rates for colon cancer at 100 per cent and for rectal cancer at 99. that for the both cancers were 30 percent, and 99 percent, respectively. The performance was found to be 05% for lung cancer when tested on the LC25000 histopathology datasets. These findings indicate the higher efficiency of the suggested hybrid model compared to the existing approaches, which raises a possibility of its practical application in cancer diagnosis. The study must be validated to ensure the model is not limited to the specific cohort used or specific clinical scenarios.

CNN and hybrid ensemble approaches are examples of the sophisticated cancer detection models that have been created recently and have shown excellent accuracy in detecting lung and colon cancers. Through enhanced feature extraction and image variance management, these models overcome the drawbacks of conventional techniques. They frequently, however, rely on particular datasets, which raises questions over their generalizability across various clinical contexts and patient demographics. In order to address these drawbacks, the suggested approach combines deep learning, multi-scale fusion, and hybrid ensemble feature extraction, with the goal of improving resilience and practicality in real-world clinical settings through the use of diverse and sizable datasets for validation.

Current colon cancer detection techniques are hindered by high false-positive rates, low sensitivity, and variable results, hindering early diagnosis. Conventional CNN-based models are not good at extracting spatial hierarchies in histopathological images and tend to miss important malignant features. Feature selection methods currently used are not effective in reducing dimensionality, resulting in redundant information and computational inefficiencies. Most diagnostic methods that are available are costly, time-consuming, and need specialized skills, making them inaccessible. This research fills these gaps by combining Capsule Networks (CapsNets) with autoencoder-based feature selection, guaranteeing enhanced feature extraction, spatial hierarchy retention, and improved classification accuracy, thus providing a more accurate and economical early detection system.

III. Problem Statement

Cancer continues to be the second greatest cause of death globally, despite significant advancements in science and healthcare over the previous forty years. More over 25% of cases are lung and colon cancers, making them among the deadliest and most common tumors. Even while early diagnosis is stressed as a crucial tactic for raising survival rates, the techniques used today are frequently expensive and time-consuming [15]. There is a desperate need for the development of an automated, precise, as well as economical method to aid in the early detection as well as classification of tissue originating from lung or colon cancer. Consequently, the proposed study aims at alleviating such deficiencies in the prediction of colon cancer through autoencoder-based feature selection and CapSNets. Thus, maintaining spatial hierarchies and detecting detailed patterns, CapsNets enhance the potential of the network to identify the malignant feature that standard CNNs would ignore. The proposed project will try to increase the efficiency of colon cancer detection by the initial preprocessing of data containing autoencoders along with the decrease in the dimensions and extraction of critical features. Such an integration approach will ensure that early detection is more accurate than it is now and; therefore, the outcomes as well as the kind of treatment a patient receives in the future will be well determined.

IV. Enhanced Colon Cancer Prediction Using CapsNets and Autoencoder-Based Feature Selection

Colon cancer is quite prevalent and can be devastating, which emphasizes the significance of early identification and precise diagnosis. Since the condition is generally asymptomatic in its early stages, early detection greatly boosts the odds of effective therapy. Early detection of colon cancer depends on advanced diagnostic models and preventive screening. Efficiently evaluating complicated histopathological images, machine learning techniques like autoencoders and CapsNets might improve the accuracy of colon cancer detection. Permitting early and accurate forecasts, these cutting-edge procedures not only increase survival rates but also lessen the psychological and physical toll on patients and their families. Furthermore, it allows early and tailored therapy to be administered, maybe one that could avoid the disease from moving to more advanced stages. Thus, incorporating such advanced technologies into the system of delivering health care, it might be possible to reduce the costs of medical treatments and improve the quality of the patients' lives due to more efficient and preventive measures. Through analyzing and identifying early signs and risk factors of colon cancer, it is possible to develop creative strategies to predict and treat it, thus stressing the importance of progress in the sphere of technology concerning health care and social security.

Fig. 1 shows a methodical procedure for utilizing histopathology scans to identify lung and colon cancer. The first step of the procedure is Data Collection, during which histological images of colon and lung cancer are acquired. Data pre-processing, which entails improving the quality of the images for analysis by shrinking, normalizing, and reducing noise, comes next. The following stage is called Feature Extraction using Auto-Encoder, in which autoencoders are used to reduce dimensionality and concentrate on important patterns while identifying and extracting the most pertinent features from the pre-processed images. After that, a Capsule Network for Model Deployment is used with these extracted characteristics, taking use of the network's capacity to maintain spatial hierarchies and precisely identify intricate patterns. Ultimately, Performance Evaluation is used to evaluate the model's efficacy and make sure that the predicted accuracy and reliability fulfill the required criteria, which in turn helps with cancer early detection and diagnosis.

A. Data Collection

The Lung and Colon Cancer Histopathological Images collection is used to create and assess cancer detection algorithms as a benchmark and as a review tool. Twenty-five thousand histological photographs, divided into five categories of lung and colon tissue, are included in it. With their 768 by 768 size, these images are perfect for creating machine learning applications. The first dataset comprises 750 verified and

HIPAA-compliant photos, comprising 250 samples of benign lung tissue, 250 samples of lung adenocarcinoma, and 250 samples of lung squamous cell carcinoma. Further, five hundred samples of colon tissue were taken, for which two fifty samples of benign colon tissue were collected and two fifty samples of colon adenocarcinomas were also collected. A new set of 25,000 photos was augmented with the help of the Augmentor program that introduced rotations, flips, zooms, etc., into photos in order to mimic variability and increase the robustness of the prediction models. This was done in a bid make the dataset more interpretable and more diverse. Due to increased augmentation and better resolution of this dataset, it is highly recommended for training and validation of ML models which employ complex approach like Autoencoders and CapsNets. Using this information, the specialists can enhance the performance of methods for cancer identification, hence enhancing the health of patients. The given dataset of lung and colon cancer samples is valuable in the further work on using IT in the fight against cancer due to its well-annotated test examples and the presence of samples of both benign and malignant tumors [16].

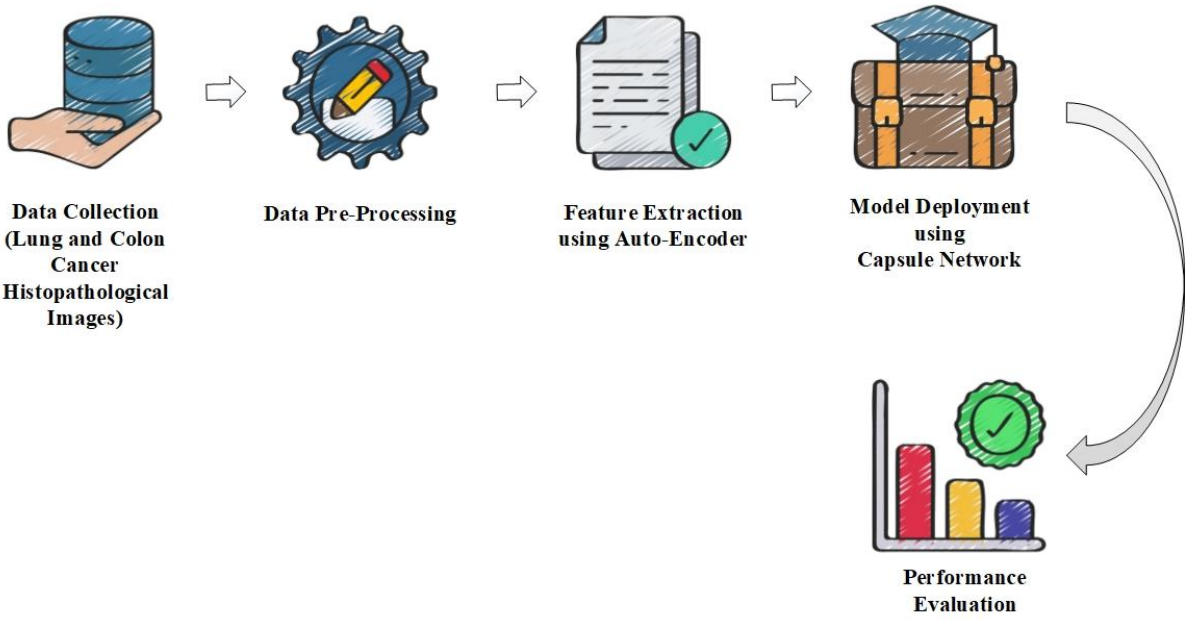


Fig. 1. Flow diagram of the proposed work.

B. Pre-Processing

It is common practice to pre-process histopathology images before delivering them to a machine learning model for examination. This entails a number of actions meant to improve the relevancy and quality of the raw photos. This procedure involves denoising as a way of reducing other unnecessary details which may probably hide other characteristics of the model, resizing of the photos to fit the model sizes and normalizing to ensure that pixel intensity levels are constant. Cleansing and normalizing, as well as pattern inclusion and non-biased, enhances the added images into the version and impacts definitely at the prediction algorithms utilized within the identity of such illnesses as lung and colon most cancers.

1) Normalization: One essential pre-processing method for getting histopathology images for ML is normalization. It is changing an images's pixel depth values to a standard scale, typically starting from 0 to 1, or to an average of zero and a general deviation of 1. Through this system, fluctuations in lighting fixtures, contrast, and coloration that can otherwise impair the model's performance are mitigated. Normalization guarantees that the model isn't impacted with the aid of unrelated elements and as a substitute concentrates at the pertinent aspects of the pics via normalizing the pixel values. In medical imaging, where constant image quality is crucial for precise analysis and diagnosis, this phase is especially crucial. Efficient normalization improves the precision and dependability of ML algorithms, resulting in improved lung and colon cancer detection and classification. Usually referred to as min-max normalization, the normalization formula is provided in (1),

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Where, the initial value of the pixel is X. The image's minimal pixel value is represented by X_{min}. The maximum pixel value in the image is represented by X_{max}. The normalized pixel value, X', will fall between 0 and 1.

C. Feature Extraction

Feature engineering is defined as data pre-processing to make it useful for the ML process and every business is now aware of the ability of ML to turn raw data into to a set of properties that can be useful in the construction of the model. Feature extraction in the case of histopathological images is about identifying and enhancing the salient features, that is, edges, textures, patterns and shapes that could indicate the presence or progression of the illness. Thereby, simplifying the data so that it can be more easily processed by computer models is the way in which the dimensionality is reduced. Thus, such methods as autoencoders might be applied to compress the image data into a lower-dimensional Latent space that maintains

the most important information. The thorough and precise feature extraction helps in enhancing the viability of the model's diagnosis and speed and its capability to categorize and predict ailments such as cancer of the lungs and colon.

1) Auto encoder: The goal of an autoencoder, a kind of artificial neural network, is to learn effective codings of input data through unsupervised learning. This is accomplished by first reconstructing the output from this representation after encoding the input into a latent-space representation. The encoder and the decoder are the two primary components of the network. The input is compressed by the encoder into a latent-space representation, which is then used by the decoder to recreate the input. Reconstruction error is to be minimized using an autoencoder in order to provide an output that closely resembles the input. This may be expressed quantitatively as reducing the mean squared error (MSE), or loss function L, between the input x and the reconstructed output xˆ in (2),

$$L(x, \hat{x}) = ||x - \hat{x}||^2 \quad (2)$$

The decoder function $x\ \hat{x} = g(h)$ maps the latent representation h back to the reconstructed input $\hat{x}$, whereas the encoder function $h = f(x)$ transfers the input x to the latent space h. The following is a summary of the complete process:

$$h = f(x) = \sigma(Wx + b) \quad (3)$$

$$\hat{x} = g(h) = \acute{o}(\acute{W}h + \acute{b}) \quad (4)$$

In Eq. (3) and (4) the weight matrices are represented by W and $\acute{W}$, the bias vectors by b and $\acute{b}$, and the activation functions by σ and $\acute{o}$. Typically, an autoencoder uses neural networks for both the encoder and the decoder, with training optimizing the weights to minimize the loss function. Autoencoders come in several designs, such as variational autoencoders (VAEs), denoising autoencoders, and sparse autoencoders, each intended for a particular use. Penalizing activations inside the hidden layers, sparse autoencoders ensure that the network learns more meaningful features by imposing sparsity restrictions on the latent space representation. This may be accomplished by including a regularization term that promotes sparsity in the loss function by (5),

$$L(x, \hat{x}) + \lambda \sum_i ||h_i||1 \quad (5)$$

Autoencoders with denoising competencies are made to address noisy information. To growth the resilience of the model, they're taught to recreate the authentic enter from a corrupted model of it. By first adding noise to the input records after which minimizing the loss characteristic between the easy enter and the output that become reconstructed from the noisy input, that is accomplished by way of (6),

$$L(x, \hat{x}) = ||x - \hat{x}_{noisy}||^2 \quad (6)$$

Fractional autoencoders (VAEs) are autoencoders with a probabilistic twist that makes them suitable for new data models. During training, a hidden signal is taken from this distribution, and the encoder outputs the parameters (mean and variance) of this distribution. Reconstruction loss along with a regularization term (KL deviation) that ensures that the reserved area distribution approximates all former distributions in the standard normal distribution of (7), forms the loss function in terms of VAEs

$$L(x, \hat{x}) = ||x - \hat{x}||^2 + KL(q(z|x)||p(z)) \quad (7)$$

In machine learning and data science, autoencoders are a vital tool because they can recognize useful representations of incoming data, facilitating streamlined and computationally efficient analysis. They are widely used for various tasks such as dimensionality reduction, anomaly detection, and generative modeling. In medical imaging, for example, it is used to extract key features from complex data, helping in tasks such as diagnosis and image reconstruction.

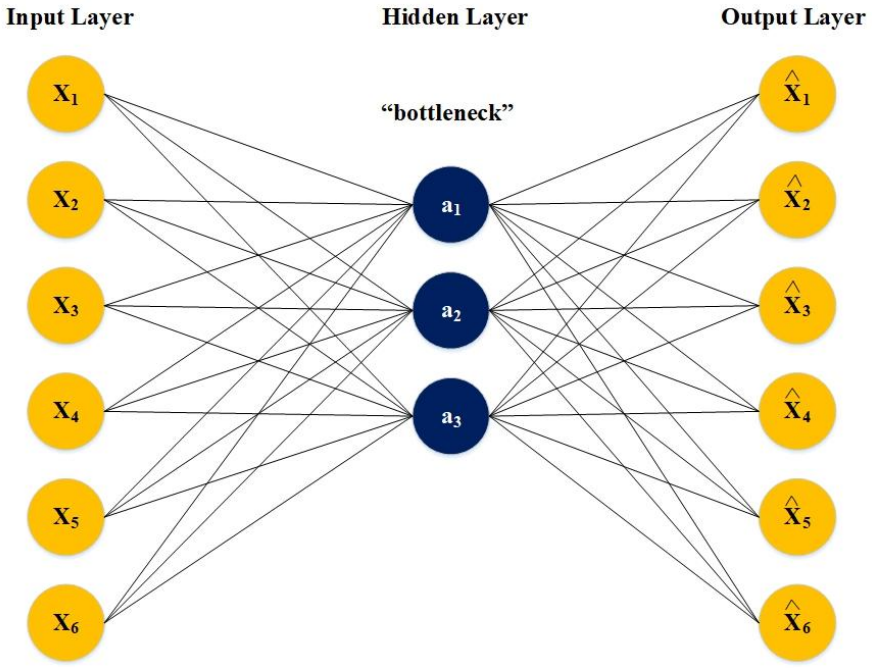


Fig. 2. Architecture of autoencoder.

Fig. 2 shows the autoencoder, a network that encodes raw input data through a series of hidden layers. The input layer receives unprocessed data, and its dimensionality is continuously reduced by the hidden layers, which identify significant patterns and characteristics. The last hidden layer, or bottleneck layer, is an indication of the input data's compressed encoding. The reconstructed output is created by the output layer, hopefully coming as near as feasible to the original data. at order to force the network to extract the most crucial features from the input data at the bottleneck layer, the autoencoder learns to reduce the reconstruction loss during training. Only the encoder portion of the autoencoder is kept after training in order to encode data that is comparable. The network is limited by regularization, denoising, short hidden layers, and activation function tuning. While adding a loss factor to the cost function encourages training in ways other than replicating the input, keeping each hidden layer as thin as feasible requires the network to take up only representative aspects of the data. Convolutional autoencoders enhance transmission and storage efficiency by reducing the dimensionality of high-dimensional image data. They are able to manage small alterations in object location or orientation and recreate missing components. Nevertheless, they have a tendency to overfit and may result in data loss, which compromises the quality of the reconstructed image. Proper regularization techniques are needed to address these issues.

The study focuses on improving colon cancer detection using CapsNets. Autoencoders play a crucial role in feature extraction from histopathological images. They reduce excessive-dimensional records to a lower-dimensional latent space, simplifying it for evaluation. The autoencoder extracts meaningful capabilities from the pics, identifying patterns, textures, and structural information indicative of cancerous and non-cancerous tissues. It also reduces noise by means of filtering out noise from histopathological pix. The autoencoder's potential to address records versions enhances the version's robustness. The autoencoder's preprocessing ensures the Capsule Network gets the maximum informative enter, improving its potential to correctly hit upon and classify cancerous tissues. This integration improves model accuracy and efficiency, leading to higher sensitivity and specificity. The autoencoder's position on this examine is to enhance early prognosis of colon most cancers through making sure specific and consultant features. Overall, the autoencoder enhances the accuracy, efficiency, and reliability of the most cancers detection model.

D. Capsule Network

A CapsNet is an artificial neural network (ANN) that imitates hierarchical connections through gaining knowledge of from the organizing concepts of biological mind systems. CapsNets are designed to mimic the hierarchical business enterprise of organic mind circuits. Basic building blocks known as tablets are used in a CapsNet to recover from regulations determined in conventional neural networks. Because tablet neurons consider each the spatial connections and the activation facts, they may be greater geared up to address changes in posture and hierarchical systems than normal neurons. Each capsule creates a collection of pose residences, consisting of orientation and position, collectively with an activation that represents a selected entity or part of an item. By enabling the network to iteratively modify the connection coefficients between them in response to the agreement of their posture parameters, capsules enable dynamic routing. Because of its ability to remember intricate spatial hierarchies and recognize subtle patterns in input, CapsNets enhance generalization. Capsules process inputs by affinely transforming the outcome into informative vectors, as opposed to neurons. Neurons function using scalars, whereas capsules use vectors. The processes involved in making artificial neurons include weighted connections, scalar activation, and sum computation. Capsules, on the other hand, undergo additional processes: input vectors are multiplied by weight matrices recorded with spatial relationships, further weight multiplication, weighted sum of input vectors, and vector output application of activation function.

1) Input vectors multiply with spatial-relationship-encoded weight matrices: The neural network's input vectors reflect the initial input or data from a previous layer. Weight matrices are multiplied across these vectors to change them. These weight matrices encode the geographical relationships within the data. When two objects are symmetrically positioned around each other and have similar dimensions, for example, the product of the input vector and weight matrix captures a high-level feature that describes this spatial arrangement. The neural network may identify and capture important correlations and features as it goes through its levels. In this instance, the weight matrix is being multiplied by the input vector.

2) Further multiplication with weights: In this phase, a capsule network's outputs from the preceding step undergo a weighted correction. While typical ANNs utilize error-based backpropagation to update weights, CapsNets use dynamic routing. The weights assigned to the synapses between neurons are determined by this unique procedure. CapsNets provide robust connections between nearby high-level and low-level capsules by dynamically changing their weights. The computation involves figuring out the precise distance between dense clusters indicating low-level capsule predictions and the outputs of the affine transform. These clusters develop and become closer together when low-level capsule predictions are comparable. As seen by the table, the high-level capsule nearest to the current prediction cluster has a bigger weight than the other capsules, which have smaller weights based on their distances.

3) Activation function application for vector output: Capsule activation functions ensure that vector outputs are dynamic and vividly represented. Squashing functions are a common choice since they preserve the vector's direction while restoring its length. The symbol for the Squashing Function is given in (8),

$$U_j = \frac{\|s_j\|^2}{1+\|s_j\|^2}\frac{s_j}{\|s_j\|} \tag{8}$$

where U_j is the output that results from applying the non-linearity function, and s_j is the sum of the input vectors. The vector s_j is compressed to a magnitude ranging from 0 to 1.

Because of this, strong hierarchical representations may be created by allowing capsules to record complex feature relationships. Because the squashing function normalizes data, it enhances resilience to changes and allows capsules to carry nuanced information necessary for complicated pattern detection in jobs like computer vision.

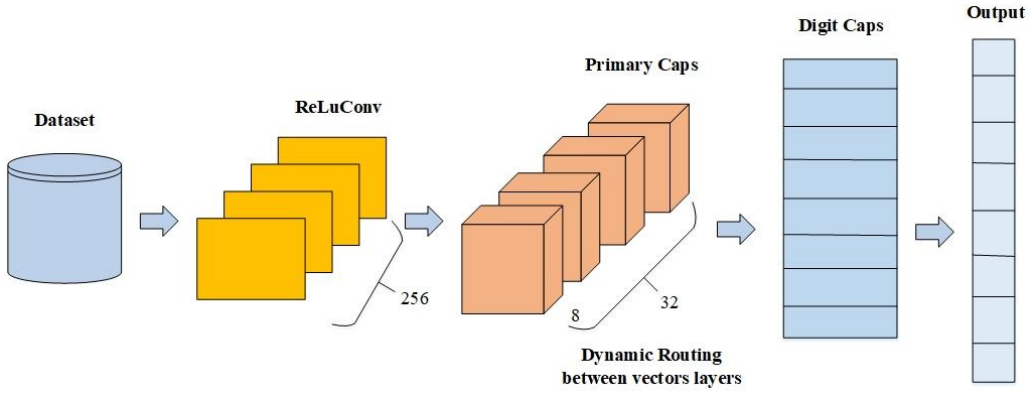


Fig. 3. Architecture of capsule networks.

Fig. 3 presents a simplified architecture of a capsule network that emphasizes the digit capsule and main layers. The input layer receives raw visual data and uses a convolutional layer to extract low-level features. After that, the main capsule layer processes the output and separates it into capsules. Each capsule produces an activation vector that indicates the existence of a particular characteristic or factor within its receptive field. The digit capsule layer, which represents a selected magnificence, is in rate of item popularity. The activation vector that every capsule produces indicates the likelihood that the class will seem in the photograph as well as its spatial connection to other components. An essential idea in CapsNets is dynamic routing, in which the primary pill layer casts votes to determine which digit pill it most carefully matches. This strengthens the settlement among compatible capsules and weakens the connections among incompatible drugs. Shooting the spatial hierarchy among functions, CapsNets are a type of neural community design that improves getting to know approximately representations. Because of dynamic routing, which makes it less complicated to attain an agreement on instantiation settings, they frequently require less information augmentation than conventional CNNs. Because CapsNets constitute vectors and routing systems, they're also more proof against adverse attacks. Additionally, they incorporate posture statistics, which complements structured representation for obligations like location estimation and item reputation.

CapsNets do have several drawbacks, though, including a lack of empirical assist, computational complexity, a probable tendency to overemphasize capsules, and intrinsic complexity. Given their latest age, CapsNets have not gone through calla lot checking out. Further research and evaluation are required to illustrate their typical effectiveness. Furthermore, due of dynamic routing, CapsNets may additionally require more processing power, main to longer training periods and better resource wishes. Applications for CapsNets may be observed in numerous domain names, which includes as clinical imaging, self-reliant vehicles, and visual anomaly detection. They are helpful in scientific imaging, assisting with troubles like organ segmentation and tumor identity. They also are useful in photograph popularity, item detection, region estimation, cybersecurity, and self-riding cars. Nevertheless, extra research and comparison are required to validate their ordinary effectiveness in diverse assignments.

The accuracy and robustness of histopathological image type are greatly improved with the aid of CapsNets inside the proposed study on boosting colon most cancers prediction the usage of CapsNets and Autoencoder-based totally function choice. CapsNets are particularly properly at retaining spatial hierarchies and connections within the image facts, which is vital for effectively recognizing difficult patterns that may be signs of malignant tumors. In evaluation to traditional CNN, which could have trouble processing orientation and attitude changes, CapsNets make better use of clusters of neurons called pills to capture and encode these spatial connections. As a result, they are able to discover characteristics at diverse abstraction stages and offer a greater complex interpretation of the photo statistics. The proposed examine makes use of CapsNets in conjunction with Autoencoders to extract features. This permits the Autoencoder to lessen dimensionality and emphasize pertinent features, whilst additionally utilising CapsNet's higher spatial awareness and sample popularity capabilities. This mixture improves the model's predictive potential, which might result in a greater specific and trustworthy categorization of histopathological photos for the identity of colon most cancers.

Algorithm 1: Algorithm for the Proposed study

Step 1: Data Collection and Preparation

- Load dataset of histopathological images
- Preprocess images

Step 2: Auto encoder for Feature Extraction

- Define Autoencoder architecture
  a. Input layer: X
  b. Encoding layers: progressively reduce dimensions
  c. Bottleneck layer: Z
  d. Decoding layers: progressively reconstruct dimensions
  e. Output layers: X’
- Split dataset into training and testing sets
- Train Autoencoder using mean squared error loss

- Extract compressed features Z from the bottleneck layer

Step 3: Capsule Network for classification

- Define Capsule network architecture
  a. Input layer: Z
  b. Convoutional Layer: extract local feature
  c. Primary Capsule Layer: convert features into capsules
  d. Digital Capsule Layer: from higher-level capsules
  e. Output layer: Class probabilities
- Combine Z with original image data
- Split combined dataset into training and testing sets
- Train capsule network using margin loss

Step 4: Integration

- Integrate Autoencoder and capsule network by feeding Z into the capsule network
- Fine-tune integrated model on training data

Step 5: Evaluation

- Validation and Testing
- Calculate Performance Metrics

V. Result and Discussion

The results of the advanced colon cancer prediction approach, which combines CapsNets with autoencoder-based feature selection, are shown in this section. The results show that this hybrid approach greatly improves the performance and reliability of the classification when compared to traditional techniques. The model makes the evaluation of histopathology images more reliable and consistent using the Autoencoder-based feature extraction with the help of the Capsule Network for determining the complex spatial relations. More details concerning the performance measures, benefits as well as the demerits of the model and how such developments aids in the early detection of colon cancer are demonstrated in this segment.

E. Training and Testing

The Fig. 4 shows the training and testing accuracy of a model over 100 epochs. The X-axis represents the number of training iterations the model has undergone, while the Y-axis represents the accuracy percentage. The figure shows the model's training accuracy on the training dataset and its testing accuracy on the testing dataset. The model's initial phase (0-20 epochs) shows rapid growth in both accuracies, indicating learning and improving performance on both datasets. The middle phase (20-60 epochs) shows slower growth in training accuracy, approaching a plateau around 60 epochs. Testing accuracy additionally improves but starts to lag behind the training accuracy, suggesting overfitting. The later phase (60-a hundred epochs) indicates an excessive schooling accuracy close to 100%, indicating superb performance on the schooling statistics. However, the trying out accuracy stabilizes at 85.9%, indicating overfitting. To cope with overfitting, techniques which include early stopping, regularization, or pass-validation could be implemented. The model's testing accuracy stabilizes at a high stage, indicating top overall performance, but there may be room for development in generalization. To deal with overfitting, techniques which include early preventing, regularization, or go-validation can be applied.

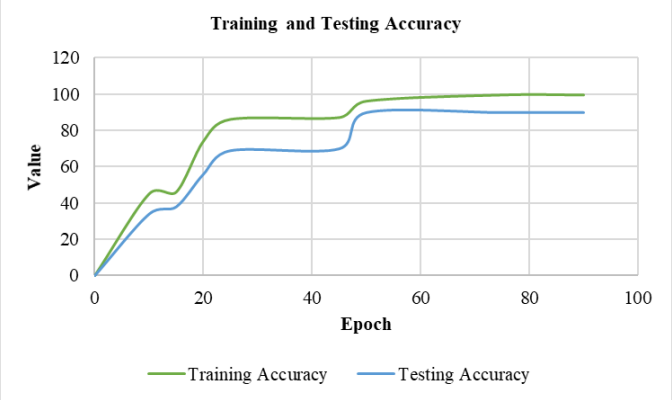


Fig. 4. Training and testing accuracy

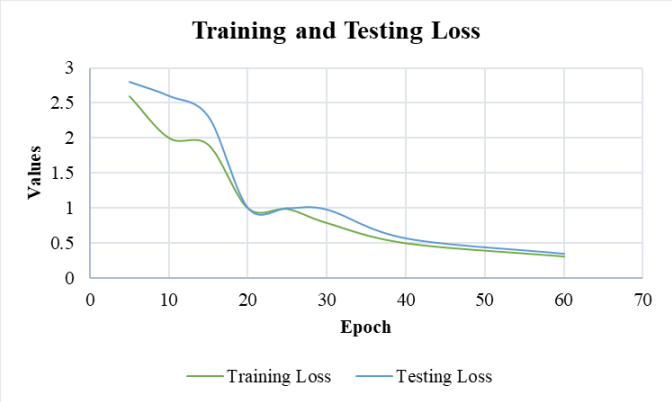


Fig. 5. Training and testing loss.

The Fig. 5 shows the training and testing loss of a version over 60 epochs. The X-axis represents the variety of schooling iterations, at the same time as the Y-axis represents the loss cost, which measures the error between anticipated and real values. The figure shows the lack of the model at the dataset and the loss at the testing dataset. The analysis shows that the version is getting to know efficiently and both training and trying out losses are decreasing, indicating top generalization overall performance. The initial phase (0-20 epochs) shows fast decreases in both losses, at the same time as the center phase (20-40 epochs) indicates a gradual lower in schooling loss and a slow lower in testing loss, indicating accurate generalization but signs of overfitting. The later phase (40-60 epochs) shows a solid and convergent loss, suggesting that the model has reached a superior point wherein further schooling does no longer significantly improve performance or motive overfitting. The graph concludes that the model is nicely-trained, achieving low mistakes charges on each schooling and trying out datasets. The near convergence of training and checking out loss in later epochs shows that the model isn't overfitting notably, maintaining true generalization performance. The stability and convergence of loss values in later epochs recommend an awesome balance among bias and variance, minimizing underfitting and overfitting.

F. Performance Metrics

Performance metrics are numerical measurements which are used to evaluate how a model or gadget plays in achieving its goal. These measures, which are relevant to ML and diagnostic

model, consist of F1 rating, accuracy, precision, and recall. TP as true positive, TN as true negative, FP as false positive, and FN as false negative are represented.

1) Accuracy: A performance statistic called accuracy counts how many of a model's predictions are accurate out of all the predictions it has made. It is computed in (9),

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

2) Precision: A performance indicator called precision counts the percentage of accurate positive predictions among all the positive predictions a model makes. It is computed in (10),

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

3) Recall: Recall quantifies the percentage of real positive cases that a model accurately detects; it is sometimes referred to as sensitivity or true positive rate. It is computed in (11),

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

4) F1 score: The F1 score is a performance statistic that offers a fair assessment of a model's accuracy by combining recall and precision into a single number. This provides a more thorough understanding of a model's performance, particularly when working with unbalanced datasets. It is the harmonic mean of accuracy and recall. The F1 score is calculated in (12),

$$F1 = \frac{Precision.Recall}{Precion+Recall} \quad (12)$$

Table I show that colon most cancers prediction with CapsNets and Autoencoder-based feature selection has extraordinary model performance. The model efficaciously classifies 99.2% of the cases with an accuracy of 99.2%, indicating its efficacy in distinguishing among samples which are malignant and people that aren't. The model's 99% accuracy suggests that it could reliably pick out affirmative conditions with few fake positives. The model's capacity to minimize fake negatives while catching the majority of real occasions is demonstrated with the aid of its 98.3% recall. The alternate-off among accuracy and reminiscence is balanced by means of the F1 score, that is 98.6% and represents the harmonic suggest of precision and recall. The resilience and dependability of the model in processing histopathological images are highlighted by those sturdy metrics. Specifically, the accuracy validates the effectiveness of merging Autoencoders with CapsNets, which effects the cautioned observe. It attests to the model's capability to both recognize the complex styles seen in the education set and generalize successfully to new units of records. Because of its brilliant accuracy, the version may be carried out nearly in clinical settings and can assist identify colon cancer early on, growing patient consequences and remedy options.

TABLE I. PERFORMANCE METRICS OF THE PROPOSED STUDY

Metrics	Efficiency
Accuracy	99.2%
Precision	99%
Recall	98.3%
F1 score	98.6%

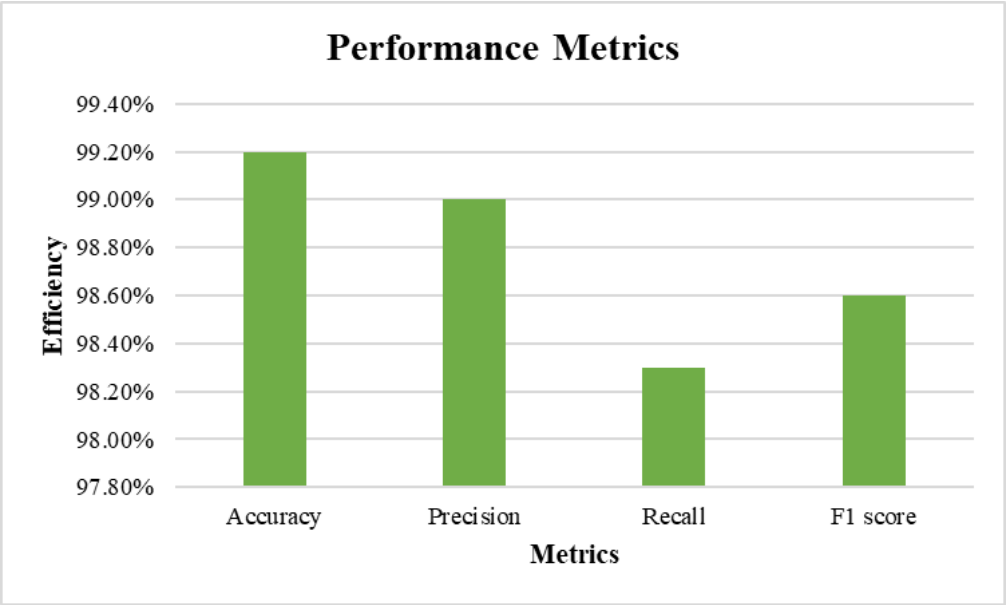


Fig. 6. Performance metrics of the proposed study.

The accuracy, precision, recall, and F1 score of the suggested colon cancer prediction model are highlighted in the Fig. 6 that displays its performance measures. At around 99.2%, the accuracy is the greatest, meaning that 99.2% of all cases are accurately classified by the model, demonstrating its overall efficacy. With 99% precision, only 99% of cases accurately categorized as positive are actually positive, reducing the number of false positives. With a little reduced recall of 98.3%, the model reduces the amount of false negatives by capturing 98.3% of all real positive cases. The F1 score, which is 98.6% and represents the harmonic mean of precision and recall, is a balanced measure of the model's accuracy that accounts for both precision and recall. Taken together, these parameters demonstrate the model's accuracy in colon cancer prediction from histopathological images as well as its dependability. The model's excellent accuracy confirms that it can generalize from training information to unseen test data, which is necessary for its application in real-world clinical settings. In clinical circumstances, it might be advantageous to prevent needless treatments by emphasizing the reduction of false positives over the capture of all real positives, as indicated by the minor decline in recall when compared to precision. The model's well-balanced performance is shown by the F1 score's proximity to both accuracy and recall. The model is robust in learning and identifying complex patterns in histopathological images, and it is also dependable in making accurate predictions, which may lead to improved early detection and diagnosis of colon cancer. This is indicated by the high accuracy and balanced precision-recall performance. Better patient outcomes and more effective clinical decision-making procedures might result from this.

The efficacy metrics of many colon cancer detection techniques are displayed in Table II. With 85% accuracy, 80% precision, 75% recall, and 77% F1 score, logistic regression offers a basic method but falls short in terms of sensitivity and overall effectiveness. Decision trees perform better but still fall short when compared to more advanced techniques, with 87% accuracy, 83% precision, 80% recall, and an 81% F1 score. Results are further improved by Random Forests, which show good overall performance with 90% accuracy, 87% precision, 85% recall, and an 86% F1 score. Gradient Boosting Machines show notable advances in identifying and categorizing malignant cells, achieving superior performance metrics with 93% accuracy, 90% precision, 90% recall, and a 91% F1 score. On the other hand, the suggested approach, which combines autoencoder-based feature selection with CapsNets, yields remarkable outcomes with 98.6% F1 score, 99% precision, 98.3% recall, and 99.2% accuracy. This strategy uses cutting-edge ML algorithms to improve feature extraction and pattern recognition, outperforming all other approaches and demonstrating its improved capacity to consistently and effectively diagnose colon cancer.

TABLE II. COMPARISON OF PROPOSED METHOD WITH DIFFERENT METHODS

Method	Accuracy	Precision	Recall	F1 Score
Logistic Regression [17]	85%	80%	75%	77%
Decision Trees [18]	87%	83%	80%	81%
Random Forest [19]	90%	87%	85%	86%
Gradient Boosting Machines [20]	93%	90%	90%	91%
Proposed Method	99.2%	99%	98.3%	98.6%

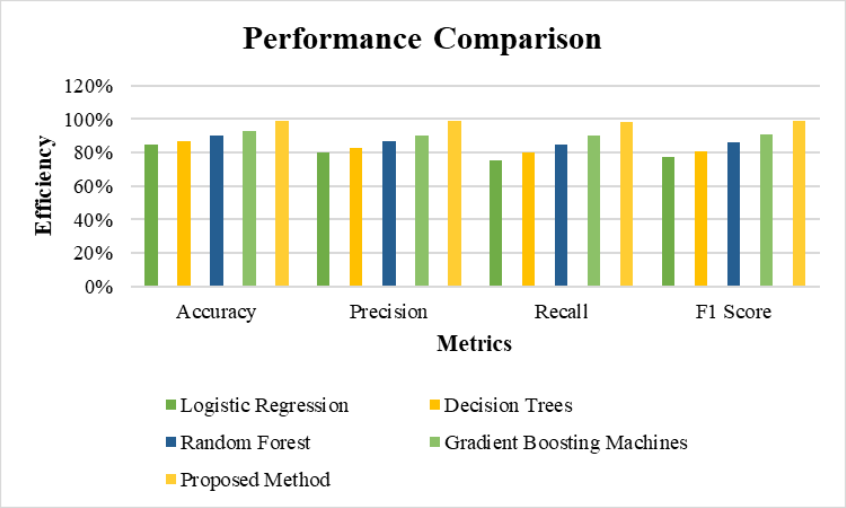


Fig. 7. Performance comparison of the proposed method with different methods.

Fig. 7 showcases a performance comparison of ML algorithms, including Logistic Regression, Decision Trees, Random Forest, Gradient Boosting Machines, and the Proposed Method. The x-axis displays metrics like accuracy, precision, recall, and F1 Score. The Proposed Method have the highest accuracy, precision, recall, and F1 Score and consistently performs well across all metrics, while Random Forest also shows strong performance in most metrics. However, it is challenging to provide a definitive interpretation without specific values and context. The visualization suggests the Proposed Method might be a promising approach, but further analysis and understanding of the data are necessary. The exact

meaning of Efficiency on the y-axis is unclear, and the dataset and problem domain used for this comparison are unknown, limiting broader conclusions.

TABLE III. COMPARISON OF PROPOSED DATASET WITH DIFFERENT DATASETS

Dataset	Accuracy	Precision	Recall	F1 Score
Gapminder Colon cancer [21]	85%	80%	75%	77%
Colon cancer [22]	87%	83%	80%	81%
Multi Cancer Dataset [23]	90%	87%	85%	86%
ICMR Dataset [24]	93%	90%	90%	91%
Lung and Colon Cancer Histopathological Images	99.2%	99%	98.3%	98.6%

The Table III compares many datasets that have been used to assess alternative approaches to the diagnosis of colon cancer. An F1 score of 77%, recall of 75%, accuracy of 85%, and precision of 80% were attained using the Gapminder Colon cancer dataset. With an F1 score of 81%, accuracy of 87%, precision of 83%, and recall of 80%, the Colon cancer dataset demonstrated increased performance. The metrics were further improved by the Multi Cancer Dataset, which achieved an F1 score of 86%, 90% accuracy, 87% precision, and 85% recall. With an F1 score of 91%, accuracy of 93%, precision of 90%, recall of 90%, and recall of 90%, the ICMR Dataset performed even better. With an amazing accuracy of 99.2%, precision of 99%, recall of 98.3%, and an F1 score of 98.6%, the suggested technique surpassed all other methods when evaluated on the Lung and Colon Cancer Histopathological Images dataset, demonstrating its better capabilities in diagnosing colon cancer.

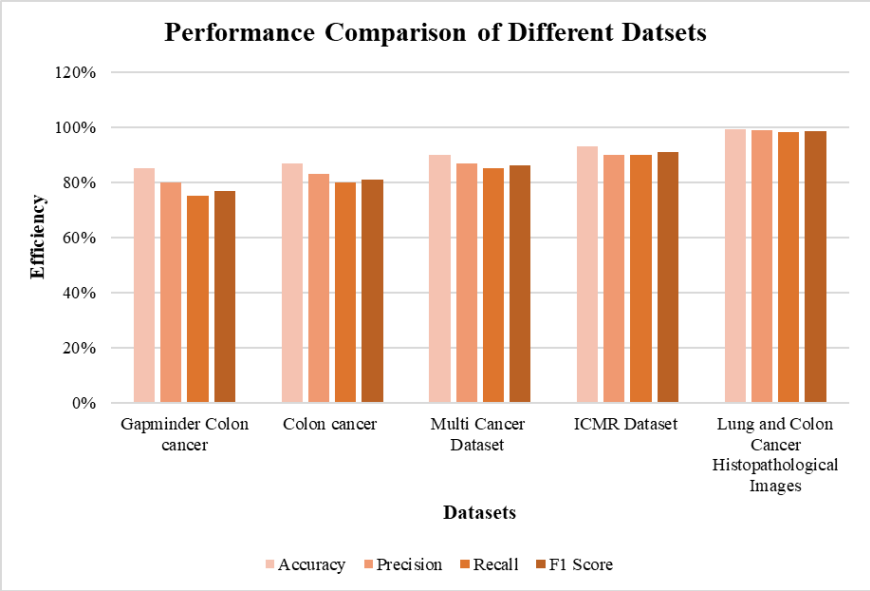


Fig. 8. Performance comparison of the proposed dataset with different datasets.

Accuracy, precision, recall, and F1 score are the four main metrics that are highlighted in the Fig. 8, which presents a visual comparison of the performance of different datasets used in colon cancer diagnosis. The least successful dataset is the Gapminder Colon Cancer dataset, which has an F1 score of 77%, recall of 75%, accuracy of 85%, and precision of 80%. With the Colon cancer dataset, performance somewhat increases to 87% accuracy, 83% precision, 80% recall, and 81% F1 score. These measures are further improved by the Multi Cancer Dataset, which achieves an F1 score of 86%, 90% accuracy, 87% precision, and 85% recall. With 91% F1 score, 90% precision, 90% recall, and 93% accuracy, the ICMR Dataset shows even better efficiency. With accuracy at 99.2%, precision at 99%, recall at 98.3%, and an F1 score of 98.6%, the suggested method which makes use of the Lung and Colon Cancer Histopathological Images dataset achieves the best results in terms of all metrics, demonstrating its better capacity to identify cancer accurately.

G. Discussion

Previous studies of colon cancer detection often struggled with the accuracy and reliability of their models. Although, traditional methods such as decision trees and logistic regression had stability and accuracy problems that increased the number of false negatives and positives Strong models needed for the accuracy of the analysis cannot be fully captured and often requires significant feature engineering [25]. The proposed method uses auto-encoder-based feature selection combined with CapsNets to address these limitations. The ability of CapsNets to preserve spatial structure and detect small patterns is essential for accurate cancer diagnosis. On the other hand, autoencoders enhance feature extraction by reducing noise and compressing high-dimensional data into a feasible hidden area that highlights some important features and most of them were captured there with significant improvements occur in accuracy, precision, recall, and F1 scores It occurs as analyzed hereafter. This approach addresses and improves the weaknesses of

existing methods, thereby providing more accurate and rapid detection of colorectal cancer that can change clinical diagnosis.

VI. Conclusion and Future Work

Research suggests that the combination of autoencoder-based feature selection and CapsNets might greatly improve the accuracy of colon cancer prediction using histopathological images. Because of the CapsNets' spatial hierarchy learning and autoencoders' strong feature extraction capabilities, this model is more successful in correctly identifying malignant tissues. Using a hybrid technique, frequent problems like overfitting are successfully reduced, classification accuracy is increased, and the model's ability to generalize from training to new data is encouraged. The findings imply that this sophisticated machine learning model has great promise for clinical use, where accurate and prompt diagnosis of colon cancer is essential for efficient treatment planning and better patient outcomes. This effective use of contemporary deep learning architectures highlights how various methods may be used to address difficult classification problems in medical image analysis.

Improving the model's flexibility and reactivity to various clinical settings should be the main goal of future research. This involves looking at cutting-edge data augmentation methods to strengthen the model's resistance to changes in the quality and quantity of histopathology images. Furthermore, investigating the incorporation of additional datasets, such genetic data or patient medical history, may provide a more thorough method of cancer diagnosis and prognosis. To fully assess the model's performance and potential, it is also crucial to apply it in actual clinical situations. Extending the study's reach will showcase the adaptability and usefulness of the concept. Lastly, to guarantee that the model satisfies clinical requirements and keeps improving patient outcomes and care, continuous cooperation with healthcare professionals is crucial.

References

[1] R. Labianca et al., “Colon cancer,” Critical reviews in oncology/hematology, vol. 74, no. 2, pp. 106–133, 2010.

[2] A. B. Benson et al., “Colon cancer,” Journal of the National Comprehensive Cancer Network, vol. 9, no. 11, pp. 1238–1290, 2011.

[3] P. F. Engstrom et al., “Colon cancer,” Journal of the National Comprehensive Cancer Network, vol. 7, no. 8, pp. 778–831, 2009.

[4] P. Greenwald, “Colon cancer overview,” Cancer, vol. 70, no. S3, pp. 1206–1215, 1992.

[5] I. M. Oving and H. C. Clevers, “Molecular causes of colon cancer,” European journal of clinical investigation, vol. 32, no. 6, pp. 448–457, 2002.

[6] J. Terzić, S. Grivennikov, E. Karin, and M. Karin, “Inflammation and colon cancer,” Gastroenterology, vol. 138, no. 6, pp. 2101–2114, 2010.

[7] S. D. Markowitz, D. M. Dawson, J. Willis, and J. K. Willson, “Focus on colon cancer,” Cancer cell, vol. 1, no. 3, pp. 233–236, 2002.

[8] M. Ali and R. Ali, “Multi-input dual-stream capsule network for improved lung and colon cancer classification,” Diagnostics, vol. 11, no. 8, p. 1485, 2021.

[9] F. E. Ahmed, “Artificial neural networks for diagnosis and survival prediction in colon cancer,” Molecular cancer, vol. 4, pp. 1–12, 2005.

[10] M. S. Kavitha, P. Gangadaran, A. Jackson, B. A. Venmathi Maran, T. Kurita, and B.-C. Ahn, “Deep neural network models for colon cancer screening,” Cancers, vol. 14, no. 15, p. 3707, 2022.

[11] Z. Tasnim et al., “Deep learning predictive model for colon cancer patient using CNN-based classification,” International Journal of Advanced Computer Science and Applications, vol. 12, no. 8, pp. 687–696, 2021.

[12] T. Babu and R. R. Nair, “Colon cancer prediction with transfer learning and k-means clustering,” in Frontiers of ICT in Healthcare: Proceedings of EAIT 2022, Springer, 2023, pp. 191–200.

[13] S. Schiele et al., “Deep learning prediction of metastasis in locally advanced colon cancer using binary histologic tumor images,” Cancers, vol. 13, no. 9, p. 2074, 2021.

[14] M. A. Talukder, M. M. Islam, M. A. Uddin, A. Akhter, K. F. Hasan, and M. A. Moni, “Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning,” Expert Systems with Applications, vol. 205, p. 117695, 2022.

[15] M. Masud, N. Sikder, A.-A. Nahid, A. K. Bairagi, and M. A. AlZain, “A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework,” Sensors, vol. 21, no. 3, p. 748, 2021.

[16] “Lung and Colon Cancer Histopathological Images.” Accessed: Aug. 06, 2024. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images

[17] G. Leonard et al., “Machine learning improves prediction over logistic regression on resected colon cancer patients,” Journal of Surgical Research, vol. 275, pp. 181–193, 2022.

[18] M. Vidya Bhargavi, V. R. Mudunuru, and S. Veeramachaneni, “Colon cancer stage classification using decision trees,” in Data Engineering and Communication Technology: Proceedings of 3rd ICDECT-2K19, Springer, 2020, pp. 599–609.

[19] Z. Yan, J. Li, Y. Xiong, W. Xu, and G. Zheng, “Identification of candidate colon cancer biomarkers by applying a random forest approach on microarray data,” Oncology reports, vol. 28, no. 3, pp. 1036–1042, 2012.

[20] A. Hage Chehade, N. Abdallah, J.-M. Marion, M. Oueidat, and P. Chauvet, “Lung and colon cancer classification using medical imaging: A feature engineering approach,” Physical and Engineering Sciences in Medicine, vol. 45, no. 3, pp. 729–746, 2022.

[21] “Gapminder Colon cancer.” Accessed: Aug. 08, 2024. [Online]. Available: https://www.kaggle.com/datasets/nancyalaswad90/gapminder-colon-cancer

[22] “Colon cancer.” Accessed: Aug. 08, 2024. [Online]. Available: https://www.kaggle.com/datasets/angevalli/colon-cancer

[23] “Multi Cancer Dataset.” Accessed: Aug. 08, 2024. [Online]. Available: https://www.kaggle.com/datasets/obulisainaren/multi-cancer

[24] “ICMR Dataset.” Accessed: Aug. 08, 2024. [Online]. Available: https://www.kaggle.com/datasets/shibumohapatra/icmr-data

[25] C.-G. Cheng, Y.-M. Tian, and W.-Y. Jin, “A study on the early detection of colon cancer using the methods of wavelet feature extraction and SVM classifications of FTIR,” Journal of Spectroscopy, vol. 22, no. 5, pp. 397–404, 2008.

Revolutionizing AI Governance: Addressing Bias and Ensuring Accountability Through the Holistic AI Governance Framework

Ibrahim Atoum
Department of Artificial Intelligence-Faculty of Science and Information Technology,
Al-Zaytoonah University of Jordan, Amman, 11733, Jordan

***Abstract*—Artificial intelligence (AI) possesses the capacity to transform numerous facets of our existence; however, it concomitantly engenders considerable risks associated with bias and discrimination. This article explores emerging technologies like Explainable AI (XAI), Fairness Metrics (FMs), and Adversarial Learning (AL) for bias mitigation while emphasizing the critical role of transparency, accountability, and continuous monitoring and evaluation in AI governance. The Holistic AI Governance Framework (HAGF) is introduced, featuring a comprehensive, five-layered structure that integrates top-down and bottom-up strategies. HAGF prioritizes foundational principles and resource allocation, outlining five lifecycle-specific phases. Unlike the OECD AI Principles, which offer a general ethical framework lacking holistic perspective and resource allocation guidance, and the Berkman Klein Center's Model, which provides a broad framework but omits resource allocation and detailed implementation, HAGF offers actionable mechanisms. Tailored Key Performance Indicators (KPIs) are proposed for each HAGF layer, enabling ongoing refinement and adaptation to the evolving AI landscape. While acknowledging the need for enhancements in data governance and enforcement, the embedded KPIs ensure accountability and transparency, positioning HAGF as a pivotal framework for navigating the complexities of ethical AI.**

***Keywords*—Artificial intelligence; framework; bias; discrimination; governance; key performance indicators**

I. Introduction

The iterative process of AI model development is crucial for enabling machines to perform human-like tasks such as problem-solving, learning, decision-making, and perception. This process enhances technological systems by developing algorithms that process large datasets, recognize patterns, and make predictions. A critical step is training, which uses labeled data to adjust model parameters and optimize performance. Therefore, the quality and quantity of data are vital for success [1]. Accurate and representative training data directly impacts model performance; biased data can lead to errors, particularly in critical fields like healthcare, finance, and criminal justice [2].

In AI, bias refers to systematic prediction distortions arising from training data and model design, while discrimination involves unfair treatment based on characteristics such as age, race, or gender. Although bias can contribute to discrimination, they are not synonymous; discrimination can occur independently of bias, and reducing bias may not eliminate discrimination [3]. Data labeling and algorithm design biases can perpetuate discriminatory patterns, and a lack of diversity in development teams can hinder identifying and mitigating these biases.

To promote fairness and equity in AI, it is essential to ensure diverse and representative training data, fair feature selection, rigorous evaluation strategies, and effective governance frameworks. These frameworks should include policies and guidelines for safe and ethical AI development, emphasizing auditing, transparency, and stakeholder collaboration [4]. Data augmentation and synthetic data generation are necessary to ensure data accuracy and representativeness, as high-quality data is crucial for avoiding biased AI models.

Four key components support responsible AI development: XAI, FMs, AL, and ongoing M&E [5]. While XAI enhances transparency, it faces challenges such as computational costs. FMs may conflict with accuracy, necessitating a holistic approach. AL can mitigate bias through adversarial examples, and continuous M&E is vital for detecting biases over time. Integrating these elements is essential for developing fair and accountable AI systems.

Governments play a critical role in promoting AI's safe and responsible implementation [6]. They can establish ethical guidelines, engage stakeholders, and enforce principles addressing bias, privacy, safety, and security. Regulation in sensitive areas like healthcare and criminal justice ensures adherence to ethical standards. Governments can also promote industry self-regulation through voluntary codes of conduct and support international cooperation to establish common standards.

This paper compares the Holistic AI Governance Framework (HAGF) with the OECD AI Principles [7] and the Berkman Klein Center's Model for AI Governance [8], focusing on the HAGF's unique contribution to AI governance. The HAGF offers a holistic, cyclical approach to resource allocation, structured around five interconnected components: establishing standards, regulating adherence, promoting AI advancement, fostering transparency and accountability, and encouraging a voluntary code of conduct. Each component includes defined roles and theoretical weights, reflecting a vision for responsible AI development and providing enhanced strategic guidance. The HAGF emphasizes the need for formal

guidelines and policies to ensure ethical AI deployment and effective resource allocation for research and development.

While the OECD AI Principles provide foundational ethical guidelines, they have been criticized for lacking a holistic approach to implementation and sufficient guidance on resource allocation, leading to potentially inconsistent interpretations and inadequate stakeholder engagement [7]. Similarly, though comprehensive, the Berkman Klein Center's Model is criticized for its breadth and lack of specific implementation measures and clear metrics for evaluating effectiveness [8]. The HAGF addresses these shortcomings by focusing on resource allocation and integrating top-down and bottom-up strategies to foster ethical AI practices. Furthermore, this paper examines key technical approaches to mitigating bias and enhancing ethical AI development, proposing KPIs for each phase to monitor effectiveness and facilitate continuous improvement.

Section II presents related work, Section III outlines the importance of gathering diverse data, Section IV explores methods for identifying bias in AI, Section V discusses the management of ethical AI practices, Section VI demonstrates study results, and Section Seven summarizes conclusions.

II. Related Work

Ethical AI frameworks are increasingly recognized for guiding responsible AI development and deployment. Notable frameworks, such as the OECD AI Principles [9, 10] and the Berkman Klein Center's Model for AI Governance [11], provide essential high-level guidance but are often criticized for lacking specific implementation details. Key shortcomings include inadequate resource allocation and insufficient stakeholder engagement, complicating the translation of principles into practice, especially regarding diverse participation [12, 13, 14, 15, 16, 17, 18].

Effectively addressing bias and discrimination in AI systems necessitates implementing concrete strategies. As highlighted by study [19] in the context of surgical care, a comprehensive approach is essential for achieving equity in sensitive areas such as healthcare [1, 5, 6]. This approach should go beyond general data-centric efforts and incorporate targeted interventions at each stage of the AI development lifecycle.

To effectively mitigate bias and discrimination, various strategies can be employed, as illustrated in Fig. 1, including enhancing data quality through data augmentation and bias detection [19, 20, 21], promoting transparency via model interpretability and audit trails [22, 23, 24], and utilizing XAI to build trust [25, 26, 27, 28, 29, 30, 31, 32]. Additionally, continuous monitoring processes and implementing fairness-aware algorithms and metrics are crucial for developing fairer AI systems and reducing discriminatory outcomes [33, 34, 29, 3, 4, 36].

Implementing these technical solutions requires integration into a comprehensive governance framework that articulates ethical principles and offers actionable steps, including resource considerations and stakeholder involvement. The growing use of AI across various sectors—such as healthcare, finance, education, and the judiciary—highlights the necessity for robust governance frameworks to tackle unique ethical challenges [1, 5, 6, 37, 38, 39, 40, 41, 42].

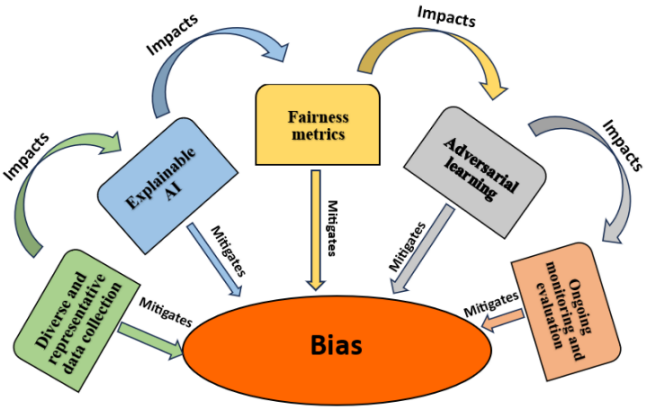


Fig. 1. Strategies for mitigating bias in AI systems.

The HAGF is proposed, which builds on existing efforts by offering a cyclical and interconnected approach to AI governance. HAGF emphasizes resource allocation and combines top-down policy development with bottom-up self-regulation, ensuring continuous refinement. Its focus on resource allocation at each phase, with specific weights assigned to sub-phases, effectively addresses implementation challenges. Critical areas include risk identification, data protection, transparency, and oversight mechanisms.

HAGF's cyclical nature facilitates continuous adaptation to the evolving AI landscape, while KPIs measure effectiveness and progress in ethical AI development. Unlike broader frameworks, HAGF offers specific guidance on data protection, providing organizations with concrete compliance mechanisms. While HAGF has notable strengths, it also recognizes limitations, including the need for deeper guidance on specific ethical issues and implementation challenges; these will be addressed through ongoing refinement and the inclusion of KPIs to track progress. Furthermore, the growing emphasis on auditing AI systems for fairness and bias aligns with HAGF's focus on continuous monitoring, addressing implicit bias, and promoting inclusivity as essential components of the framework.

III. Inclusive Data Collection

Diverse data is essential for developing fair and inclusive AI. Collecting information from various sources and involving multiple stakeholders helps identify potential biases during design and training, enhancing AI accuracy and fairness. In contrast, limited data diversity can lead to overfitting, reduced performance, and the perpetuation of biases against marginalized groups, as illustrated in Fig. 2. The figure underscores these interconnected challenges, including ethical concerns. Therefore, addressing these issues requires careful attention to data collection, model development, and evaluation.

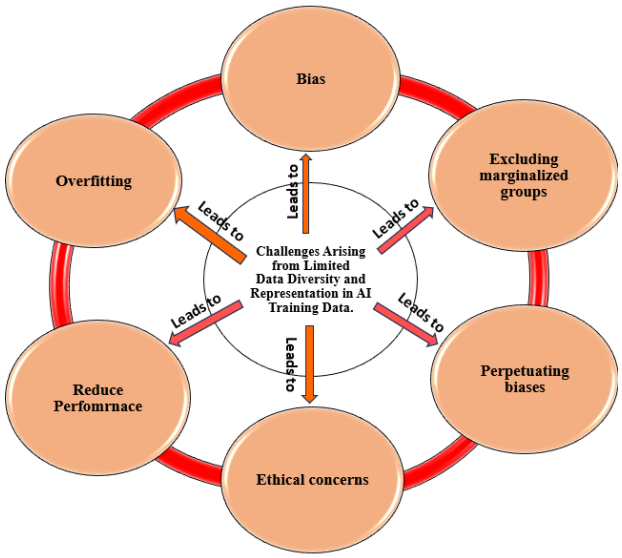


Fig. 2. Challenges arising from limited data diversity and representation.

To address these problems, carefully curating the data to ensure accuracy and representativeness and using techniques such as data augmentation and synthetic data generation to increase diversity and representation. Data curation, the practice of meticulously selecting, organizing, and maintaining data for long-term quality and usability [40], encompasses collection, validation, transformation, integration, and archiving activities. This ensures a trustworthy and well-documented resource for effective analysis and reuse. Building on this foundation, data augmentation transforms existing data to create new, diverse examples, ultimately enhancing model accuracy and robustness [20]. These transformations, tailored to the data type and the problem at hand, can help reduce overfitting and address imbalanced datasets by generating new examples for underrepresented classes, thereby improving overall model performance.

High-quality data is essential for training AI models (as discussed earlier), but challenges arise in acquiring diverse data [40]. Synthetic data generation can address limited or biased training data by creating variations from existing information to improve model fairness and robustness [20]. However, collecting representative data poses difficulties [21, 24, 26, 43, 44]. Biased training data leads to biased AI models, a significant concern for critical applications [45]. Data governance efforts, like audits and fairness benchmarks, aim to ensure the use of representative data, promoting fair and inclusive AI decision-making [46].

IV. Unveiling Bias in AI: XAI, FMs, AL, and M&E

As AI permeates our lives, ensuring fairness and trust is paramount. Three guardians stand watch over responsible development: XAI acts as a transparency lens, revealing how AI makes decisions. FMs, watchful guardians, identify potential biases for equal treatment by AI. Finally, AL strengthens models by mimicking real-world challenges. M&E continuously safeguards against bias creep, ensuring ongoing fairness through regular checks. This ensures AI remains fair and accountable. Despite the oversight role of these guardians, they come with challenges, as outlined in Table I.

Opaque AI systems can lead to biased decisions. XAI cuts through this, revealing AI's reasoning and becoming a powerful tool for fighting bias in critical healthcare fields [27]. Fig. 3 illustrates a typical AI model's limitation: it does not explain its predictions. XAI addresses this by adding an explanation layer and building trust through transparency.

TABLE I. Summary of the Challenges of AI Bias Mitigation Technology

Diverse and representative data collection	***XAI***	***FMs***	***AL***	***On-Going Monitoring***
Data may be incomplete or biased	Computationally expensive	contextual challenges and considerations.	Resource intensive	Resource-intensive.
difficult and computationally expensive	Explanation-challenged	Trade-offs between fairness and performance	limited efficacy	Data bias and incompleteness.
The potential for privacy concerns arises from biases and power imbalances in data collection	Explanations bias	vulnerability to manipulation	over-specialization	Limitations in detecting emerging issues
Potential biases and power imbalances in data collection	Privacy implications	Fairness checklist pitfall	Ethical concerns arise from the possibility of malicious use of AI	The opacity of AI systems limits transparency and interpretability, raising privacy concerns.

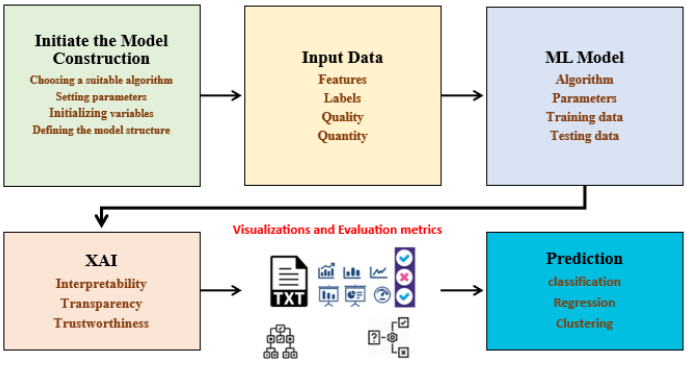


Fig. 3. The placement of XAI in AI model construction.

The National Institute of Standards (NIST) defines transparency, interpretability, explicability, and fairness as key principles for trustworthy AI [28]. These principles are essential for fostering user trust and ensuring AI systems operate ethically and accountable. They align with XAI, which employs techniques like decision trees and other methods to make AI reasoning more transparent for users [29]. By providing clear insights into how decisions are made, XAI helps demystify complex algorithms, enabling users to understand better and challenge AI outputs. This transparency is crucial for individual users, regulatory compliance, and societal acceptance of AI technologies, as illustrated in Fig. 4.

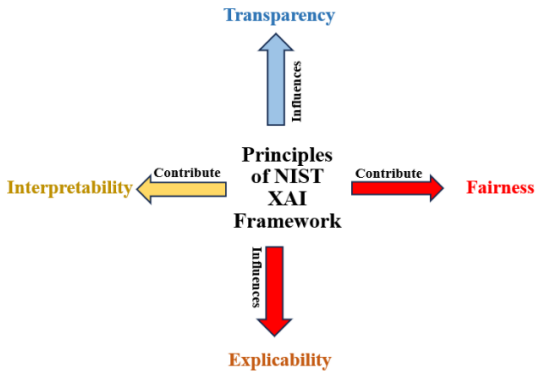


Fig. 4. NIST Key components of AI governance.

However, XAI faces challenges like computational cost, limited explanations, privacy concerns, and unclear explanations that hinder adoption [30]. A multi-disciplinary approach combining machine learning, human-computer interaction, psychology, and ethics is crucial. Balancing technical progress with societal needs and ethics is key to successful XAI [30]. This means developing transparent, interpretable, and ethical XAI solutions prioritizing fairness, accountability, and privacy. A multi-stakeholder approach involving users, experts, and marginalized groups is necessary. Only then can the potential of XAI for trusted and inclusive AI that benefits society be unlocked.

FMs like Equal Opportunity (EO) ensure fair and unbiased AI [31, 32]. These quantitative measures assess bias by examining how AI models treat different groups. For example, EO focuses on consistent True Positive Rates (correct identification of true positives) across groups. Other metrics like Demographic Parity (DP) ensure similar positive outcomes across groups, while Sufficiency (Calibration/Predictive Parity) checks for accurate positive probability predictions [47]. However, FMs face challenges. Defining and measuring them can be difficult; they might conflict with accuracy, and focusing solely on them can lead to a rigid fairness approach. Despite these limitations, FMs are essential for promoting fairness, transparency, and accountability in AI development. It must be recognized that FMs are not a standalone solution [47]. Over-reliance on them risks oversimplification, and manipulation is a potential concern. A balanced approach that considers FMs within a broader fairness definition and measurement framework is necessary. This collaborative effort, acknowledging trade-offs and integrating fairness throughout the AI lifecycle, is key to building more equitable and trustworthy AI systems.

AL tackles bias in AI by introducing cleverly crafted training data designed to fool the system and expose its weaknesses [48, 49]. These examples, like images or text, are false positives during training. By encountering these errors, AI models learn to become more robust against mistakes caused by biased data. AL offers benefits like improved accuracy and reduced cyberattack vulnerability [50]. However, it's not a silver bullet. While promising in improving AI robustness and reducing bias, AL's effectiveness depends on context. It can be computationally expensive and may not address all biases, particularly those rooted in deeper societal issues [51].

Moreover, concerns about overfitting and potential misuse exist. Therefore, AL is most effective as part of a broader strategy for fair AI, alongside continued research focused on responsible innovation and ethics [51]. This combined approach unlocks AL's potential while mitigating risks, allowing AI to harness adversarial training responsibly to create fairer and more trustworthy technologies.

AI systems require continuous M&E to prevent bias from creeping in and ensure ongoing fairness. Regular audits, testing, and user engagement are crucial to identifying and addressing potential biases [36]. FMs, like demographic parity and equal opportunity, can track how outcomes are distributed across different groups and highlight any disparities [36]. Human oversight, through approaches involving humans in the decision-making loop, can help review and intervene in potentially biased decisions [36].

However, challenges exist, including resource-intensive monitoring, requiring expertise and time to analyze data [42]. Biased or incomplete surveillance data can lead to inaccurate conclusions [37]. New biases that emerge over time or issues not present in the training data may also be missed [38]. The lack of transparency in some AI systems can make it challenging to identify bias [52].

Finally, the lack of transparency in some AI systems can make it challenging to identify bias [52], and privacy concerns arise when collecting and analyzing sensitive data [52].

V. Ethical and Responsible AI Management

Governments can promote the safe and responsible implementation of AI by establishing ethical guidelines and standards [54] [55] [56], consulting experts, engaging stakeholders, and formulating clear ethical principles, all of which monitor compliance and implement sanctions to build public trust and ensure the usefulness of AI. Ethics, in general, is a set of principles that help us distinguish between right and wrong. AI ethics is the process of studying how to improve the beneficial impact of AI while minimizing negative consequences. Rapid changes in AI systems raise profound ethical concerns by embedding biases, threatening human rights, and more.

Different international organizations [57] have issued some basic requirements for an AI system to be considered trustworthy. These organizations initially define values necessary for an ethical AI model and move beyond this towards an actionable policy. The most common requirements are explainability, reliability, robustness, security, accountability, privacy and data governance, human agency,

and oversight, legality, fairness, and safety, as shown in Fig. 5; all of these requirements should be evaluated throughout the life cycle of designing an AI model.

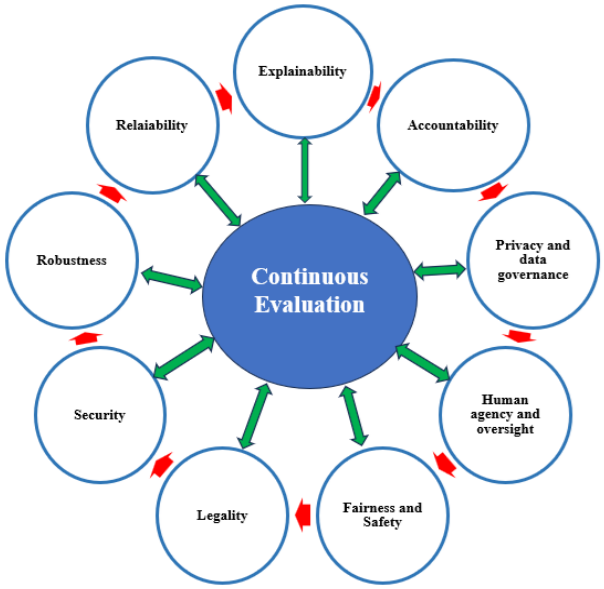


Fig. 5. Basic requirements for an AI model to be considered trustworthy.

Governments play a crucial role in shaping AI's responsible development and use. This includes establishing regulations to address data protection, privacy, and the use of AI in sensitive areas like healthcare and criminal justice. These regulations would focus on data protection, transparency, accountability, and ethical AI use, mitigating potential risks.

These regulations would focus on data protection, transparency, accountability, and ethical AI use, mitigating potential risks. Furthermore, governments can encourage industry self-regulation through voluntary codes of conduct and oversight mechanisms, potentially overseen by regulatory bodies that conduct audits. International collaboration can further solidify these efforts by establishing common standards.

However, responsible AI development shouldn't stifle innovation. Governments can promote AI advancement by investing in research and development, funding academic research, and fostering public-private partnerships. Additionally, supporting educational and training programs equips the workforce with the skills to navigate this evolving technological landscape [41]. This multifaceted approach ensures responsible AI development while fostering innovation and economic growth.

VI. Results

HAGF, presented in Fig. 6, offers a cyclical and interconnected approach to AI governance, emphasizing resource allocation and a multi-faceted strategy combining top-down policy development, bottom-up self-regulation, and continuous refinement. This comprehensive framework establishes a unified approach to AI governance, incorporating a coherent set of definitions and principles for responsible AI utilization. HAGF aligns with both model governance and rights-based models, focusing on ethical AI development, stakeholder engagement, and accountability. It provides a systematic approach to overseeing AI systems throughout their lifecycle.

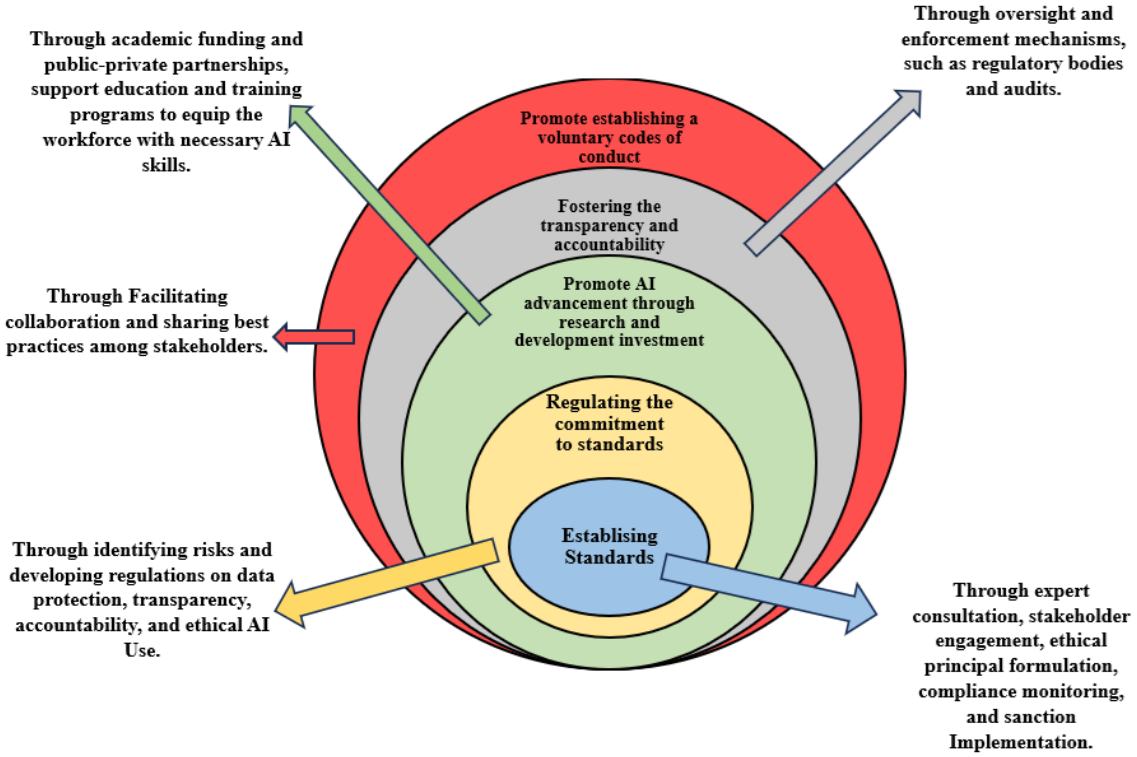


Fig. 6. HAGF – AI Governance framework.

The HAGF framework consists of five interconnected and cyclical components: establishing standards, regulating commitment to those standards, promoting AI advancement, fostering transparency and accountability, and encouraging a voluntary code of conduct. Each of these components plays a specific role in advancing ethical AI governance.

The phases are assigned theoretical weights, as illustrated in Fig. 7, to clearly convey the authors' prioritized vision for responsible AI development and deployment. These weights act as a powerful tool for clarifying the conceptual relationships between the components, providing strategic guidance for implementation, and stimulating discussion and debate regarding prioritization.

These assigned weights enhance the rigor and transparency of the HAGF framework, moving beyond abstract principles to deliver a concrete, actionable, and transparent representation of the authors' vision for ethical AI governance. By emphasizing these weights, the framework illustrates the importance of each component and fosters a deeper understanding of how they interconnect to promote a responsible AI ecosystem.

Phase 1 allocated 20%, focusing on developing a strong foundation for ethical guidelines for AI. This moderately resource-intensive phase involves forming a steering committee, conducting comprehensive research, formulating clear standards, and gathering feedback from diverse stakeholders. Notably, the sub-phase weights prioritize core values, with the highest allocation to formulating ethical guidelines (6%). Furthermore, stakeholder engagement (5%) was emphasized, recognizing the importance of engagement from all stakeholders. In addition, expert consultation (4%) informs the guidelines on best practices, while compliance monitoring (3%) and sanctions enforcement (2%) focus on developing the framework. It is important to note that implementation and enforcement will be addressed in subsequent phases.

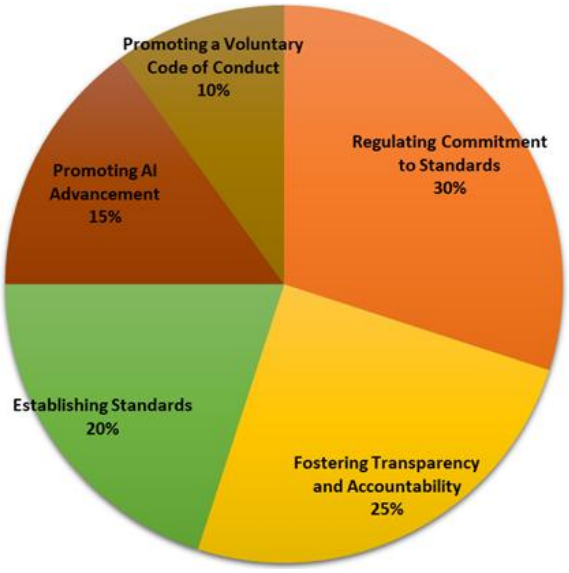


Fig. 7. HAGF phase weights.

The second phase has been assigned the highest weight, with 30% dedicated to translating ethical guidelines into action through robust compliance mechanisms. This resource-intensive phase emphasizes risk identification and the development of regulations for data protection, transparency, accountability, and ethical AI use. Due to the complexity of compliance, ongoing monitoring, regular audits, and potential legal action are necessary. Key focus areas in this phase include risk identification and assessment (9%), data protection regulations (8%), transparency and explainability regulations (7%), accountability mechanisms and sanctions (4%), and collaboration with authorities and dispute resolution (2%). This distribution effectively reflects the relative effort and significance of ensuring compliance with ethical AI standards.

15% has been allocated to the third phase, which aims to foster AI innovation through targeted investments. Notably, Funding Allocation (6%) is the primary focus, receiving the largest share, highlighting the importance of strategically distributing financial resources. In addition, Public-Private Partnerships (4%) leverage both resources and expertise, while Talent Development (3%) works to build a skilled workforce. Ethical Research Practices (1%) ensure that ethical standards are maintained throughout the research process, and Evaluation and Impact Assessment (1%) facilitate continuous learning and improvement. As a result, this distribution emphasizes funding while also recognizing the significance of partnerships, talent, ethics, and evaluation.

In the fourth phase, 25% of the weight is allocated, shifting the focus to building trust through responsible AI practices. Oversight and Enforcement Mechanisms (10%) are essential for ensuring compliance with established guidelines. Furthermore, Regular Audits (8%) assess adherence to these guidelines and pinpoint areas for improvement, while Public Disclosure (7%) enhances transparency by making information about AI systems accessible. This distribution prioritizes oversight and audits, highlighting their importance for accountability, and acknowledges the role of public disclosure in fostering trust.

Finally, 10% of the weighting is allocated to the fifth phase, which emphasizes encouraging self-regulation and best practices in AI. Facilitating Collaboration Among Stakeholders (5%) is vital for uniting diverse perspectives and fostering consensus. Additionally, Sharing Best Practices (3%) allows organizations to learn from each other and implement effective strategies. Promoting Inclusivity (2%) ensures that the codes encompass a broad range of viewpoints. Consequently, this distribution prioritizes collaboration and sharing best practices as essential drivers for developing and adopting meaningful voluntary codes of conduct.

A summary of the weighting assigned to each sub-phase within HAGF is provided in Fig. 8. This figure illustrates a project prioritizing compliance, risk management, and transparency, with Oversight & Enforcement (10%) and Risk Identification (9%) receiving the heaviest weight. Data Protection and Regular Audits (8% each) are also strongly emphasized. Transparency & Explainability (7%) and Ethical Principles and Funding (6% each) are moderately weighted. Stakeholder Engagement (5%), Accountability & Sanctions, Public-Private Partnerships, and Expert Consultation (4% each) are important but less emphasized. Supporting activities

include Talent Development (3%), Public Disclosure (3%), Collaboration (2%), Ethical Research (1%), Evaluation (1%), and Inclusivity (2%). The focus is on robust mechanisms and ongoing monitoring of ethical AI practices.

HAGF aligns with several emerging best practices in AI governance. Its emphasis on ethical development, stakeholder engagement, and accountability resonates with the principle of human-centered AI. The framework prioritizes transparency and accountability, addressing the growing demand for explainable AI (XAI). Furthermore, it implicitly considers risks at various stages of the AI lifecycle and acknowledges the importance of diverse stakeholder involvement, promoting a multi-stakeholder approach. Its cyclical nature facilitates continuous refinement, ensuring the framework remains relevant and effective.

Model Governance Frameworks, including HAGF, the Berkman Klein Center's framework, and the OECD AI Principles, provide blueprints for organizations seeking to implement responsible AI practices. These frameworks offer structured, principled approaches to governance, applicable across various contexts, and serve as models for navigating the complex challenges of ethical AI. While not legally binding or technically specific, they share a commitment to fostering ethical AI practices, each with unique attributes, as highlighted in Table II.

HAGF, in particular, offers significant ethical guidelines for AI development, emphasizing interconnectedness and resource allocation. It promotes stakeholder collaboration and underscores the importance of transparency and accountability, laying a solid foundation for responsible AI practices and positioning HAGF as a valuable contributor to the field.

HAGF framework offers significant advantages through its actionable implementation, breaking down ethical AI governance into specifically weighted sub-phases that prioritize critical areas and inform resource allocation. This explicit weighting compels strategic thinking, clarifying the relative importance of each component. Unlike other frameworks that provide broad guidance, HAGF delivers detailed direction, particularly in data protection, equipping organizations with concrete mechanisms for compliance. Its cyclical and adaptable nature fosters continuous improvement, ensuring relevance in a rapidly evolving landscape.

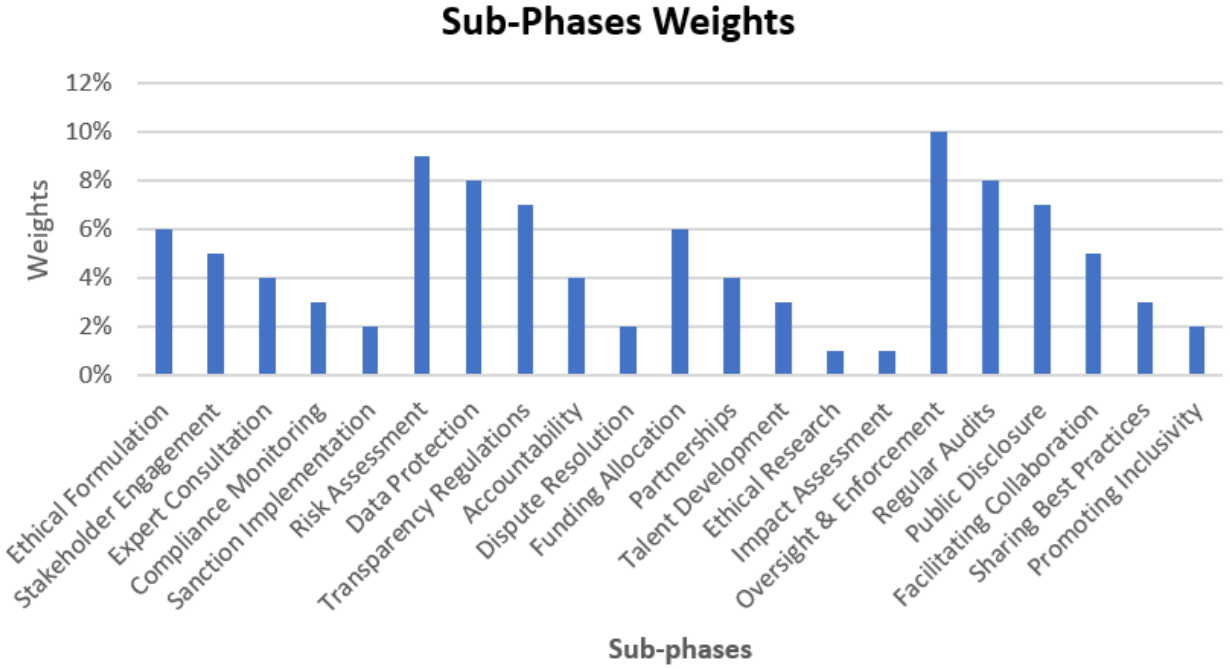


Fig. 8. Percentage weighting of sub-phases in HAGF.

TABLE II. COMPARING AI GOVERNANCE FRAMEWORKS: HAGF, OECD, AND BERMAN KLEIN

Framework	*Scope*	*Focus*	*Holistic Approach*	*Resource Allocation*	*Practical Focus*
HAGF Framework	Provides ethical guidelines for AI development and use, may offer more specific guidance	Interconnectedness and resource allocation.	Yes, it considers various interconnected components	Emphasizes the importance of funding and support	Provides a more concrete and actionable framework
OECD AI Principles	Provides general ethical guidelines for AI development and use	A general approach to ethical considerations	No	It does not explicitly address resource allocation	It gives more general principles
Berkman Klein Center's Model	Provides a comprehensive framework for AI governance	A holistic approach, considering various aspects of AI governance	Yes	It does not explicitly address resource allocation	Provides a general framework with guidance on implementation

However, HAGF has areas for potential improvement. It would strengthen its applicability by providing more detailed guidance on specific ethical issues, such as enhancing its focus on data governance with clear directives for responsible data practices. While its current complexity may pose challenges for some organizations, these issues are manageable and can be addressed through further refinement. HAGF's reliance on consistent funding and expertise highlights its potential for growth and adaptation. Measuring the effectiveness of its ethical guidelines and fostering public trust will enable it to evolve with rapid technological change.

Improvements in its enforcement mechanisms, particularly concerning voluntary codes of conduct, are also essential. As illustrated in Table III, a comprehensive set of KPIs has been developed to facilitate ongoing improvement by measuring the effectiveness of each HAGF phase and tracking progress toward ethical AI development. These KPIs cover critical aspects such as standards development and adoption, regulatory compliance, AI research investment, public trust, stakeholder engagement, and workforce impact. This clear mechanism for monitoring HAGF's performance helps identify areas for enhancement and ensures continuous progress in ethical AI practices, fostering accountability and transparency.

This holistic, cyclical approach, coupled with its emphasis on resource allocation and integration of top-down and bottom-up governance strategies, positions HAGF as an effective tool for promoting responsible AI development and use. Addressing the identified weaknesses and aligning more closely with best practices will further enhance HAGF's effectiveness in fostering ethical AI practices. Recognizing these limitations, as reflected in the insights from Table III, offers a balanced perspective on HAGF's potential and challenges.

TABLE III. KPIS FOR EVALUATING AN AI GOVERNANCE FRAMEWORK

Phase	***Key Performance Indicator (KPI)***
Establishing standards	- Number of AI-related standards and guidelines developed and published. - Level of stakeholder engagement in the standard-setting process. - Clarity and Comprehensibility of the established standards. - Adoption rate of the established standards by AI developers and users. - Number of revisions made to the standards. - Time taken to develop and publish the standards
Regulating the commitment to standards	- Compliance rate with established AI standards. - Number and severity of non-compliance incidents. - Effectiveness of enforcement mechanisms (e.g., audits, sanctions). - Level of public trust in the AI governance framework. - Number of AI-related regulations enacted. - Time taken to resolve non-compliance incidents. - Level of industry engagement in the regulatory process.
Promoting AI advancement through research and development investment	- Amount of funding allocated for AI research and development. - Number of successful AI research projects funded. - Number of new AI-based products or services developed. - Growth of the AI industry in revenue, employment, and innovation. - Collaboration rate between academia and industry on AI projects. - Number of AI-related patents filed. - Return on investment (ROI) for AI research funding.
Fostering transparency and accountability	- Number of AI systems with publicly available information on their algorithms and data sources. - Level of public trust and confidence in AI systems. - Number of successful cases of AI-related accountability measures (e.g., investigations, legal actions). - Frequency and quality of AI impact assessments. - Number of AI explainability tools or frameworks developed. - Number of AI bias detection and mitigation strategies implemented. - Level of public awareness and understanding of AI systems.
Promoting establishing a voluntary code of conduct.	- Number of organizations that have adopted voluntary codes of conduct for AI. - Level of adherence to the codes of conduct by organizations. - Positive impact of codes of conduct on ethical and responsible AI development. - Public perception of the effectiveness of voluntary codes of conduct. - Number of industry-led initiatives promoting responsible AI. - Level of diversity and inclusion in the development and governance of AI systems. - Number of self-assessments or audits conducted by organizations to ensure compliance with codes of conduct.

VII. CONCLUSION

Bias and discrimination in AI systems present significant challenges; however, emerging technologies such as Explainable AI (XAI), Fairness Metrics (FMs), and Adversarial Learning (AL), combined with robust governance frameworks, offer promising solutions. This article introduced the Holistic AI Governance Framework (HAGF), a comprehensive and cyclical approach to ethical AI governance. HAGF comprises five interconnected components: establishing standards, regulating adherence to those standards, promoting AI advancement, fostering transparency and accountability, and encouraging voluntary codes of conduct. By prioritizing resource allocation and integrating top-down policies with bottom-up self-regulation, HAGF addresses the complexities of

AI governance. Its emphasis on expert consultation, stakeholder engagement, and ethical principle formulation lays a strong foundation for effective standards. Additionally, focusing on risk identification, data protection, and accountability through regulation and oversight cultivates a culture of compliance. Investment in AI innovation and the promotion of voluntary codes of conduct supports both responsible practices and technological advancement. To ensure HAGF's effectiveness and adaptability in an evolving AI landscape, a robust suite of Key Performance Indicators (KPIs) is proposed, measuring standards, regulatory effectiveness, public trust, societal impact, and research advancement. Ongoing research, stakeholder engagement, and continuous evaluation and refinement of HAGF—particularly regarding data rights and algorithmic accountability—will be crucial for realizing a just and equitable AI future.

REFERENCES

[1] I. Castiglioni, L. Rundo, M. Codari, G. Di Leo, C. Salvatore, M. Interlenghi, et al., "AI applications to medical images: From machine learning to deep learning," Phys. Med., vol. 83, pp. 9–24, Mar. 2021. DOI: 10.1016/j.ejmp.2021.02.006

[2] C. Rudin, "Stop explaining black-box machine learning models for high stakes decisions and use interpretable models instead," Nat. Mach. Intell., vol. 1, no. 5, pp. 206–215, May 2019. DOI: 10.1038/s42256-019-0048-x

[3] X. Ferrer, T. Nuenen, J. M. Such, M. Cote, and N. Criado, "Bias and discrimination in AI: A cross-disciplinary perspective," IEEE Technol. Soc. Mag., vol. 40, no. 2, pp. 72–80, 2021. DOI: 10.1109/MTS.2021.3056293

[4] Z. Chen, "Ethics and discrimination in artificial intelligence-enabled recruitment practices," Humanit. Soc. Sci. Commun., vol. 10, no. 1, pp. 1–12, 2023. DOI: 10.1057/s41599-023-02079-x

[5] S. Reddy, S. Allan, S. Coghlan, and P. Cooper, "A governance model for the application of AI in health care," J. Am. Med. Inform. Assoc., vol. 27, no. 3, pp. 491–497, Mar. 1 2020. DOI: 10.1093/jamia/ocz192.

[6] Dankwa-Mullan, I., Ndoh, K., Akogo, D., Rocha, H. A. L., & Juaçaba, S. F. (2025). Artificial Intelligence and Cancer Health Equity: Bridging the Divide or Widening the Gap. *Current Oncology Reports*, 1-17.

[7] Prabhakar, P., Pati, P. B., & Parida, S. (2025). Navigating Legal and Ethical Dimensions in AI, IoT, and Cloud Solutions for Women's Safety. In *Developing AI, IoT and Cloud Computing-based Tools and Applications for Women's Safety* (pp. 155-191). Chapman and Hall/CRC.

[8] D. F. Engstrom and A. Haim, "Regulating Government AI and the Challenge of Sociotechnical Design," Annu. Rev. Law Soc. Sci., vol. 19, no. 1, p. 19, 2023. DOI: 10.1146/annurev-lawsocsci-120522-091626

[9] Da Mota, M. (2024). Toward an AI Policy Framework for Research Institutions. *Artificial Intelligence*.

[10] Fu, Y., & Weng, Z. (2024). Navigating the ethical terrain of AI in education: A systematic review on framing responsible human-centered AI practices. *Computers and Education: Artificial Intelligence*, 100306.

[11] Papagiannidis, E. (2024). Responsible AI governance in practice: The strategic impact of responsible AI governance on business value and competitiveness.

[12] Khan, M. A. (2024). *Responsible AI Principles in Product Management Practices* (Doctoral dissertation, PhD thesis, Aalborg University).

[13] Challoumis, C. (2024, October). Building a sustainable economy-how ai can optimize resource allocation. In *XVI International Scientific Conference* (pp. 190-224).

[14] Constantinides, M., Bogucka, E., Quercia, D., Kallio, S., & Tahaei, M. (2024). RAI guidelines: Method for generating responsible AI guidelines grounded in regulations and usable by (non-) technical roles. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW2), 1-28.

[15] Leslie, D., & Perini, A. M. (2024). Future Shock: Generative AI and the international AI policy and governance crisis.

[16] Kaminski, M. E., & Malgieri, G. (2024). Impacted Stakeholder Participation in AI and Data Governance. *Forthcoming (2024-25)*.

[17] Lee, S. U., Perera, H., Liu, Y., Xia, B., Lu, Q., Zhu, L., ... & Whittle, J. (2024). Responsible AI Question Bank: A Comprehensive Tool for AI Risk Assessment. *arXiv preprint arXiv:2408.11820*.

[18] Porter, Z., Habli, I., McDermid, J., & Kaas, M. (2024). A principles-based ethics assurance argument pattern for AI and autonomous systems. *AI and Ethics*, *4*(2), 593-616.

[19] A. Ahmad, et al., "Equity and Artificial Intelligence in Surgical Care: A Comprehensive Review of Current Challenges and Promising Solutions," BULLET: Jurnal Multidisiplin Ilmu, vol. 2, no. 2, pp. 443–455, 2023.

[20] Paproki, A., Salvado, O., & Fookes, C. (2024). Synthetic Data for Deep Learning in Computer Vision & Medical Imaging: A Means to Reduce Data Bias. *ACM Computing Surveys*.

[21] J. S. Park, et al., "Understanding the representation and representativeness of age in AI data sets." Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 2021. DOI: 10.1145/3461702.3462590

[22] Zhang, Kuan, et al. "Mitigating bias in radiology machine learning: 2. Model development." Radiology: Artificial Intelligence 4.5 (2022): e220010. DOI: 10.1148/ryai.220010

[23] Gunning, David, et al. "XAI—Explainable artificial intelligence." Science robotics 4.37 (2019): eaay7120. DOI: 10.1126/scirobotics.aay7120.

[24] Landers, Richard N., and Tara S. Behrend. "Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models." American Psychologist 78.1 (2023): 36. DOI: 10.1037/amp0000972

[25] Karim, Md Rezaul, et al. "Explainable AI for Bioinformatics: Methods, Tools and Applications." Briefings in Bioinformatics 24.5 (2023): bbad236. DOI: 10.1093/bib/bbad236

[26] S. Akhai, "From Black Boxes to Transparent Machines: The Quest for Explainable AI." Available at SSRN 4390887 (2023). DOI: 10.2139/ssrn.4390887

[27] Eke, C. I., & Shuib, L. (2024). The role of explainability and transparency in fostering trust in AI healthcare systems: a systematic literature review, open issues and potential solutions. *Neural Computing and Applications*, 1-36.

[28] Radanliev, P. (2025). AI Ethics: Integrating Transparency, Fairness, and Privacy in AI Development. *Applied Artificial Intelligence*, *39*(1), 2463722.

[29] Kostopoulos, G., Davrazos, G., & Kotsiantis, S. (2024). Explainable Artificial Intelligence-Based Decision Support Systems: A Recent Review. *Electronics*, *13*(14), 2842.

[30] Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., ... & Stumpf, S. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, *106*, 102301.

[31] Schoenherr, Jordan Richard, et al. "Designing AI using a human-centered approach: Explainability and accuracy toward trustworthiness." *IEEE Transactions on Technology and Society* 4.1 (2023): 9-23.

[32] P. Phillips, C. Hahn, P. Fontana, A. Yates, K. Greene, D. Broniatowski, et al., Four Principles of Explainable Artificial Intelligence, NIST Interagency/Internal Report (NISTIR). Gaithersburg, MD: National Institute of Standards and Technology, 2021 [online], https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=933399. Accessed Jul. 17, 2023., DOI: 10.6028/NIST.IR.8312.

[33] Bellamy, Rachel KE, et al. "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias." arXiv preprint arXiv: (2018).

[34] C. Dwork, et al., "Fairness through awareness." Proceedings of the 3rd innovations in theoretical computer science conference. 2012. DOI: 10.1145/2090236.2090255

[35] Almajali, M. H., Nasrawin, L., Alqudah, F. T., Althunibat, A. A., & Albalawee, N. (2023). Technical Service Error as a Pillar of

Administrative Responsibility for Artificial Intelligence (AI) Operations. *International Journal of Advances in Soft Computing & Its Applications, 15*(3).

[36] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification." Conference on fairness, accountability and transparency. PMLR, 2018.

[37] Cascella, Marco, et al. "Crossing the AI Chasm in Neurocritical Care." Computers 12.4 (2023): 83. DOI: 10.3390/computers12040083

[38] J. S. Brownstein, B. Rader, C. M. Astley, and H. Tian, "Advances in Artificial Intelligence for infectious-disease surveillance," N. Engl. J. Med., vol. 388, no. 17, pp. 1597–1607, Apr. 27 2023. DOI: 10.1056/NEJMra2119215

[39] Qatawneh, A. M., Lutfi, A., & Al Barrak, T. (2024). Effect of Artificial Intelligence (AI) on Financial Decision-Making: Mediating Role of Financial Technologies (Fin-Tech). HighTech and Innovation Journal, 5(3), 759-773.

[40] Abu Afifa, M. M., Nguyen, T. H., Le, M. T. T., Nguyen, L., & Tran, T. T. H. (2024). Accounting going digital: a Vietnamese experimental study on artificial intelligence in accounting. *VINE Journal of Information and Knowledge Management Systems*.

[41] D. Schiff, "Education for AI, not AI for Education: The Role of Education and Ethics in National AI Policy Strategies," Int. J. Artif. Intell. Educ., vol. 32, no. 3, pp. 527–563, 2022., doi:DOI: 10.1007/s40593-021-00270-2

[42] G. Said, et al., "Adapting Legal Systems to the Development of Artificial Intelligence: Solving the Global Problem of AI in Judicial Processes," International Journal of Cyber Law, vol. 1, p. 4, 2023.

[43] H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, et al., "Scientific discovery in the age of artificial intelligence," Nature, vol. 620, no. 7972, pp. 47–60, Aug. 2023. DOI: 10.1038/s41586-023-06221-2

[44] Y. Chen, E. W. Clayton, L. L. Novak, S. Anders, and B. Malin, "Human-centered design to address biases in artificial intelligence," J. Med. Internet Res., vol. 25, p. e43251, Mar. 24 2023. DOI: 10.2196/43251

[45] D. Hartmann et al., "Addressing the Regulatory Gap: Moving Towards an EU AI Audit Ecosystem Beyond the AIA by Including Civil Society," arXiv preprint arXiv:2403.07904, 2024.

[46] M. Roshanaei, "Towards best practices for mitigating artificial intelligence implicit bias in shaping diversity, inclusion and equity in higher education," Educ. Inf. Technol., pp. 1–26, 2024.

[47] N. Zhou, Z. Zhang, V. N. Nair, H. Singhal, and J. Chen, "Bias, fairness and accountability with artificial intelligence and machine learning algorithms," Int. Stat. Rev., vol. 90, no. 3, pp. 468–480, 2022. DOI: 10.1111/insr.12492

[48] Pagano, Tiago P., et al. "Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods." Big data and cognitive computing 7.1 (2023): 15. DOI: 10.3390/bdcc7010015.

[49] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples." arXiv preprint arXiv: (2014).

[50] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning." Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018. DOI: 10.1145/3278721.3278779

[51] D. Madras, et al., "Learning adversarially fair and transferable representations." International Conference on Machine Learning. PMLR, 2018.

[52] Nwakanma, Cosmas Ifeanyi, et al. "Explainable artificial intelligence (xai) for intrusion detection and mitigation in intelligent connected vehicles: A review." Applied Sciences 13.3 (2023): 1252. DOI: 10.3390/app13031252

[53] Dhirani, Lubna Luxmi, et al. "Ethical dilemmas and privacy issues in emerging technologies: a review." Sensors 23.3 (2023): 1151. DOI: 10.3390/s23031151

[54] J. Cahill, V. Howard, Y. Huang, J. Ye, S. Ralph, and A. Dillon, "Intelligent Work: Person Centered Operations, Worker Wellness and the Triple Bottom Line," Commun. Comput. Inf. Sci., vol. 1421, pp. 307–314, 2021., doi:DOI: 10.1007/978-3-030-78645-8_38

[55] M. Cole, C. Cant, F. Ustek Spilda, and M. Graham, "Politics by Automatic Means? A Critique of Artificial Intelligence Ethics at Work," Front. Artif. Intell., vol. 5, p. 869114, Jul. 15 2022., doi:DOI: 10.3389/frai.2022.869114

[56] "Ethics of Artificial Intelligence." UNESCO, 26 Sept. 2023, www.unesco.org/en/artificial-intelligence/recommendation-ethics.

[57] "Ethics Guidelines for Trustworthy AI." Shaping Europe's Digital Future, 8 Apr. 2019, digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Enhanced Early Detection of Diabetic Nephropathy Using a Hybrid Autoencoder-LSTM Model for Clinical Prediction

U. Sudha Rani[1], Dr. C. Subhas[2]
Research Scholar, JNTUA Ananthapuramu, India[1]
Professor, JNTUA CEK[2]

***Abstract*—Early detection and precise prediction are essential in medical diagnosis, particularly for diseases such as diabetic nephropathy (DN), which tends to go undiagnosed at its early stages. Conventional diagnostic techniques may not be sensitive and timely, and hence, early intervention might be difficult. This research delves into the application of a hybrid Autoencoder-LSTM model to improve DN detection. The Autoencoder (AE) unit compresses clinical data with preservation of important features and dimensionality reduction. The Long Short-Term Memory (LSTM) network subsequently processes temporal patterns and sequential dependency, enhancing feature learning for timely diagnosis. Clinical and demographic information from diabetic patients are included in the dataset, evaluating variables such as age, sex, type of diabetes, duration of disease, smoking, and alcohol use. The model is done using Python and exhibits better performance compared to conventional methods. The Hybrid AE-LSTM model proposed here attains an accuracy of 99.2%, which is a 6.68% improvement over Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression. The findings demonstrate the power of deep learning in detecting DN early and accurately and present a novel tool for proactive disease control among diabetic patients.**

Keywords*—Autoencoder-LSTM; Diabetic nephropathy; early disease detection; machine learning; clinical data analysis; hybrid models

I. Introduction

Diabetes is a serious issue for general health. Approximately millions of individuals worldwide suffer with diabetes. Globally, there were 463 million diabetics in 2019, and 700 million are predicted by 2045 [1]. Kidneys, eyes, nerves, skin, and heart can all be harmed by it. The most frequent cause of kidney failure in diabetic people is diabetes nephropathy (DN)[2]. DN patients have surged in tandem with the exponential rise in the frequency of diabetes patients. Consequently, the mortality rate of DN has also gone up. Consequently, it's critical to identify DN patients early in order to prevent the related illnesses. Early detection of DN by diagnostic markers is crucial, as it might impede the loss of renal function and mitigate unfavorable consequences. Micro albuminuria, or the presence of minute quantities of protein albumin in the urine, is recognized as the first indication of the onset of diabetes mellitus. On the other hand, a significant amount of renal damage has been documented to occur even prior to the development of micro albuminuria [3]. A number of complicating factors, including exercise, urinary tract infections, acute illnesses, and heart failure, are linked to albuminuria. Moreover, it has been documented to transpire in the urine of individuals following a regular diet, suggesting that albuminuria is not a reliable indicator for precisely forecasting diabetic kidney disease.

Elevated blood vessel glucose levels are critical to the development of diabetic neuropathic pain. Through an excess of reactive oxygen species (ROS), hyperglycemia causes problems with metabolism in mitochondria and the sugar metabolic pathway [4]. Glycation at high glucose concentrations creates adducts that are covalent with plasma proteins. AGEs, or advanced glycation end products, are one of these events and a significant risk factor for complications from diabetes. Podocytes are an essential component of the glomerular filtration barrier, and they may become aberrant after extended exposure to hyperglycemia. The loss of podocytes is one of the earliest glomerular morphologic alterations, and it is essential to the emergence of DN. Clinically speaking, diabetic individuals with DN have proteinuria and decreased kidney function [5]. DN patients can be maintained with blood pressure and glucose management, but many eventually develop renal failure [6]. Therefore, it will be crucial to comprehend the pathophysiology of DN and create novel biomarkers in order to diagnose DN early. Nowadays diagnostic methods for conditions such as DN present various issues that stem from the reliance on clinical sign and/or biomarkers associated with the late stage of the disease. Such reliance can lead to lack of timely treatment/fulfillment of early milestones/health concerns prevention. Also, there are other diagnostic tests whereby biopsy or blood samples are taken and this causes discomfort or is risky and therefore people are discouraged from frequent checkup. These traditional tools may also be insensitive and non-specific and therefore result in false positive or negative results as the traditional biomarkers may not necessarily mimic the early stages of DN. Secondly, certain diagnostic techniques for instance, specialized imaging may be expensive as well as unavailable in most health facilities and more so in developing countries. There is also a certain degree of subjectivity and variability in diagnostics, this is due to the facts that using subjective methods such as clinical impression or an interpretation of test results the results of the diagnostics can be significantly different depending a healthcare provider's experience.

However, current approaches of advanced diagnostic targets do not seem to have a perfect solution to these

challenges; but, the machine learning (ML) models for medical diagnostic are a potential solution to these challenges [7]. The large volume of complex data gathered in every healthcare practice may be analyzed by an ML model to search for patterns that can be missed by a clinician and that can predict DN and its development even before its first clinical manifestation. They also enable constant, non- invasive monitoring through trailing with wearable devices or EHR, which offer an instant read out of the patient's status. Through utilization of big and heterogeneous databases, it is possible to enhance the diagnostic precision and increase the abilities to distinguish between the diseases that are similar and make more accurate prognosis depends on the individual parameters of the patient [8]. Due to scalability and cost implications, ML models provide coherent diagnosis support irrespective of the healthcare facility, hence the need to employ specialized personnel and equipment especially in setting with limited accessibility to the same [9]. Moreover, a model for prediction based on the DN-related parameters must be created. As an automated model construction method, ML-based approaches have taken the lead in the domains of medical imaging, human interaction, and healthcare. ML-based approaches are mainly employed for early identification and prediction/detection of different healthcare conditions, such as diabetes, carcinoma, and kidney damage, in order to increase classification accuracy. The popularity of ML-based approaches has skyrocketed recently. While a significant amount of research has been conducted and newly created ML-based algorithms have garnered attention, the hunt for ways to improve classifier accuracy has never ended [10]. Thus, one of the key elements that will determine how accurate the classifiers are is the selection of an ML-based model. Over many years, researchers have worked very hard to create useful models to enhance precision of categorization. Medical data categorization remains a difficult problem for machine learning-based classifiers, despite the quick advancement of computational intelligence theories.

Furthermore, with the use of ML models, patients can be diagnosed and treated, based on the specific characteristics like genetics, lifestyle, and presence of a variety of diseases. It can also take disparate data from clinical, genomic and environmental domains and make application of the data to provide new and unique insights into the disease causes, drivers and therapeutic outcomes. Taken together, the existing diagnostic practices have certain drawbacks, whereas the machine learning models open a vast range of possibilities that help improve the detection rates at the initial stages, refine the diagnostics, and transform the healthcare services to be more individualized and easily scalable. These ML-based classifiers, however, are still unable to categorize patients accurately due to their unsatisfactory accuracy. But the goal of this work is to use ML-based approaches to create a prediction model based on the risk variables associated with DN. Feature extraction techniques also identify DN risk variables. Here are four primary contributions of the proposed Hybrid Autoencoder-LSTM model for detecting Diabetic Nephropathy (DN).

- The study introduces the use of an Autoencoder for effective dimensionality reduction, which helps in isolating essential features. This step not only simplifies the dataset but also enhances the model's ability to focus on the most relevant information, thereby improving the overall accuracy and interpretability of the model.
- The combination of Autoencoder and LSTM architectures leverages the strengths of both models. The Autoencoder efficiently handles feature learning and noise reduction, while the LSTM network excels at sequential data analysis. This hybrid approach provides a comprehensive framework for DN detection, offering improved prediction accuracy compared to conventional machine learning models.
- The study provides a comprehensive performance evaluation of the AE-LSTM model, including metrics such as reconstruction loss, classification accuracy, precision, recall, and F1-score. The comparative analysis with other methods highlights the AE-LSTM model's superior performance and its potential advantages in handling complex, high-dimensional healthcare data.
- The use of advanced optimization techniques like the Adam Optimizer, along with appropriate loss functions (MSE for the Autoencoder and binary/categorical cross-entropy for the LSTM), ensures efficient and effective training of the model. This contributes to achieving high performance metrics, such as accuracy and precision, in the detection and diagnosis of DN.

The rest of the contents are listed in the following order. An introduction is given in Section I. The literary portions are shown in Section II. This is the problem statement found in Section III. The hybrid model-based modeling and analysis approach is covered in Section IV. The results are compiled and the performance indicators are shown in Section V. Section VI offers further research and a conclusion.

II. Related Works

Kim et al. [11] developed the initial stages diagnostic biomarkers to detect DN as a means of DN intervention. In the investigation, Zucker diabetes-related fatty rats were used to model the DN phenotype. The results showed that in addition to significantly raised serum levels of blood glucose, BUN, and creatinine, DN rats also exhibited severe renal injury, fibrosis, and microstructural changes. Moreover, the urine of DN rats emitted higher concentrations of kidney injury molecule-1 (KIM-1) and neutrophil gelatinase-associated lipocalin (NGAL). New DN biomarkers were discovered by transcriptome analysis. Moreover, they were discovered in DN patients' urine. The findings showed that the onset of diabetic nephropathy was associated with an up-regulation of CXCR6 expression levels in rat urine, renal tissue, and clinical samples. In essence, our discovery offers direct evidence that CXCR6 was elevated in urine as diabetic nephropathy progressed. The results therefore imply that the CXCL16/CXCR6 pathway may be involved in the development of end-stage renal disorders. Using these results, a unique therapeutic approach to treating renal fibrosis can be developed. It is unclear, therefore, how CXCR6 contributes to the development of diabetic nephropathy. To investigate the underlying process in DN, more research is needed.

A new deep learning model is presented by Singh et al. [12] for the early identification and prediction of chronic kidney disease. This project aims to construct a DNN and evaluate its performance relative to other state-of-the-art machine learning techniques. Any information that were absent from the database during testing were substituted with the mean of the relevant attributes. The optimal parameters of the neural network were then established by configuring them and carrying out several trials. The most important characteristics were selected using Recursive Feature Elimination (RFE). Hemoglobin, specific gravity, serum creatinine level, blood vessel count, albumin, packed cell volume, and high blood pressure were important features in the RFE. To categorize them, machine learning models were given a range of attributes. The technique could be useful to nephrologists in detecting CKD. The model's testing on limited data sets was one of its limitations. In the future, large amounts of more complex and representative CKD data will be gathered to determine the severity of the illness and enhance the model's performance.

The goal of the Kang et al. [13] study was to create and assess a DL model that uses retinal fundus images to identify early renal function deterioration. This retrospective analysis includes patients who had color fundus imaging and renal function testing. From the images, a DL model was built to identify renal impairment. An estimated rate of glomerular filtration of less than 90 mL/min/1.73 m2 was considered early renal function impairment. The AUC and ROC curve were used to assess the performance of the model. For the whole population, the model's AUC was 0.81.Using retinal fundus images, the deep learning algorithm in this work makes it possible to identify early renal function deterioration. Only for individuals with increased blood HbA1c levels, the model demonstrated greater accuracy is the drawback.

A unique ML model for CKD prediction was put out by Arif et al. [14] It included a number of pre-processing stages, selection of features, a hyper parameter optimization method, and ML algorithms. In order to tackle the difficulties encountered with medical datasets, we utilize sequential data scaling along with robust scaling, z-standardization, and min-max scaling, as well as iterative imputation for missing values. The Boruta method is used for feature selection, while ML algorithms are used to create the model. The model performed exceptionally well with good accuracy when assessed on the UCI CKD dataset. The method, which combines novel pre-processing techniques, the Boruta feature selection, the k-nearest-neighbour algorithm, and grid-search cross-validation (CV) for hyper parameter tuning, shows promise in improving the early identification of CKD. This study emphasizes how machine learning approaches might enhance therapeutic support networks and lessen the influence of ambiguity around the prognosis of chronic illnesses. The study's primary drawback was its dependence on a single dataset the UCI CKD dataset, which has a significant number of missing values.

By combining a number of easily accessible clinical variables with retinal vascular measures, Shi et al. [15] developed a new DKD diagnostic approach for individuals with type 2 diabetes. Xiangyang Central Hospital's 515 consecutive type-2 diabetes mellitus patients were included. Patient diagnoses of DKD were used to separate patients into two groups: the training and testing set, with a random seed of 1. While the ML was developed using data from the training set, the MLA was validated using data from the testing set. The model's performances were assessed. When compared to other classifiers, the random forest classifier-using MLA performed at its best. Verified, the accuracy was 84.5%. Retinal vascular alterations may help in DKD screening and identification, according to a novel machine learning method for the disease's diagnosis that was constructed using fundus images and eight readily accessible clinical data. The sample used in the study limits the generalizability of the model to broader and more diverse populations.

In the Zhang et al. [16] study, membranous nephropathy was diagnosed by combining deep learning techniques with blood and urine Raman spectra. Following baseline correction and data smoothing, the training set was supplemented with Gaussian white noise at varying decibel levels to enhance the data. The assessment results of the ResNet, AlexNet, and GoogleNet models for membranous nephropathy were then obtained by feeding the amplified data into them. As per the experimental findings, AlexNet emerged as the most proficient deep learning model for both samples. All three models were able to attain an accuracy of 1 in classifying serum data pertaining to patients with membranous kidney damage and the unaffected group, and above 0.85 in differentiating urine data. The test results described above show how powerful deep learning methods can be when used in combination with serum- and urine-based Raman analysis to accurately and quickly diagnose individuals with membranous nephropathy. The limitation is the high accuracy reported would not be achievable in a more diverse and unstructured clinical environment, where data quality and characteristics can vary widely.

Several limitations to the application of ML and DL models in diagnosing kidney related diseases are presented in the reviewed literature. A typical issue is that the number and types of datasets are often limited and often only a single dataset may be used in developing the model and hence the range and the variety of data populations might not be very wide. Also, some research works' drawbacks were associated with data quality and data loss, where data missing was a major issue that came up to a need of imputation or data augmentation. In this case, the models' applicability for different conditions or groups may be limited since the improvement was noted only in patients with raised HbA1c levels. In addition, the very high levels of accuracy found in this and similar studies, again in precise experimental settings, may not be highly representative of the variability and range of clinical datasets, again influencing the model's results in the clinic. Lastly, the strong focus on concrete characteristics that include the use of retinal fundus images or Raman spectra may also suggest the models' drawback in conditions when such data is irrelevant or missing.

III. Problem Statement

DN is categorized as a usual complication of diabetes the result of which may culminate into end-stage renal disease if not diagnosed. The current diagnostic processes that are based on biomarkers and imaging oftentimes diagnose DN at the most severe stages, thereby limiting the interventional and treatment

procedures[16]. The reason why diagnosis is usually made later in the course of DN is partly attributed to the slow and progressive nature of kidney damage in such patients when conventional techniques are used. Thus, addressing the mentioned problem of the shortage of diagnostic techniques, this paper proposes and compares an Autoencoder-LSTM model for the diagnosis of diabetic nephropathy. The proposed hybrid model intends to integrate the autoencoder' s ability of decreasing the dimensionality of data and finding hidden beneficial aspects as well as the feature superior to recognizing temporal relations in the series data which is LSTM network. Using this model for Clinical and patient data analysis, the study aims at finding the feeble signs of DN that are masked normally. The latter goal is to create a less invasive and more precise diagnostic tool for early identification of DN, which in turn will allow for proper treatment to be given, enhance the patient's outcomes, and perhaps decrease the likelihood of transitioning to the more significant level of nephropathy.

IV. Proposed Methodology of Hybrid Autoencoder-LSTM model for Early Detection of Diabetic Nephropathy

For the detection of DN, this study uses the Hybrid Autoencoder-LSTM model that uses feature learning and temporal pattern analysis. The methodology encompasses several key stages: acquisition of data as well as preparation, designing the model as well as learning rate schemes for the model. First, the input dataset is formed by using clinical data of patients with diabetes, retrieved from prior researches. This dataset must then be cleaned to deal with any issues pertaining to missing values as well as the normalization of its features, and where needed, feature selection. The Autoencoder component accomplishes the Dimensionality reduction of the input data for the aim of segregating necessary features from the noisy ones. These condensed features are then passed to the LSTM network which learns on the sequences to identify relationships with time of DN. The connection of these two parts is to improve the accuracy of the model through the application of the advantages of the architectures of deep learning. When training the model, several loss functions are used; MSE for autoencoder and binary or categorical cross-entropy for LSTM based on the classification task. The Adam Optimizer helps in achieving the steady state for the value and the model goes through epochs for the optimization. The DN diagnosis and accurate identification of patients with the condition is sought through this integrated, extensive methodological approach that aims to enhance the optimality of the model. It is demonstrated in Fig. 1 given below.

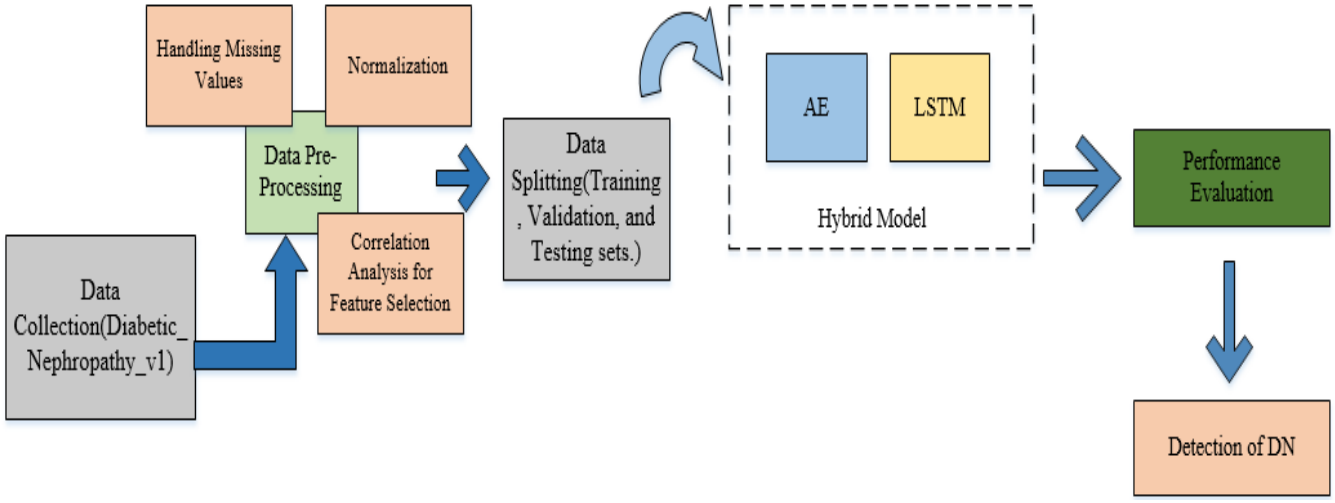


Fig. 1. Hybrid AE-LSTM model block diagram.

A. Dataset Collection

The dataset which is "Diabetic_Nephropathy_v1" contains clinical and demographic data of DN and related diseases[17]. The dataset consists of 767 patient records. These variables involve the patient's sex, age, type of diabetes, duration of the disease, DR and DN, and smoking and drinking habits; as well as glucose levels, HbA1c, body mass index (BMI), and blood pressure. These parameters include height, weight, body mass index, systolic blood pressure, diastolic blood pressure, glycated haemoglobin, fasting blood glucose, blood triglycerides, C-peptide, total cholesterol, high-density lipoprotein cholesterol and low-density lipoprotein cholesterol respectively. More so, details on medication use including insulin, metformin, as well as lipid-lowering drugs among others are incorporated. These variables are expected to be used for analyzing the associations between them and the development of diabetic nephropathy, the results of which are planned to be used for developing better predictors of the disease and increasing the general knowledge on the subject.

B. Data Pre-processing

1) Handling missing values: It is essential to manage the cases of missing values because, for example, they affect the accuracy of a machine learning model, as well as the results of making predictions. It is recommended to impute the missing values with the median or mode of the particular columns for numerical variables since median imputation is less influenced by outliers than the mean imputation. For categorical features it is optimal to use simple form of imputation for the missing data, which is to replace it with the most frequently used category. This approach ensures that the employed dataset is also strong and minimizes the chances of a mistake being made on the

model. Feature selection was performed using correlation analysis and domain knowledge to identify the most relevant clinical variables, discarding those that were redundant or had low significance for DN.

2) Normalization: Most preprocessing acts like standardization on the input variables are significant for models like Autoencoder and LSTMs as they are dramatically affected by the scale of input variables needed. This process helps to make all features at a similar scale to make better and more accurate models and avoid instabilities in the model. Min-Max Scaling standardizes the features to a range of 0 to 1, which is a good practice in terms of equal scaling of features and can contribute to the enhancement of the model's performance and convergence. The formula for Min-Max Scaling is

$$Y_{Scaled} = \frac{Y - Y_{min}}{Y_{max} - Y_{min}} \quad (1)$$

Where Y is the original feature value, Y_{min} is the minimum value of the feature, and Y_{max} is the maximum value. This normalization ensures that all features contribute equally to the model, which is crucial for algorithms sensitive to feature scaling.

3) Correlation analysis for feature selection: Begin by examining the correlation between each feature and the target variable, which in this case is diabetic nephropathy. This analysis helps in identifying features that have a strong relationship with the target variable. Features that exhibit high correlation with the target are likely to provide significant predictive value. For example, if one use correlation coefficients or statistical tests to quantify these relationships. Features with low or negligible correlations can be discarded to simplify the model and improve its interpretability.

C. Integration of Hybrid AE-LSTM Model for the Detection of Diabetic Nephropathy

A multilayer neural network called an auto encoder produces desirable outputs that are similar to inputs with less modification—that is, results that are similar to inputs that have some reconstruction error [18]. Generally, autoencoder is applied in dimensionality reduction since it provide efficient dimensionality reduction in cases of large and more specifically medical data sets due to its high dimensionality reduction rate yet efficiency in preserving important features of the data. In contrast to model like PCA it is capable of learning local non-linear manifold structure of data and thus is more appropriate for high dimensional clinical records. Further, it aids in the reduction of noise to ensure the model concentration on the most important characteristic which is important for tasks such as DN development prediction. It is the same case with autoencoder as they also function in an unsupervised way, which can be useful when processing big data with little or no labels at all. The auto encoder encrypts the input and then utilizes unsupervised learning to reconstruct or decode the output.

The encoder, reconstruction loss, bottleneck, and decoder are the four main parts of a generic auto encoder. The encoder helps to remove characteristics from the input by shrinking the data into an encoded form. The bottleneck layer is the layer that has the fewest characteristics and compressed incoming data. The decoder makes sure that the input and output are the same by helping the model to rebuild the result from the encoded representation. The last metric used to evaluate the decoder's performance and determine how closely the output resembles the original input is Reconstruction Loss.

Additionally, training is done using back propagation, which further minimizes reconstruction loss. This minimum loss serves as an example of the goal that AE aspires to accomplish. The input y that the encoder will compress Eq. (2).

$$y = E(x) \quad (2)$$

Decoder will make an effort to replicate the input. D as x' = D (E (x)).

$$loss(E, D) = \frac{1}{n}\sum_{j=1}^{n} x^i - D(E(x^i)))^2 \quad (3)$$

In this instance, the reconstruction loss equals the difference between the encoded and decoded vectors. The MSE is one method for calculating the reconstruction loss. It is stated in the above-mentioned Eq. (3). The hybrid AE-LSTM's architectural diagram is shown in Fig. 2.

LSTM was created using sophisticated recurrent neurons. In an LSTM, every recurrent neuron may be thought of as a single cell state [19]. For the temporal analysis, the study use LSTM since it is capable of processing sequential data and it can capture long term dependencies, this is because tracking the health status of the patients require tracking their status over a period of time. The LSTMs are built in a way that they do not suffer from vanishing gradient problem and this makes the model to retain information from the previous time step, which is extremely important when predicting medical conditions. For hyper-parameters, learning rate, batch size and number of hidden units were appropriately selected from cross-validation in order to achieve high learning rate but low over-fitting. These choices were further optimized to make sure that the model does well as far as the training data is concerned as well as the unseen data. An LSTM determines its current state by using its data from the previous state, much like a conventional RNN does. The LSTM uses three gates to control the current neuron: the forget gate, update gate, and output gate.

A LSTM can connect current data with historical knowledge. An LSTM is coupled to three gates: an output gate, an input gate, and a forget-about gate. The new and last states are represented by the symbolsQ_t andQ_{t-1}, respectively for the input, and p_t and p_{t-1} for the existing and prior outputs.

Eq. (4), Eq. (5) and Eq. (6) explain the LSTM input gate idea.

$$j_t = \sigma(\mathrm{Z}_j \cdot [z_{t-1}, y_t] + b_j) \quad (4)$$

$$\tilde{Q}_t = \tanh(Z_j \cdot [p_{t-1}, \mathrm{p}_t] + \mathrm{b}_j) \quad (5)$$

$$\mathrm{Q}_t = f_t \mathrm{Q}_{t-1} + j_t \tilde{Q}_t \quad (6)$$

To decide which of the data points y_t and p_{t-1} should be added, where Eq. (4) use a sigmoid layer to filter them. Combining the long-term storage data,$\tilde{Q}_t$ with the present moment informationQ_{t-1}, results in Eq. (6). $\tilde{Q}_t$ Displays a tanz

output, whereas $Z_j$indicates a sigmoid results. The bias of the LSTM input gate is denoted by $b_j$in this instance, while Z_j denotes the weight matrices. Consequently, because of the LSTM's forget gate, the dot product and sigmoid layer may pass information selectively. A certain probability is used to decide whether to delete relevant data from a previous cell.

Use Eq. (7) to determine whether to preserve relevant data from an earlier cell with a particular option. Z_f Represents the weighted matrix b_f the offset, and σ the sigmoid term.

$$f_t = \sigma\left(Z_f \cdot [c_{t-1}, y_t] + b_f\right) \quad (7)$$

The states needed for the following equations are determined by the output gate of the LSTM Eq. (8) and Eq. (9) states provided by the inputs y_t and p_{t-1} .After the final output is produced, it is multiplied by the state decision vectors Q_t that transmit new data via the tanz layer.

$$R_t = \sigma(Z_o \cdot [P_{t-1}, y_t] + b_o) \quad (8)$$

$$p_t = R_t \tanz(Q_t) \quad (9)$$

When using the Autoencoder-LSTM, the transition from the Autoencoder to LSTM is carried out systematically with regards to features and sequence. First, the bottleneck layer of the autoencoder, which contains the compressed and the higher level features of the input is used as input to the LSTM network. This transfer of feature favours LSTM to process data that has undergoes post processing hence removing noise and unnecessary details. Since these are the features passed to the LSTM network, the LSTM makes its computations in a sequential manner which is vital in the learning of temporal characteristics and structures that are useful in diagnosing Diabetic Nephropathy (DN). The last state of the LSTM network is used for the purpose of predicting the probability of DN or for distinguishing between the patients who have DN and those who do not have DN based on temporal pattern learning integrated into the model from the sequences.

The integration of the hybrid model has the following advantages. Thus, by reducing the dimensionality of input data, the first stage of the autoencoder employs a K value to facilitate the LSTM's identification of relevant patterns to the task. This organizational improvement make the model lighter and thus enhances its functionality. Second, the temporal features as processed by the LSTM show important sequential characteristics that can indicate early signs of DN that are not observed by simpler models, allowing for a better understanding of the diseases' evolution.

Optimizations of the autoencoder and LSTM network is done during the training phase so as to get the best performance. In the autoencoder, mean squared error (MSE) loss function is used to minimize the errors in data reconstruction; this way, only the noise is eliminated, and important details are preserved. With relation to the LSTM network, the selection of the loss function is contingent upon the nature of the classification; in cases where the classification is categorically classified as several classes, as opposed to binary cross entropy, which is utilized in cases where the classification is either true or false, the binary cross entropy loss function is employed. For optimizing weights in the model and to make enhancements, specialized algorithms like Adam is used. The training process takes several epochs; the used number of epochs and the batch size is defined depending on the data volume and available computational power. Syllable stress and domains' size tuning makes sure that the model properly learns and functions well in regard to new data.

This detailed approach will assist in established structure to enhance the Autoencoder-LSTM model, by having dimensionality reduction then LSTM in achieving accurate Diabetic Nephropathy prediction. Fig. 2 illustrates the architecture of proposed hybrid AE-LSTM is given below.

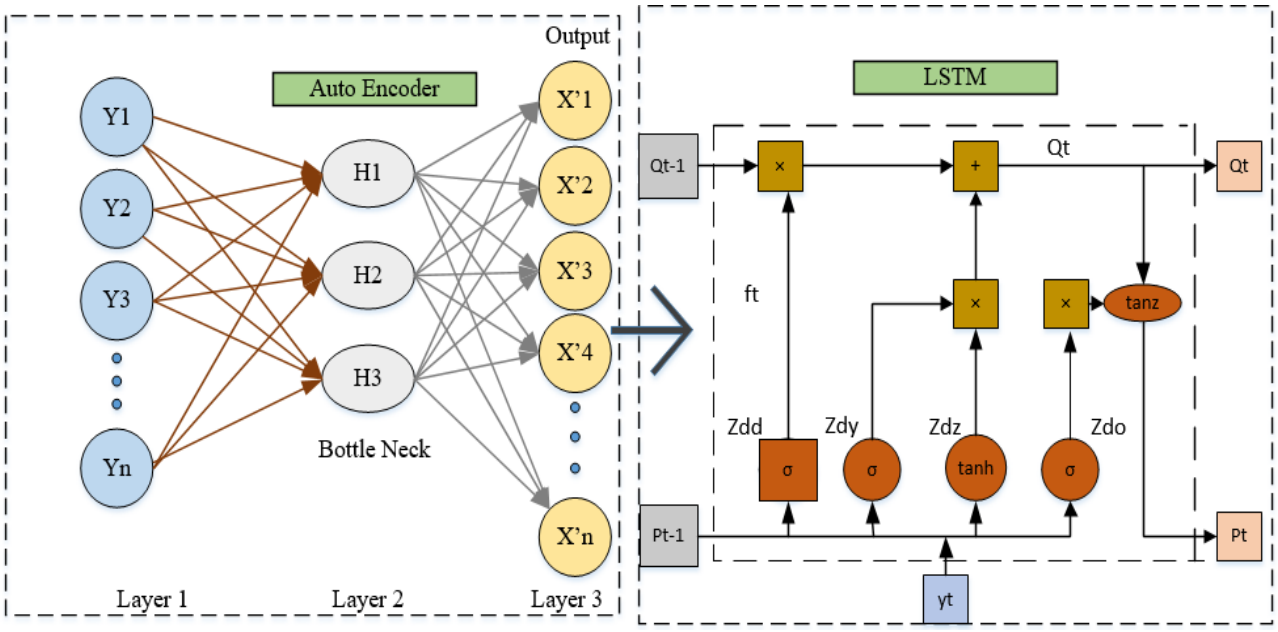


Fig. 2. Hybrid AE-LSTM architecture.

V. Results and Discussion

The results section of the study, the authors provided distilled information about the results from the predictive modeling and disease detection study using the proposed Hybrid Autoencoder-LSTM model. The accuracy of the presented model was compared to several simple and complex models based on the sub belongs and more traditional machine learning as well as deep learning techniques. The performance evaluation was based on the model's metrics as the model successfully predicted the occurrence of DN. The methodology employed data from clinical records and qualified the model implementation and analysis on a Windows 10 environment using python programming language. In line with such findings, the model excels in the identification of the DN onset; this is due to the utilization of the autoencoder dimensionality reduction and the LSTM network with the capability of recognizing temporal patterns. The following variables were used in a measure of the efficiency and predictive capability of the specified model.

A. Auto Encoder's Reconstruction Loss

Reconstruction Loss is one of the ways of evaluating a model particularly an autoencoder to whether compress the data and then decompress it to get the preservation quality. It measures the degree of distortion of output signal in comparison with the input signal end product. It is then computed by comparing the two, sometimes via calculating MSE or Binary Cross-Entropy, etc. Hence, reconstruction loss defines how much information has been lost during the encoding and decoding process, and lower value of it would mean better reconstructions and therefore a better performance of the model in terms of preserving important features of the input data. Eq. (10) expressed it.

$$\text{Reconstruction Loss} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x_i})^2 \qquad (10)$$

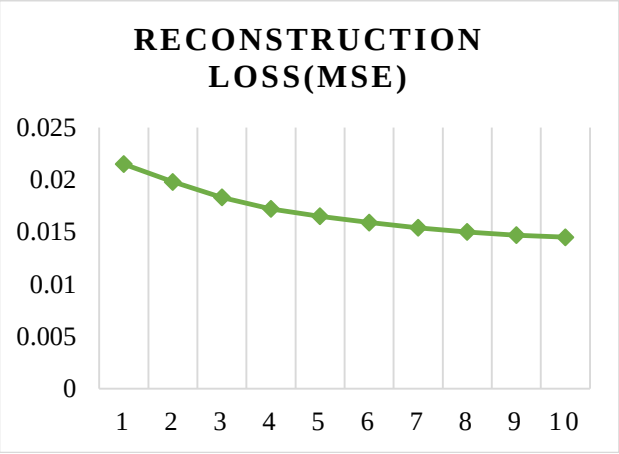


Fig. 3. Reconstruction loss of the proposed AE-LSTM approach.

The Fig. 3 will indicate the reconstruction loss in the Autoencoder, which should decrease over time, thus demonstrating the Auto encoder's ability to learn how to encode and decode the data. The gradual reduction in MSE shows that the present Autoencoder component continues to reduce the dimensionality of the data and retain crucial characteristics efficiently.

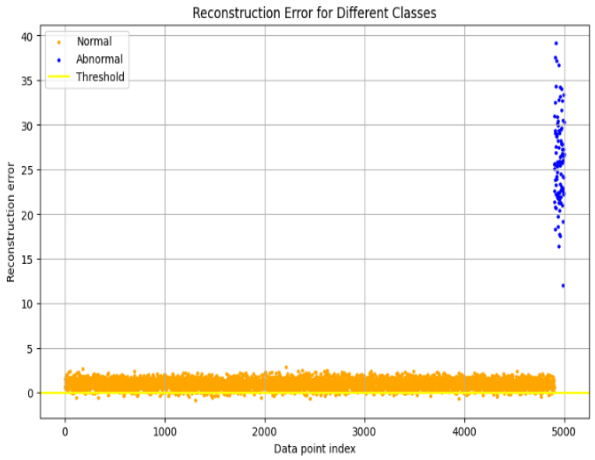


Fig. 4. Reconstruction error for different classes.

Fig. 4 shows the reconstruction errors for normal and abnormal classes, with normal data points clustered around a value of 1 and abnormal points centered around 25. The threshold line, set at 0, visually separates the error ranges for both classes. Fig. 5 shows the histogram of reconstruction error is given below.

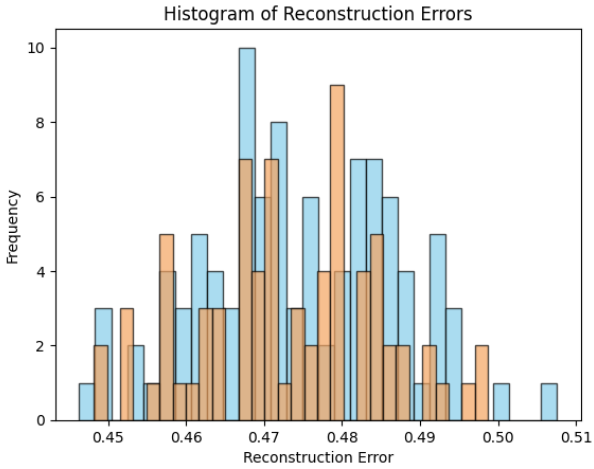


Fig. 5. Histogram of reconstruction error.

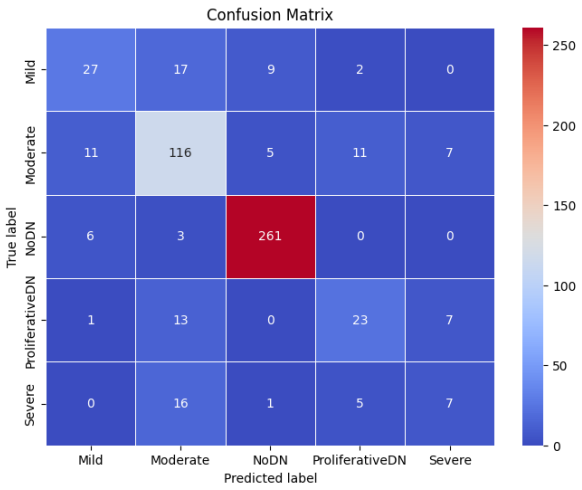


Fig. 6. Confusion matrix.

Fig. 6 shows a heat map of the confusion matrix, displaying the performance of a classification model across five categories: It includes Mild, Moderate, No DN, Proliferative DN and Severe. The diagonal values are the correctly predicted patterns giving high accuracy for some of the top categories such as "No DN" (261 instances correctly classified). Other discrepancies such as the 17 Mild examples classified as Moderate as well as the 16 Severe specimens also classified as Moderate can also be observed from off- diagonal values. Different colors define importance of the data and stress on the distribution of errors and correct predictions.

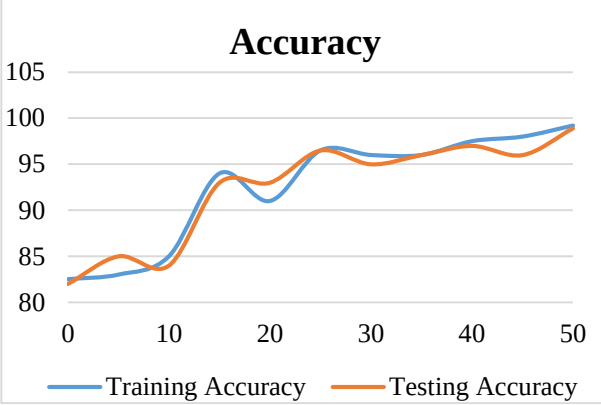


Fig. 7. Training and testing accuracy of the proposed AE-LSTM approach.

Fig. 7 shows that the performance metrics indicating the increase in model's ability to identify relevant patterns for DN detection have escalated exponentially. Thus, it is not surprising that the accuracy recorded during this study was a phenomenal 99.2%% shows that the model is very well trained and is in condition to be able to make sound predictions with the help of encoded features and temporal data. The training accuracy graph shows them gradually rising up to a certain point proving that the model gains better and better understanding of the training data set during the training process. The testing accuracy curve is also smooth and similar to the training accuracy which might depict a good generalization of the model with unseen data. Hence the little difference between the training and testing accuracy shows that the model performed very well by reducing the risk of over fitting as well as improving the robustness of the model.

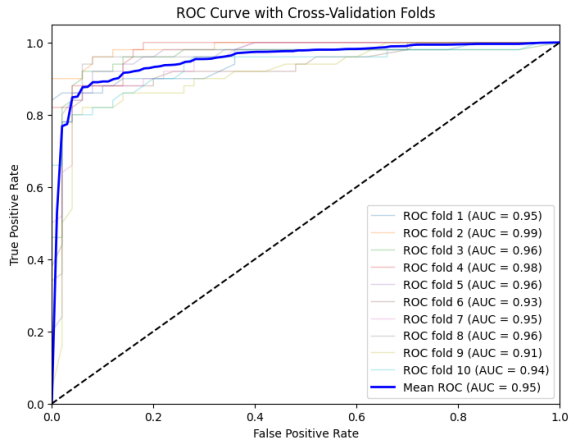


Fig. 8. ROC of the proposed approach.

Fig. 8 shows the ROC curve for the proposed AE-LSTM approach, illustrating how well the model distinguishes between positive and negative cases. The high AUC suggests that the AE-LSTM approach is effective in correctly classifying the data, making it a reliable method for early detection and diagnosis. The results demonstrate that using autoencoder for feature extraction combined with LSTM for classification is a successful strategy in this study.

B. Performance Metrics

1) Accuracy: A method's accuracy is measured by the proportion of test cases it can identify correctly on a certain test set. It is computed as follows in Eq. (11)

$$Accuracy = \frac{RN+RP}{RP+AP+RN+AN} \quad (11)$$

2) Precision: Precision is the ratio of all positively recognized cases to the total number of properly identified positive occurrences by the model. It is quantified in Eqn. (12) is as follows.

$$Precision = \frac{True\ Positives}{(True\ Positives+False\ Positives)} \quad (12)$$

3) Recall: Recall is the idea of the positive cases that the framework correctly detects. It is calculated as follows in Eq. (13).

$$Recall(sensitivity) = \frac{True\ Positives}{True\ Positives+False\ Negatives} \quad (13)$$

4) F1-Score: When there is a large difference in the number of students in one class compared to the other, the F1 score might be helpful. To access the F1 score is apply the Eq. (14) as follows.

$$F1\ Score = 2 \times \frac{(Precision*Recall)}{(Precision+Recall)} \quad (14)$$

When assessing someone, one should consider their F1 score as it provides a useful and unbiased means of gauging recall and accuracy.

C. Consideration with Other ML Approaches

Table I concern the relative comparison of the overall effectiveness of the considered classification algorithms. It can also be seen that, the proposed AE+LSTM method has the maximum accuracy of 99%. 2%, and the EWs are 98. 75%, 98. 92%, with an excellent SW of 98. 79%. These results indicate that the AE+LSTM is not only able to pinpoint the true positive instances but it also keeps a good harmony between Precision and Recall. On the other hand, the Accuracy index of the SVM is fairly impressive recording a 98. Achieves a 96% level of accuracy, solid values for the precision and recall rates, however has a slightly lower F1-score. The same can be said about the Multivariate Logistic Regression method, which also yields the 0.95 of accuracy, but does not rank as high as the AE+LSTM. RF method also behaves well, but has the lowest performance indicators with accuracy equal to 85%, and less TS, PR, and F1-MACV. Thus, it is established that the AE+LSTM approach performs better as compared to the other methods studied for the classification of the data, especially with reference to precision and recall, which are quite

significant signs of how effectively the model is useful for a wide range of data patterns. It is illustrated in Fig. 9. The proposed Hybrid Autoencoder-LSTM model significantly improves early DN detection by enhancing feature extraction and temporal pattern learning. Its ability to reduce dimensionality while maintaining critical diagnostic information ensures higher accuracy than traditional methods. This advancement enables timely medical intervention, improving patient outcomes. The study highlights AI's potential in transforming predictive healthcare with scalable and reliable diagnostic models.

TABLE I. EXISTING METHODS AND SUGGESTED METHOD COMPARISON

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1Score (%)
RF[20]	85	81.60	82.24	81.42
SVM[21]	98.96	91.78	94.08	90.21
Multivariate Logistic Regression[22]	95	92.78	90.82	90.21
Proposed AE+LSTM	99.2	98.75	98.92	98.79

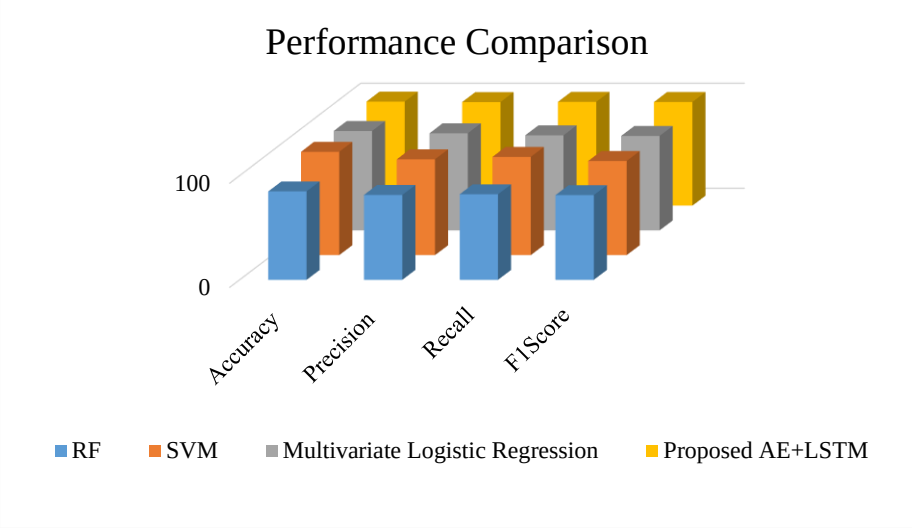


Fig. 9. The performance evaluations of AE+LSTM with conventional approaches.

D. Discussion

The findings showed that one of the better ways of early prediction of DN has been established to be the Hybrid Autoencoder-LSTM model given its efficiency in handling the otherwise elaborate medical diagnoses. Autoencoder whose function here is to decrease the dimensionality of the image data leaving vital features intact and removing background noise; this improves LSTM's ability to learn and identify temporal features essential for DN diagnosis. The LSTM's sequences and the capability to store information from the previous states also enables it to follow the advancement of DN in time. Compared to other classical machine learning systems like RF[20], support vector machines[21], and logistic regression[22], this hybrid model has better result. This benefit derived from Autoencoder dimensionality reduction becomes very useful in helping LSTM isolate on the most significant features, thus lowering the levels of computationally intensity and enhancing model analyzability. The reduction in reconstruction loss over time reflects the autoencoder' s capability to efficiently compress and reconstruct the data while preserving important features. This capability is crucial for enhancing the LSTM's performance by focusing on relevant temporal patterns without being overwhelmed by irrelevant data. The successful reduction in reconstruction errors, particularly the distinction between normal and abnormal classes, further validates the effectiveness of the autoencoder in filtering out noise and emphasizing critical features. Such findings imply that deep learning architectures are more valuable for the intricate medical diagnosis. In a clinical sense, the predictability of DN would definitely alter decisions that are made on a patient as this would enable the clinicians to prevent or at least delay the occurrence of the disease since there would be time to plan on how to handle the situation. Since the proposed model has achieved good values for precision and recall, it means that the identified patients might indeed be at high risk of developing DN, and therefore, timely intervention might help improve the patient's condition. The Hybrid Autoencoder-LSTM performs efficiently in facilitating early Diabetic Nephropathy (DN) prediction through dimensionality reduction and temporal pattern discovery. Autoencoder removes noise and retains core features, improving LSTM's potential for DN progress tracking. Its accuracy is superior compared to RF, SVM, and Logistic Regression with reduced complexity. The decrement in reconstruction loss reflects its optimization in feature extraction, which makes it perform classifying tasks with better efficiency. High recall and precision values validate its feasibility for actual clinical use, allowing for prompt interventions to enhance patient outcomes.

VI. CONCLUSION AND FUTURE SCOPE

In conclusion the findings and analysis of the Autoencoder-LSTM model for the identification of DN in the early stages have shown the prospect of strengthening diagnostic performance. It is suggested that integration of an autoencoder with an LSTM network reduces the dimensionality and contains profitable sequential pattern understanding for the later

component in improving the identification of early signs of DN. The study's results underscore the model's potential for early detection of DN, which could lead to improved patient outcomes through timely medical intervention. Despite these achievements, future work should focus on several key areas to further enhance the model's applicability and performance. This includes validating the AE-LSTM model on diverse datasets to ensure generalizability across different populations, exploring additional feature extraction techniques to improve model robustness, and investigating the integration of the model into clinical decision-support systems for real-time applications. It is suggested that in future studies, the sample ought to be diverse, and the data collected should be followed up over time to see how to improve on the given model. Further, it may be beneficial to experiment with state of the art methods like incorporating attention mechanism, or integrating hybrid model with other modalities of data like genomics or Imaging data to enhance the diagnostic accuracy for early detection. Practical emphasis and the integration of technological innovations into clinical practices will also play an important role in converting these progressive changes into valuable outcomes for patient and stakeholders' experiences.

REFERENCES

[1] S. M. Ganie, M. B. Malik, and T. Arif, "Performance analysis and prediction of type 2 diabetes mellitus based on lifestyle data using machine learning approaches," Journal of Diabetes and Metabolic Disorders, vol. 21, no. 1, p. 339, Jun. 2022, doi: 10.1007/s40200-022-00981-w.

[2] I. Păunică et al., "The Bidirectional Relationship between Periodontal Disease and Diabetes Mellitus—A Review," Diagnostics, vol. 13, no. 4, Art. No. 4, Jan. 2023, doi: 10.3390/diagnostics13040681.

[3] Ganie et al., "An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators," Healthcare Analytics, vol. 2, p. 100092, Nov. 2022, doi: 10.1016/j.health.2022.100092.

[4] E. Eftekharpour and P. Fernyhough, "Oxidative Stress and Mitochondrial Dysfunction Associated with Peripheral Neuropathy in Type 1 Diabetes," https://home.liebertpub.com/ars, Sep. 2022, doi: 10.1089/ars.2021.0152.

[5] P. González, P. Lozano, G. Ros, and F. Solano, "Hyperglycemia and Oxidative Stress: An Integral, Updated and Critical Overview of Their Metabolic Interconnections," International Journal of Molecular Sciences, vol. 24, no. 11, Art. No. 11, Jan. 2023, doi: 10.3390/ijms24119352.

[6] Pathak et al., "Mechanistic approach towards diabetic neuropathy screening techniques and future challenges: A review," Biomedicine & Pharmacotherapy, vol. 150, p. 113025, Jun. 2022, doi: 10.1016/j.biopha.2022.113025.

[7] H. S. R. Rajula, G. Verlato, M. Manchia, N. Antonucci, and V. Fanos, "Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment," Medicina, vol. 56, no. 9, Art. No. 9, Sep. 2020, doi: 10.3390/medicina56090455.

[8] S. Dixit, A. Kumar, and K. Srinivasan, "A Current Review of Machine Learning and Deep Learning Models in Oral Cancer Diagnosis: Recent Technologies, Open Challenges, and Future Research Directions," Diagnostics, vol. 13, no. 7, Art. No. 7, Jan. 2023, doi: 10.3390/diagnostics13071353.

[9] S. Mishra and A. Tyagi, "The Role of Machine Learning Techniques in Internet of Things-Based Cloud Applications," 2022, pp. 105–135. doi: 10.1007/978-3-030-87059-1_4.

[10] J.-P. O. Li et al., "Digital technology, tele-medicine and artificial intelligence in ophthalmology: A global perspective," Progress in Retinal and Eye Research, vol. 82, p. 100900, May 2021, doi: 10.1016/j.preteyeres.2020.100900.

[11] K.-S. Kim et al., "Identification of Novel Biomarker for Early Detection of Diabetic Nephropathy," Biomedicines, vol. 9, no. 5, Art. No. 5, May 2021, doi: 10.3390/biomedicines9050457.

[12] V. Singh, V. K. Asari, and R. Rajasekaran, "A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease," Diagnostics, vol. 12, no. 1, Art. No. 1, Jan. 2022, doi: 10.3390/diagnostics12010116.

[13] E. Y.-C. Kang et al., "Deep Learning–Based Detection of Early Renal Function Impairment Using Retinal Fundus Images: Model Development and Validation," JMIR Medical Informatics, vol. 8, no. 11, p. e23472, Nov. 2020, doi: 10.2196/23472.

[14] M. S. Arif, A. Mukheimer, and D. Asif, "Enhancing the Early Detection of Chronic Kidney Disease: A Robust Machine Learning Model," Big Data and Cognitive Computing, vol. 7, no. 3, Art. No. 3, Sep. 2023, doi: 10.3390/bdcc7030144.

[15] S. Shi et al., "The automatic detection of diabetic kidney disease from retinal vascular parameters combined with clinical variables using artificial intelligence in type-2 diabetes patients," BMC Med Inform Decis Mak, vol. 23, no. 1, Art. No. 1, Dec. 2023, doi: 10.1186/s12911-023-02343-9.

[16] X. Zhang et al., "Rapid diagnosis of membranous nephropathy based on serum and urine Raman spectroscopy combined with deep learning methods," Sci Rep, vol. 13, no. 1, p. 3418, Feb. 2023, doi: 10.1038/s41598-022-22204-1.

[17] "Diabetic_Nephropathy_v1." Accessed: Sep. 18, 2024. [Online]. Available: https://www.kaggle.com/datasets/wangting2023/diabetic-nephropathy-v1

[18] Wei Song, "A new deep auto-encoder using multiscale reconstruction errors and weight update correlation," Information Sciences, vol. 559, pp. 130–152, Jun. 2021, doi: 10.1016/j.ins.2021.01.064.

[19] R. C. Staudemeyer and E. R. Morris, "Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks," Sep. 12, 2019, arXiv: arXiv:1909.09586. Accessed: Nov. 21, 2023. [Online]. Available: http://arxiv.org/abs/1909.09586

[20] S. M. H. Sarkhosh, A. Esteghamati, M. Hemmatabadi, and M. Daraei, "Predicting diabetic nephropathy in type 2 diabetic patients using machine learning algorithms," Journal of Diabetes and Metabolic Disorders, vol. 21, no. 2, p. 1433, Dec. 2022, doi: 10.1007/s40200-022-01076-2.

[21] L. Muflikhah, F. A. Bachtiar, D. E. Ratnawati, and R. Darmawan, "Improving Performance for Diabetic Nephropathy Detection Using Adaptive Synthetic Sampling Data in Ensemble Method of Machine Learning Algorithms," J. Ilm. Tek. Elektro Komput. Dan Inform, vol. 10, no. 1, p. 123, Feb. 2024, doi: 10.26555/jiteki.v10i1.28107.

[22] S. M. Hosseini Sarkhosh, M. Hemmatabadi, and A. Esteghamati, "Development and validation of a risk score for diabetic kidney disease prediction in type 2 diabetes patients: a machine learning approach," J Endocrinol Invest, vol. 46, no. 2, pp. 415–423, Feb. 2023, doi: 10.1007/s40618-022-01919-y.

A Review of Cybersecurity Challenges and Solutions for Autonomous Vehicles

Lasseni Coulibaly[1], Damien Hanyurwimfura[2], Evariste Twahirwa[3], Abubakar Diwani[4]
African Center of Excellence in Internet of Things, University of Rwanda, Kigali, Rwanda[1, 2, 3]
Department of Computer Science and IT, The State University of Zanzibar, Zanzibar, Tanzania[4]

***Abstract*—With the continuously increasing demand for new technologies, many concepts have emerged in recent decades and the Internet of Things is one of the most popular. IoT is revolutionizing several aspects of human life with a large range of applications including the transportation sector. Based on IoT technologies and Artificial Intelligence, new-generation vehicles are being developed with autonomous or self-driving capabilities to handle transportation in future smart cities. Regarding human-based errors such as accidents, traffic congestion, and disruptions, autonomous vehicles are presented as an alternative solution to increase traffic safety, efficiency, and mobility. However, by transferring from a human-based to a computer-based driving style, the transportation area is inheriting existing cyber-security challenges. Due to their connectivity and data-driven decision-making, the security of autonomous vehicles is a high-level concern since it involves human safety in addition to economic losses. In this paper, a comprehensive review is conducted to discuss the security threats and existing solutions for autonomous vehicles. In addition to that, the open security challenges are discussed for further investigations toward trusted and widespread deployment of autonomous vehicles.**

Keywords*—Internet of Things; smart transportation; autonomous vehicles; cybersecurity

I. Introduction

Ashton, in 1999, introduced the idea of connecting Radio Frequency Identification (RFID) Tags to the Internet which enabled the interconnection of conventional objects to handle autonomous tasks, and therefore, led to the new concept of the Internet of Things (IoT) [1]. With the help of IoT technologies, the transportation area is having a significant evolution from conventional mechanical vehicles to next-generation smart vehicles capable of collecting environmental data, process and communicating them, to make intelligent driving decisions without human assistance [2], [3]. Due to this new orientation toward intelligent transportation systems (ITS), the transport sector has become attractive for various interdisciplinary researchers and industries to work in the industrialization and deployment processes of autonomous vehicles (AV).

An AV is a computer on wheels that is equipped with a multitude of software and electronic components (such as sensors, processing units, and transmission modules), and capable of performing driving tasks on its own. In 2021, the Society of Automotive Engineers (SAE) provided an official updated reference which describes the evolution of vehicles in five automation levels, known as Level 1: driver-assistance; Level 2: partial-automation; Level 3: conditional-automation; Level 4: high-automation; and Level 5: full-automation [4]. In level 1, the human driver has full control over the vehicle's driving mode, but some computing systems are embedded to assist the driver in monitoring the environment (e.g. measuring of distance between vehicles to produce collision alerts, over-speeding alerts, …). In level 2, the human driver has partial control over the vehicle's driving mode, where some vehicular functions are controlled by automated systems (e.g. executing steering and acceleration functions). In level 3, most of the vehicle's driving functions are automated to enable a self-driving mode, but the human driver must necessarily respond to feature requests and take driving control in some situations. In level 4, the vehicle is highly automated and capable of handling driving functions without requiring much intervention from the human driver. In level 5, the human driver has no control over the vehicle's driving mode, and all the driving functions are performed by computing systems in all situations.

The main objective of introducing AVs is to avoid human-based errors that are mostly the cause of traffic issues like accidents, and therefore, to increase traffic safety, mobility, and efficiency [5], [6]. However, the transfer of the driving mode from human hands to computer systems also exposes vehicles to existing cybersecurity challenges, where attackers can gain access to AVs and control them for malicious ends making it urgent to consider their security at a high level for trustworthy developments and deployments. In this paper, a comprehensive review is conducted to give a state of the art of vehicular security, which provides the following key contributions:

- First, it describes the architecture of AVs, highlighting the role of different components, their security and privacy requirements, and the threats and vulnerabilities that can affect their normal operations.
- Second, it explores the literature to identify the reported attack methods against AV components as well as existing solutions that can be used to mitigate attacks.
- Third, it discusses the existing countermeasures to highlight their advantages and limitations, points out open challenges then proposes research directions for further investigations.

In the rest of this paper, Section II provides an overview of the architecture of AVs. The literature review is reported in Section III. Section IV discusses the existing security methods

for AVs. Section V proposes new research directions that can be considered to provide efficient security methods for AVs. The paper is concluded in Section VI.

II. Architecture of Autonomous Vehicle: Overview

The internal components of AVs can be classified into three main layers including the input layer, processing and control layer, and communication layer as depicted in Fig. 1. The input components collect data from the environment which are used by processing components to make decisions for controlling the vehicular functions. The communication components serve as interfaces allowing the interaction between different internal components and also between the vehicle and external entities such as other vehicles, personal devices, and infrastructures.

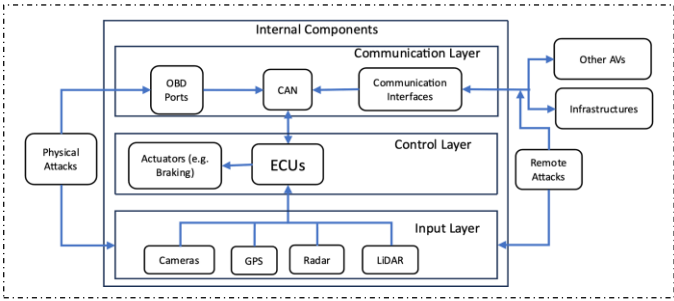


Fig. 1. Components of autonomous vehicles.

A. The Input Layer

AVs are equipped with a large number of sensors to collect specific data from the environment serving as input for control units [7]. Some commonly used sensors are Cameras, "Light Detection and Ranging (LiDAR), Global Positioning System (GPS), Radio Detection and Ranging (Radar), Ultrasonic", etc.

LiDAR sensors are used to detect surrounding objects by sending light waves and calculating the distance based on reflected signals [8]. In the same logic, Radar sensors measure the distance and speed of objects by sending electromagnetic waves in the radio domain and sense the reflected signals. Used for obstacle detection, the Radar works better in bad weather compared to the LiDAR and both are most used for long-range detection whereas ultrasonic sensors are preferable for short-range measurements based on sound waves. No matter the performance of obstacle detection sensors, they cannot identify the colour of a traffic light. Therefore, image sensors (cameras) are used to provide vision capability to the AV and identify different entities in the environment. GPS sensors operate by receiving radio signals from three or more satellites to determine the geographical location [9]. Therefore, GPS sensors are necessary for AVs to localize them and find routes between different locations. These sensors are useful in many applications and serve in AVs to observe the environment and make intelligent decisions to control the vehicle for safe driving and efficient navigation as illustrated in Fig. 2.

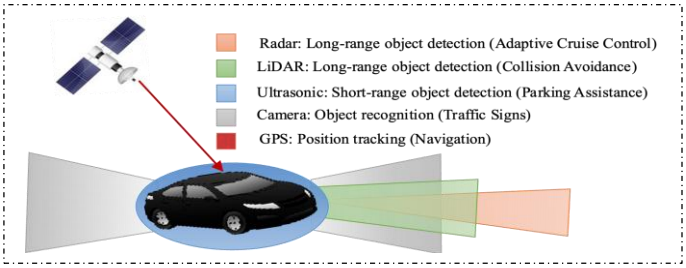


Fig. 2. Use of sensors in AVs.

B. The Control Layer

Electronic Control Units (ECU) represent the brain of AVs. ECUs are embedded systems that receive input signals from other components mainly sensors, process them, and decide the behaviour of vehicular functions [10]. As illustrated in Fig. 3, several types of ECU are used in AVs to perform specific tasks ensuring that vehicular functionalities are well-controlled and operational [11]. These include engine control, speed control, body control, tire-pressure monitoring, transmission modules also called telematics ECUs, and other internal measurement systems [12]. The ECUs can work together or independently based on the required action to take. For example, after detecting a pedestrian or other obstacles, the braking and speed control systems can collaborate to avoid collisions.

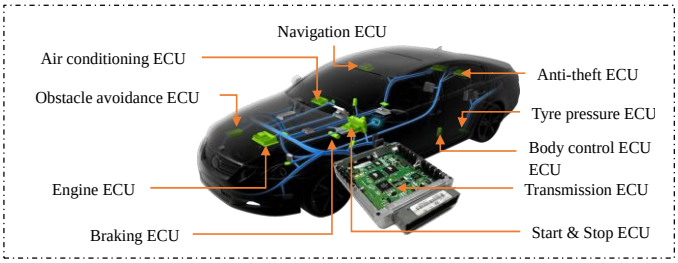


Fig. 3. Embedded ECUs in AV.

C. The Communication Layer

The internal components of AVs such as sensors, ECUs, and actuators, are interconnected through the in-vehicle network also known as "Controller Area Network (CAN)", where they exchange data to perform tasks together [7]. Short-range technologies are mostly privileged for in-vehicle communications to establish wireless connections between sensors and ECUs to reduce complex wiring. Also, external devices can be physically connected to the vehicle through onboard diagnostic (OBD) ports to access the CAN data for diagnostics on components or firmware updates [13]. The OBD ports can be found in any modern vehicle including existing AVs and are commonly used to access and update embedded software in the vehicle's control units.

For external communications, multiple AVs can form a network following the paradigm of vehicular ad-hoc networks (VANET) [14], where each AV can communicate with others and with surrounding communications infrastructures such as roadside units (RSU) as illustrated in Fig. 4.

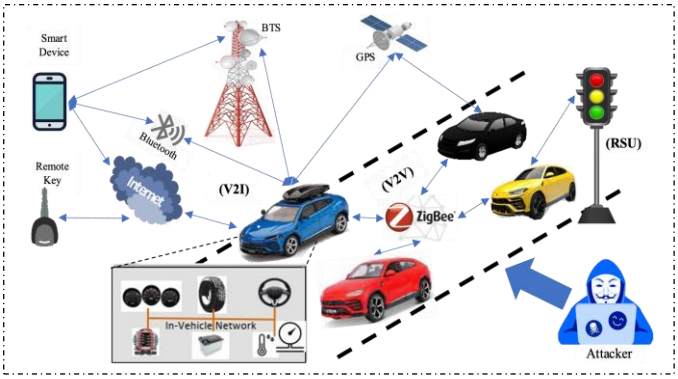


Fig. 4. Network of autonomous vehicles.

Typically, VANETs achieve four main types of communication listed as follows:

- Vehicle-to-Vehicle (V2V) communication: This allows vehicles to share traffic safety information and their status information like speed, position, and direction, to maintain good and safe driving conditions [15].
- Vehicle-to-Infrastructure (V2I) communication: This allows vehicles to interact with RSUs (V2R) for traffic safety information such as accidents, congestion alerts, and various warning messages. In this phase, vehicles can also communicate with other infrastructures such as satellite and Cellular for receiving their navigation-related data and other remote communications [16].
- RSU to RSU (R2R) communication: This allows nearby RSUs to interact and share network status information.
- Vehicle to Everything (V2X) communication: The V2X represents all the vehicular interactions with a large range of communication entities, such as smart devices, smart homes, pedestrians, clouds, computers, cellular networks, etc. [17]. This also includes V2V and V2I.

The vehicular interactions use different types of wireless communications protocols including short-range technologies (e.g. ZigBee, Bluetooth, and ultra-wideband (UWB)); medium-range technologies (e.g. Wi-Fi and "dedicated short-range communication (DSRC)"); and long-range technologies (e.g. Cellular Communications) [18]. The DSRC also known as "Wireless Access in Vehicular Environments (WAVE)" is adopted for V2V and V2R communications whereas cellular technologies are preferable for other V2I communications [19]. In addition to that, the AV uses the WIFI interface to interact with cloud and mobile applications for remote control [20].

D. Security and Privacy Requirements for AVs

The input, processing, and communication layers work together to create a well-functioning driving capability for AVs and their communication environments [21], but compromising any layer can destabilize the vehicle leading to harmful damage with direct consequences on human safety and economy [22]. Therefore, the security of AVs is a high priority and should cover all their internal and external interaction aspects. Some common security requirements are given as follows:

1) Availability: The internal components of AVs must remain accessible for collecting, processing, or transmitting data to ensure continuous operability. Also, the vehicular networks must stay available for receiving and sending safety-related information even during critical conditions such as high mobility. Therefore, AVs must be secured against attacks that can result in availability issues.

2) Authentication: This is a primary security measure where each node must be able to identify the source node that has sent a given message before going through further interactions. Therefore, vehicular networks must be secured against intrusions of malicious nodes to prevent attacks. For the sake of real-time requirements, rapid authentication methods are preferable to minimize communication delays.

3) Confidentiality and integrity: The exchanged messages between different nodes must only be accessible by the authorized members and each node must be able to verify that the received message was not modified or altered during transmission. The cryptographic algorithms are commonly used to achieve confidentiality and integrity requirements, but again, rapid encryption methods are necessary for vehicular applications to avoid added communication delay.

4) Privacy and anonymity: As a large amount of data is collected by AVs, processed, or transmitted over the network, the private information of users must not be exposed to unauthorized parties. This requires strict protection of identification information against potential privacy leakages.

5) Monitoring: In presence of multiple attacks, vehicular networks must be controlled to identify malicious nodes and actively remove them from vehicular communications through an appropriate authority. Therefore, real-time monitoring methods are required to efficiently prevent potential attacks.

E. Security and Privacy Threats Against AVs

Depending on the attack opportunity, attackers can reach AV components using remote interfaces or through physical access [23]. In the remote attack, any component capable of interacting with the surroundings can be vulnerable where an attacker can perform different types of attack aiming to steal information, to control a vehicular function, or to interrupt an operation [9]. In the physical access attack, the attacker can inject malicious codes into the vehicle's system using the onboard ports, physically damage a component, or insert an additional fake component to transmit wrong data into the vehicle's system [24], [25]. Some common attack vectors on AV components and communications are given as follows:

1) Sensor spoofing: The attacker manipulates and generates fake signals stronger enough to force sensors to detect and transmit wrong data [26]. This attack aims to control the decision-making of a targeted ECU which will receive the collected data and make wrong decisions. For example, if the GPS sensor detects stronger signals from the attacker, the navigation ECU of the vehicle can decide to

follow a different trajectory which is intended by the attacker as illustrated in Fig. 5.

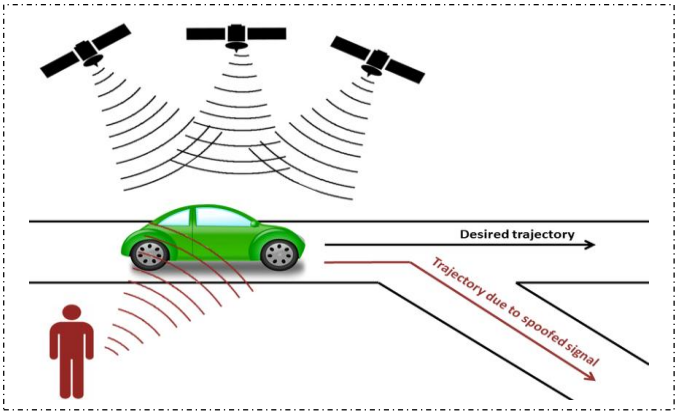


Fig. 5. Illustration of AV's GPS spoofing attacks [27].

2) Sensor jamming: It consists of blocking the sensor's perception by sending noise signals to interfere with normal signals [28]. This attack can interrupt operations of an ECU that depends on data from the targeted sensor. For example, if obstacle detection sensors are not able to collect data, the speed control ECU can decide emergency braking which can lead to accidents and traffic congestion as illustrated in Fig. 6.

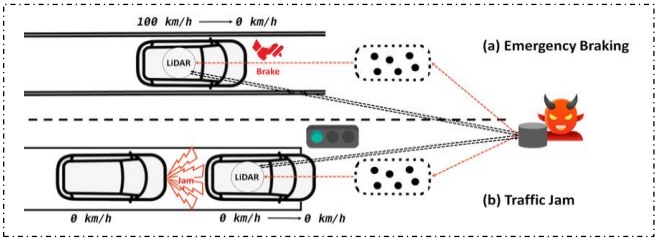


Fig. 6. Illustration of AV's sensor jamming attacks [29].

3) Blinding and adversarial images: The attacker can use strong light beams to blind or confuse the perception of the targeted camera [22]. Adversarial images generally target the machine learning models that are used for image recognition, where attackers manipulate images with adversarial samples that appear to be normal to human eyes but can cause huge confusion to the model producing incorrect outputs [30]. For example, if the attacker manipulates a stop road sign, the vehicle can misinterpret the captured image as a speed limitation and therefore speed up instead of slowing down, which can lead to harmful accidents as illustrated in Fig. 7.

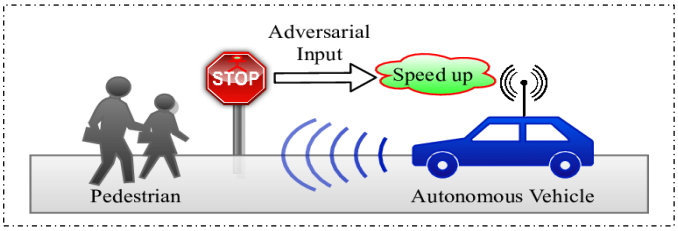


Fig. 7. Illustration of AV's camera attacks [31].

4) Malware and message injection: The attacker runs malicious code in the AV's system using the OBD ports, by flashing into its memory or through the process of firmware updates [32]. Also, the attacker can inject fake information through vehicular communications and force vehicles to take action on wrong data performing the intended activities [33]. These attacks aim to execute a specific task in a targeted ECU or interrupt its normal functionality as illustrated in Fig. 8.

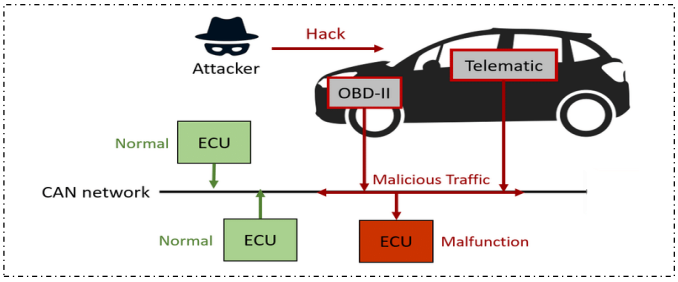


Fig. 8. CAN network attacks [34].

5) OBD Attack: Historically, dedicated handheld tools are used to scan information through OBD ports but most modern OBD devices such as Telia Sense [35] and AutoPi [36], allow connection with personal computers and smartphones for self-diagnostics purposes. As shown in Fig. 9, an attacker can use a compromised OBD device to access the vehicle's system which can allow executing a malicious program in targeted ECUs to control the vehicular functions.

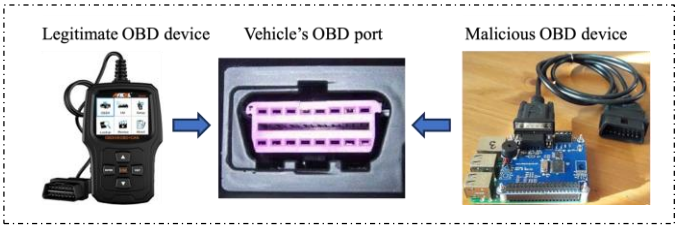


Fig. 9. OBD attacks.

6) DoS (Denial of Service) and DDoS (Distributed DoS): This attack aims to create unavailability of a service. It can target protocols and networks by sending excessive bad traffic packets to disrupt communications [37]. This attack can isolate a targeted vehicle from communicating with others or block the entire vehicular network leading to unwanted traffic conditions. The DoS and DDoS attacks can target AVs in both the internal and external communication aspects.

7) Eavesdropping: The attacker can intercept the exchanged information between different entities and secretly analyse or modify them to serve a malicious goal [37], [38]. The Man in The Middle attack is a typical example where the attacker modifies data between different entities and lets them believe that they are communicating with each other. This attack aims to access private information or to gain control over the vehicle's behaviour.

8) Message replay: The attacker can retransmit the past traffic status information in the vehicular network to mislead vehicles into believing that the action is currently happening

[28]. This attack can create traffic disorder and allow attackers to attempt their goal causing accidents or congestions.

9) Black hole: The attacker can manipulate the network informing other nodes that the malicious node has the shortest path to their destination [39]. After receiving packets, the malicious node will drop all of them to cause data loss blocking other nodes from receiving safety information. This attack can disrupt vehicular communications with the illusion of a normal driving environment.

10) Sybil: This attack consists of creating multiple fake identities in the network known as Sybil nodes which are used to transmit fake information to other nodes [28], and therefore, provoke the illusion of a busy network.

11) Physical damage: The attacker can target a specific component in the vehicle and destroy it [21]. This attack aims to interrupt the normal operations of the targeted component.

III. Literature Review

With the rapid development of connected and autonomous vehicles, several studies and hypothetical demonstrations in the literature have raised security weaknesses. In recent studies reported in [40] and [41], connected vehicles were targeted by more than 1300 reported cyberattacks from 2010 to 2023 and the analysis has shown that the attack frequency increased by 225% from 2018 to 2021, where 85% of attacks used remote interfaces and 54.1% were done by malicious actors leading to system control, vehicle theft, and unauthorized access to private data of users. This section exposes attack methods against AVs and the proposed countermeasures.

A. Sensors: Attacks and Defences

In general, sensors simply collect data and transmit them for further processing without authentication [21]. This fact makes sensors vulnerable to spoofing and jamming attacks.

1) LiDAR Sensors

Attacks: Authors in study [42] reported a successful jamming attack on a LiDAR sensor, the "ibeo LUX3 model", which consisted of sending higher intensity light to the LiDAR and blocking the acquisition of legitimate reflected light waves. Two variations of attack against LiDAR sensors were also demonstrated in study [43], where authors first showed the possibility of manipulating different LiDAR sensors of the same AV to perceive objects farther or closer than their real locations, by recording signals sent from one LiDAR sensor and then relaying those signals to the other LiDAR sensor. The second attack was to send fake signals to the targeted LiDAR sensor and make the vehicle believe that it was approaching a large obstacle. Authors in study [44] also performed this second attack on a LiDAR sensor, the "VLP-16 model", and explained that most LiDAR sensors are vulnerable to this attack especially those with large receiving angles.

Defences: Authors in study [45] proposed a new LiDAR scheme to detect jamming attacks based on random modulation of light waves. This modulation creates four polarization states (horizontal, vertical, diagonal, and anti-diagonal) for the photon and the jamming attack is detected based on a comparison of distances measured from the states. However, the authors acknowledged that their proposed scheme cannot filter jamming signals out of legitimate signals. Authors in study [46] proposed a method to detect fake input signals from LiDAR sensors by using the previous data frames to build a momentum model. However, building this model would require high computational power and time which is not adequate for real-time and resource-constraint applications such as AVs. The use of multiple LiDAR sensors was proposed in [44], to have overlapping views of the vehicle's surroundings or to reduce the signal-receiving angle for each sensor. This technique can reduce attack chances by preventing spoofing attacks on all the sensors at the same time, but requires a high number of LiDAR devices to cover all the vehicle's surroundings and therefore increases the cost. Authors in study [47] also proposed embedding identification data onto LiDAR's light waves by modulating them together, which allows sensor nodes to authenticate the received signals and therefore prevent LiDAR spoofing attacks. An experiment was conducted in [48], and the authors concluded that it becomes more difficult to succeed in LiDAR spoofing attacks when object detection is based on machine learning (ML) models. Authors in study [49], also confirmed the effectiveness of using ML models to detect LiDAR spoofing attacks. A LiDAR spoofing mitigation algorithm was proposed in [50] to detect adversarial objects and non-existing obstacle attacks where authors claimed correct attack detection based on simulation results.

2) Radar sensors

Attacks: A spoofing attack on a Radar device, the "Ettus Research USRP N210 model", was experimented with [51], by recording the broadcasted Radar signals to modify their phases and re-broadcast them to Radar sensors. This caused incorrect distance calculations and resulted in perceiving objects at a 15-meter distance while the real distance was 121 meters. Authors in study [52] demonstrated that it was possible to manipulate an object's velocity together with distance using spoofing attacks on FMCW Radar sensors, by designing an adversarial Radar to simulate two scenarios of attack provoking emergency braking and acceleration in a victim vehicle. Authors in [53] performed a similar attack on an FMCW Radar using a semi-passive modulated transponder and reported that it is possible to confuse a radar perception with ghost targets at different distances and velocities by simply changing the modulation frequency of the transponder without the need to use complex techniques. Also, authors in [54] demonstrated a jamming attack on Radar sensors for manned and unmanned aerial vehicles (UAV) where the attacker can modify the amplitude and the frequency of the recorded signals and then re-broadcast them to the Radar sensor to cause failure in object detection.

Defences: To mitigate spoofing and jamming attacks, a "physical challenge-response authentication (PyCRA)" was proposed in study [55], which sends random signals called challenging signals in the Radar sensing environment and detects fake signals based on a noise threshold. The PyCRA shuts down the sensing signals at random times, which was criticized to potentially affect AV safety-critical components by the authors in [56], who proposed an alternative method

called "Spacio-Temporal Challenge-Response (STCR)" and claimed to achieve better performance by transmitting challenging signals in random directions together with sensing signals instead of shutting them down. Once a malicious signal is detected, the reflected challenging signals are used to identify the attack directions and exclude them. Authors in [57] experimented with an unsupervised deep-learning method on Radar system data to detect manipulation attacks and found an accuracy of 88% detection rate. They defended that their technique could be used in AV's Radar systems to mitigate spoofing attacks, by learning the correlation between categorical and numerical features from Radar signals.

3) GPS Sensors

Attacks: The GPS spoofing attacks were analysed in study [58], where authors explained how it can be easy for an attacker to carry out GPS spoofing attacks by using hardware capable of generating stronger GPS signals, broadcasting them to GPS receivers in a chosen environment to force them to switch from legitimate satellite signals and manipulate their location calculations. In 2013, a man was arrested in New Jersey for using a GPS device that was interfering with GPS ground-based receivers of Newark's Liberty Airport [59]. This device was able to block surrounding GPS receivers from receiving legitimate GPS signals and he claimed using it in the company truck simply to hide from his employer. In [60], authors demonstrated a successful GPS attack using a low-cost device assembled from conventional components and were able to manipulate the navigation data of 38 real cars out of 40 participants to follow a wrong predetermined destination without being noticed. The authors discussed that this attack may not succeed when the driver is familiar with the location but it represents a high risk for self-driving vehicles. Authors in study [61] also demonstrated a new approach to GPS attacks that can succeed in manipulating navigation routes on vehicles where security mechanisms are used such as internal navigation system (INS). The technique consists of exploiting existing navigation data between the vehicle's start and destination points to identify similar routes with the original trajectory using an algorithm and then forcing the vehicle to follow the most similar trajectory. The authors claim that this attack can be successfully executed due to the negligible inconsistencies between the original and the spoofed routes. Also, authors in study [62] proposed a spoofing generator that cancels all legitimate GPS signals and allows surrounding GPS receivers to collect the attacker's generated signals. The authors defended that their spoofing generator can cover all open-sky satellites making this attack difficult to detect based on a comparison of signal consistency from different GPS receivers.

Defences: To prevent GPS attacks, the use of multiple antennas was proposed in study [63], to receive GPS signals and measure their phase differences to detect spoofing attacks but this technique would be inefficient under attack methods as presented in study [62]. Authors in study [64], proposed the use of coding in GPS systems to reduce jamming attacks where GPS signals are encoded and modulated by the satellite before transmitting to receivers that will then demodulate and decode to recover the original GPS signals. However, this method requires changes in GPS satellites which is very difficult and also, the authors acknowledged that their method is less effective when the jamming signals are too strong than the legitimate ones. In many applications, a technique known as "Receiver autonomous integrity monitoring (RAIM)" is used in which, the observed GPS signals are compared with the expected signals to determine the integrity of the received signals [65]. The RAIM uses a pseudo-ranging measurement to produce several GPS positions based on redundant signals [66]. However, the Advanced RAIM (ARAIM) used as an extension for other navigation systems beyond GPS, was criticised in study [67], for having availability issues when one or more satellites cannot be reached. Later, authors in [68] proposed a solution to improve the availability of ARAIM up to 98.75%. Authors in study [69], proposed integrating transmission signatures into GPS satellites which will allow GPS receivers to authenticate the received signals. This method can easily help to detect spoofing attacks but it would involve higher costs for changes in satellites. A rotation-based technique was proposed in study [70], for GPS receivers that can help to determine the angle of arrival of GPS signals from different satellites and compare them to detect spoofing attacks. Authors in study [71], proposed a GPS spoofing detection technique for vehicular GPS receivers based on the Doppler Shift associated with them. In their approach, authors intentionally perturbed the vehicle's velocity and observed the changes in Doppler Shift value if they were consistent with velocity variations or not. Due to the unpredictability of these variations, spoofing signals cannot follow the changes. A GPS spoofing mitigation technique was proposed in study [72], based on the Isolation Forest that consists of detecting the attack and isolating the compromised GPS receiver before correcting it using the location data of roadside units. The authors claimed to achieve good results but this method would require the use of multiple GPS receivers to avoid service interruption. Authors in study [73] and study [74], demonstrated through simulation that machine learning and deep learning algorithms can achieve detection of both GPS spoofing and jamming attacks with high accuracy.

4) Image Sensors (Cameras)

Attacks: A blinding attack was experimented in study [43], on a car's camera (MobilEye C2-270) where authors projected different light beams on it. First, an LED matrix of 940nm 5*5 and an LED spot of 850nm were used and able to blind the camera from perceiving images which took 5 seconds to recover later. A 650nm laser was then used on it to achieve the same results but the camera never recovered again. Authors in [75] conducted similar experiments to permanently blind a camera and concluded that both LED and Laser beams can blind cameras with enough intensity but infrared beams can make exceptions due to their narrow frequency band. Attackers can also use other methods such as manipulating images to cause incorrect predictions of road signs by ML-based algorithms as described in studies [76], [77], [78], [79]. In 2017, Google researchers created stickers with patterns and attached them to some important road objects such as speed limitations and stop signs [80]. The authors claimed that the stickers were able to provoke incorrect predictions in the used algorithms. Authors in study [81], experimented with similar attacks by decorating stop signs with many black-and-white stickers and found a failure of 100% of the algorithm to

recognize the stop signs with a fixed camera and 84.8% with a moving camera on a vehicle. Authors in study [82], experimented with a blinding attack using electromagnetic waves to interfere with cameras and were able to cause incorrect observation of stop signs. This attack is hard to detect because it does not require a physical modification. A similar experiment was conducted in study [83], using invisible infrared lights which affected the captured image's pixels with a magenta colour in ambient light. The authors succeeded in perturbating cameras on the Tesla Model 3 using off-the-shelf IR light sources and confirmed the effectiveness of their attack in various settings.

Defences: To mitigate camera blinding attacks, the authors in [43] proposed two solutions. The first consists of using multiple cameras in the AV to capture redundant images and avoid single-camera failure. This strategy makes it difficult to attack all cameras simultaneously due to the limited beam widths of LED and Laser spots but can increase the cost according to the number of cameras. The second solution is to integrate a light filter into cameras that can cut near-infrared lights. This strategy can be implemented at a low cost but lacks experiments to confirm its effectiveness. The authors in [83], proposed to implement infrared light filter-based software on cameras as mitigation to their attack. Authors in [84], proposed the use of ML algorithms to predict images and compare them with the captured images and claimed that this technique can help to detect blinding attacks and take the required actions before any damage. The use of machine learning models as a solution against adversarial image attacks has also been discussed in the literature. In these models, three major aspects are considered including pre-processing input images as detailed in [85], [86], [87], [88], training with adversarial image samples as proposed in [89], [90], [91], and detecting adversarial inputs using run-time information as described in study [92]. These models can be integrated into ECUs which receive their input images from cameras for the AV's vision.

B. Control and Processing Units: Attacks and Defences

The internal components of the AV can exchange information through the CAN network, where ECUs receive their input data from the different sensors, communication interfaces, and/or other ECUs [93], [94], [95]. Therefore, any failure from the input source can directly influence the ECU to give incorrect output. Also, attackers can observe the CAN messages and inject malicious data through OBD ports or telematic interfaces to target a specific ECU.

1) ECU Attacks: Authors in study [96] experimented with an attack by connecting a laptop to a vehicle's OBD port to access the CAN network and run a custom code named CarShark in targeted ECUs, which was able to compromise their initial functions. They warned that no security measures were applied during vehicle software updates through the OBD port. Authors in study [97] have also shown that an attacker can develop a malicious program and let a vehicle owner download it as a self-diagnostic application which will allow the control of ECUs through OBD connexions. Authors in study [98], were able to access a vehicle's ECUs through Bluetooth and long-range connections allowing them to analyse the firmware and execute their codes. The authors reported that an attacker can use the same process to remotely inject malicious codes in a targeted ECU and compromise its functions. In Black Hat 2015, some researchers demonstrated a successful attack on a Jeep Cherokee ECUs using remote interfaces [99]. The authors were able to control the vehicle's braking, steering, and acceleration systems. Authors in study [100], focused a study on discovering weaknesses in the deployed access control and communication mechanisms on Tesla vehicles, the "P85 and P75 models", and demonstrated how ECUs can be remotely controlled by sending malicious packets to the CAN via wireless technologies. Authors experimented with their attack, to remotely control the steering ECU of the Tesla Model S 75 [101]. Authors in study [102], demonstrated CAN vulnerabilities using experimental fuzzy-testing and reported that an attacker can easily access the CAN network and control a targeted ECU of the vehicle with necessary protocol analysis tools.

2) ECU Defences: Many mitigation solutions have been proposed in the literature to prevent ECU attacks. Authors in [103], proposed an attack detection technique which calculates entropy during normal and abnormal CAN communications to detect suspicious activities. The authors in study [104], proposed a technique to monitor all messages in the CAN network where each ECU will use a flag to indicate a message transmission time, and therefore, detect unauthorized messages based on the time threshold. This method was criticized in study [105], because it requires modifications in every ECU in the vehicle, then proposed to use identity checking of ECUs and observe the frequency of their message transmissions. If a significant change in frequency is detected from an ECU, then it can be considered compromised. A similar technique was proposed in stuy [106], to detect abnormal massages based on interval measurements of periodic CAN messages. Authors in study [107], proposed to use a machine learning-based device that can be connected to OBD ports to detect malicious patterns from the CAN traffic data and disable the messages when an attack is detected to prevent ECUs from being compromised. Authors in study [108], proposed to implement a hardware-based protocol that can achieve both CAN access authentication and message encryptions. Authors in [109], proposed a technique to monitor the correlations between ECU messages and estimate the behaviour of the vehicle. In this method, a specific ECU is detected as compromised when there is a sudden change in its messages but a sudden change in the vehicle's behaviour would mean multiple ECU attacks. Authors in study [110], proposed an Intrusion Detection System (IDS) to detect CAN network attacks based on ML algorithms. They experimented with their model using a CAN dataset and claimed to successfully classify DoS and Fuzzing attacks with high accuracy. Authors in study [111], proposed a secure boot scheme based on cryptographic algorithms that can protect the CAN network from malicious software being executed by the

vehicle's ECUs. After experiments, the authors claimed to achieve good performances with the Cipher-based MAC (CMAC) and the elliptic curve digital signature (ECDS) algorithms in terms of authentication and execution speed. Authors in study [112], also proposed an ML-based anomaly detection for CAN networks using the deep autoencoder method and claim to achieve high detection accuracy of up to 99.98%.

C. Vehicular Communications: Attacks and Defences

Vehicles are mobile and their interactions with various external entities make them vulnerable to several cyberattacks.

Attacks: Authors in study [113] conducted a simulation on a group of cooperative driving AVs where they experimented with an attack to compromise one of the vehicles and then used it to transmit false information in the vehicular network, which resulted in sudden disturbances in vehicles' speeds. A DoS attack was also experimented in study [114], by saturating a V2I network channels with excessive noise messages through simulation, where authors showed that this kind of attack in practice, can block all vehicles from sending messages in the network and therefore interrupt the vehicles' cooperation. In study [40], a group of researchers conducted a study to explore the security of automotive APIs, telematic systems, and the infrastructures that support them. The authors discovered multiple vulnerabilities across 19 major global suppliers and original equipment manufacturers (OEM) and exploited them to remotely control vehicles and access sensitive data.

1) Authentication defence mechanisms: Authors in study [115], proposed an authentication method for V2V communications in VANET, where vehicles periodically broadcast their presence information to others and record the received announcements to determine a neighbouring group, and then identify malicious nodes by sharing the composition of groups. Authors in study [116] proposed a V2I authentication method called "Security Credential Management System". The method was based on a public-key infrastructure and claimed to provide good privacy protection, but it suffers from high computation and communication delays.

Authors in study [117] proposed an authentication method to achieve group signatures for short-term communications in VANET based on the Boneh-Shacham algorithm. Authors in study [118] also proposed an authentication system for VANET which generates pseudonyms based on vehicles' IDs through public key cryptography and then uses these pseudonyms in the authentication processes for privacy-preserving purposes. They used an ID-based signature for V2I authentication and an ID-based online/offline signature for V2V authentication and defended the feasibility and efficiency of their method in vehicular networks based on performance evaluations.

Authors in study [119] proposed a lightweight authentication method for handover in V2X communications where each vehicle can be allocated a temporary identity from its home network and then use that identity when moved to a new network. The authors claimed to achieve better performance with low computation overhead through simulations. Authors in [120] proposed a security technique for vehicular LTE networks which can mutually authenticate vehicle nodes and preserve their privacy. They evaluated their method to have better performances in terms of communication cost, security level, and less computation. A privacy management algorithm based on hybrid cryptography was proposed in study [121], to ensure trusted communication between vehicles. The authors used an asymmetric identity-based digital signature and claimed to achieve better performances in terms of communication latency, computation and storage overheads.

Authors in study [122] proposed a self-checking authentication method for VANET where vehicles and RSUs can verify each other without including a Trusted Authority (TA). Initially, the TA is responsible for the registration of all vehicles and RSUs before they join the vehicular network environment and therefore TA will intervene in the vehicle's authentication process through RSUs. This method proposed to allocate a group signature to vehicles at their first authentication from one RSU domain and then use the same signature for authentication in other domains without going through the whole process. Authors claim that this method meets security requirements and benefits from faster authentication.

Authors in study [123] proposed a multifactor authentication process for AVs and claimed to achieve good security checks without revealing sensitive information of users. Authors in study [124] also proposed a multifactor authentication for remote vehicle diagnosis and maintenance which requires both biometric and password verifications from the vehicle's owner or the Service Centre to ensure legitimate access to the system. Through performance analysis of the technique, the authors claimed that it achieves a robust security level. An edge-based vehicular authentication architecture was proposed in study [125], where different vehicles can be grouped to form a vehicular cloud. The authors claim easier attack detections using deep learning algorithms in this technique that offer a lightweight authentication of vehicles for secure V2V communications.

Authors in study [126] proposed an identity-based cryptographic method for V2V authentications and security key agreement, where the ID of each vehicle is used as its public key, which can expose this method to privacy leakage. A Blockchain-based One-Time authentication method was proposed in study [127], for V2X communications. In this method, the identities of nodes are encrypted before sharing instead of revealing the real identities, and different proofs are generated to authenticate nodes which are verified through a noninteractive blockchain. Based on security analysis, the authors claim to achieve secure V2X authentications with reduced delay. Authors in study [128], proposed an aggregate and continuous authentication technique using federated learning for VANET applications. Based on the edge devices as learning centre between vehicles and RSUs, the authors claimed to achieve a secure and privacy-preserving authentication with reduced communication overheads.

2) Confidentiality defence mechanisms: In study [129], a "cryptographic mix-zone (CMIX)" algorithm was proposed to secure exchanged data in vehicular networks. In CMIX, the encryption process is based on a group secret key, where the same key is shared between all the vehicle nodes in the network to save time from individual key sharing. However, all the security is compromised when an attacker can intercept the encryption key during its broadcasting.

Authors in study [130] proposed a game theory technique known as the Markovian game to achieve secure communication, where each vehicle in the network is considered a player and players are either data holders (DH) or data requesters (DR). In this game, each vehicle earns income according to the provided services and uses that income to buy access to private data. If a DR node wants to access private data from a DH node, it will propose a motivation price then the DH will decide a privacy concession according to its satisfaction. The problem with this game is that the DH cannot verify if the DR is a malicious node or not before privacy concession. Therefore, the algorithm should consider other parameters to prevent network intrusion.

Authors in study [131] proposed a secure V2X communication method based on a hash chain of secret key cryptography and claimed to provide secure messaging between vehicles at low cost. The authors in [132] proposed a batch verification method using the "Paillier cryptographic algorithm" to solve privacy issues in VANETs, in which, vehicles can cooperate to identify malicious users without disclosing sensitive information. This method can achieve privacy-preserving communications but does not guarantee the confidentiality of exchanged data.

3) Network monitoring defence mechanisms: Authors in study [133], proposed a security algorithm to detect malicious vehicles based on their behaviours in the network and then isolate them from the rest of entities. This method can reduce the chances of successful sybil attacks but isolating a vehicle node from the network as a prevention technique can represent a danger in some situations, especially in complex traffic. Authors in study [134] proposed an intrusion detection system to detect attack scenarios against vehicular networks including packet duplication, selective forwarding, resource exhaustion, wormhole, black hole, and Sybil attacks. After simulation, the authors claim that the proposed algorithm can provide good attack detection accuracy with minimum detection time.

Authors in study [135] proposed an intrusion detection system for vehicular networks based on "deep neural network (DNN)" to detect attacks. Using an unsupervised training, the DNN algorithm could accurately classify normal packets and attacked packets and able to detect malicious events against the vehicle as a result. Therefore, this method can perform better in vehicular networks when many attacked packets from different attack scenarios are used in the training process of the DNN.

Authors in study [136] proposed a voting technique to identify rogue nodes in VANET. In this method, two vehicles vote for each other when they can communicate without any security issues. The trust level of a vehicle in the network is evaluated according to the number of gathered votes, where a vehicle with a small value of a vote is considered a rogue node and a potential source of attack. This method can achieve good performance in a fixed number of nodes since it is an experience-based system but does not perform in scalable network scenarios like vehicular networks, where vehicles can join a locale network and leave at any time due to their mobility. Authors in study [137] proposed a coalitional security game to detect malicious nodes in vehicular networks based on Dempster-Shafer's theory. The game consists of building trusted relationships between vehicles based on their reputation, experience, and knowledge. However, attackers can target a vehicle with experience and a good reputation and use it to perform attacks in the network. Also, vehicles can be in new environments at any time due to their mobility without previous experience gained from that environment.

Authors in study [138] proposed an intrusion detection system using ML models to detect "Distributed Denial of Service (DDoS)" attacks in V2I communications. The authors claimed to achieve good detection accuracy after various testing. The same approach was proposed in study [139] based on the Support Vector Machine algorithm where authors achieved high DDoS detection performances through simulations. Authors in study [140] proposed an attack detection mechanism for AVs that works both in online and offline modes. The offline phase is used to establish parameters based on which, the detection of attacks and responses are executed in the online phase.

IV. Discussion

Security is a very active research area for information technologies, and the introduction of AVs has increased the interest where each newly published work can offer fresh approaches to system security problems. However, most of the existing security standards are still facing challenges in addressing issues in this cutting-edge technology where additional critical parameters are being considered regarding real-time communication, resource constraint computation, user privacy, fault detection, network scalability, quality of service, etc. [141]. The previous security mechanisms presented in the literature are discussed in this section to point out the open security challenges for further investigations.

A. Sensors

Sensors represent the perception elements that collect data for AVs and are used to make decisions. Therefore, the security of sensors is crucial and the proposed attack mitigation methods should consider some requirements including:

- Detection of adversarial signals such as spoofing and jamming attack signals, and filter them to allow good perception of legitimate signals under attack situations;
- Availability: the solution should not disrupt other services during its execution and should have faster execution to meet the real-time functionality of AVs;

The defence strategies related to the security of sensors are discussed in Table I.

B. The In-Vehicle Network

In the internal network of the vehicle, the components interact to perform driving tasks together, and any successful attack or dysfunction can compromise the vehicle's normal operations. Therefore, the security measures should include the following requirements:

- Authentication: Every component should be identified and trusted before accessing the CAN network;
- Attack detection: suspicious activities in the CAN network should be identified, and compromised nodes should be excluded from sending data in the CAN.
- Availability: the solution should not disrupt other services during its execution and should have faster execution to meet the real-time functionality of AVs;

The defence strategies related to the security of CAN networks are discussed in Table II.

TABLE I. DISCUSSION OF SECURITY METHODS FOR SENSORS

Attack Types	Target components	Proposed Defense Strategies	Contributions	Limitations
Sensor Spoofing and Jamming	LiDAR, Radar, GPS	1) "Prevention of spoofing and jamming attacks" using multiple sensors [44], [63]. 2) "Detection of spoofing and jamming attacks" based on signals' directions [55], [56]. 3) "Detection of spoofing and jamming attacks" using the ML models [48], [49], [48], [49], [50], [57], [73], [74]. 4) "Detection of spoofing and jamming attacks" based on signals' authentication [45], [47], [64], [69].	1) The use of multiple sensors provides an overlapping coverage of the vehicle's surroundings, which can therefore reduce attack chances since it becomes difficult to compromise all sensors together. 2) The detection of attack directions can help to reject signals arriving from them and therefore prevent attacks. 3) The ML models increase the detection accuracy of spoofing and jamming signals. 4) The authentication of signals through modulation or signature methods, helps to identify attack signals from legitimate ones.	1) The main challenge with the use of multiple sensors to mitigate spoofing or jamming attacks is its implementation increases the cost. 2) The detection of attack signals based on their directions can be performed when the nodes are fixed, which makes it inefficient for vehicular applications due to the mobility of vehicles. 3) The ML-based attack detection methods lack experiments to determine their efficiency in realistic AV environments. 4) The authentication of signals requires computational modifications in sensor nodes, which represents a high implementation cost and also involves higher computational delays.
		Open challenges: The spoofing and jamming attacks against sensors remain a serious challenge in the context of AVs. Most of the existing solutions aim to detect spoofing and jamming signals, but there is a lack of efficient methods to filter them and allow sensors to correctly collect legitimate signals in attack situations. Therefore, further investigations are needed to guarantee the security of sensors for AV applications.		
Sensor Blinding and Adversarial images	Cameras	1) Prevention of blinding attacks using multiple cameras [43]. 2) Prevention of blinding attacks using light filters in cameras [43], [83]. 3) Detection of blinding and adversarial image attacks using ML models [84], [85], [86], [87], [88], [89], [90], [91], [92].	1) Multiple cameras provide overlapping views of the vehicle's surroundings to capture redundant images, which makes it difficult to blind all cameras together, and therefore, reduce attack chances. 2) The integration of light filters into cameras can help to cut near-infrared lights and therefore prevent blinding attacks. 3) The ML models increase the detection accuracy of adversarial image attacks.	1) The main challenge with the use of multiple cameras to mitigate blinding attacks is that it represents a high implementation cost. 2) The implementation of light filters needs to be experimented on real cameras to validate their effectiveness for AV applications. 3) Complex and adequate dataset of adversarial examples are still needed to train ML models and extensively experiment them on realistic cameras to determine their effectiveness.

TABLE II. DISCUSSION OF SECURITY METHODS FOR THE CAN NETWORK

Attack Types	Target components	Proposed Defense Strategies	Contributions	Limitations
Malware and Message Injection	CAN network, ECUs	1) "Detection of CAN network attacks" based on entropy calculation [103]. 2) "Detection of CAN network attacks" based on message transmission time [104]. 3) "Detection of CAN network attacks" based on message frequency [105], [106], [109]. 4) "Detection of CAN network attacks" based on authentication of ECUs [105]. 5) "Detection of CAN network attacks" based on ML models [107], [110], [112]. 6) "Prevention of	1) The changes in entropy can identify irregular activities, which can be useful in detecting CAN traffic anomalies. 2) The of message transmission times comparison is useful in identifying malicious data if the transmission time is higher than expected, which can particularly prevent message replay attacks without heavy computations. 3) The frequency of messages can help to determine the behaviour of different nodes in the CAN network, and therefore, detect abnormal actions. 4) The authentication of ECUs can prevent other ECUs from receiving malicious data from unidentified or illegitimate nodes. 5) The ML models can detect complex attacks with high accuracy. They can analyze large volumes of data and easily adapt to evolving attack patterns using updated datasets. This can monitor the CAN network in real-time. 6) The encryption of messages can prevent unauthorized access to data and guarantee	1) The main challenge with entropy calculation is its high sensitivity to data distribution, which must be well modelled to provide meaningful entropy values. Also, small changes in data distribution can lead to significant changes in entropy values making it practically inefficient for vehicular security. 2) The use of message transmission time can negatively impact the network in some unexpected situations such as transmission delays due to congestions or routing issues can falsely flag legitimate messages as malicious. Also, this can be vulnerable when the attacker manipulates the time. 3) The attack detection based on the frequency of messages may only be useful for the security of sensors, which collect data at a regular rate. However, this method is inefficient for ECUs that randomly transmit messages based on the needs. 4) The authentication of ECUs requires computational modifications in each ECU of the vehicle, which represents a high implementation cost and also involves higher delays. 5) The challenge with ML models is that their Training and deployment require significant computational resources, which may not be feasible in constrained environments like ECUs. Also, there is a lack of experiments to determine the effectiveness of the ML models on realistic CAN networks. 6) Many existing CAN networks lack built-in support for encryption, which requires significant hardware or software

		CAN network attacks" using encryption techniques [108], [111].	the integrity of exchanged information in the CAN network.	upgrades. Also, it would be challenging to implement and execute encryption algorithms in CAN networks because of their limited processing power and memory.
		Open challenges: The AVs remain vulnerable to malware and message injection attacks through OBD ports or telematics. Traditionally, the OBD ports are protected by a physical lock, which does not guarantee effective security. However, no security method was found in the literature that can distinguish legitimate OBD devices from malicious OBD devices when connected to the OBD port. Therefore, further investigations are needed to guarantee the security of ECUs for vehicular applications.		

C. Vehicular Communications

Each AV is an autonomous system susceptible to joining a network environment where it will interact with everything using wireless interfaces. The exchanged information between the vehicle with the outside world will determine its driving behaviour in traffic, making it vulnerable to various attacks. Therefore, the V2X communications protocols should include the following requirements:

- Authentication: Every entity including vehicles, smart devices, and RSUs, should be identified and verified as trusted before accessing the vehicular network;
- Data protection: Exchanged information between entities in the vehicular network should be authentic and confidential to avoid unauthorized access;
- Attack detection: The vehicular network should be controlled to detect suspicious activities, and exclude compromised entities from sending data in the network;
- Compatibility: The solution should not disrupt other services during its execution.
- Computational efficiency: The solution should have faster execution with low implementation cost to meet real-time and resource-constraint functionality of AVs;
- Scalability: The solution should accept the changes in the number of network entities.

The defence strategies related to the security of V2X communications are discussed in Table III.

TABLE III. DISCUSSION OF SECURITY METHODS FOR VEHICULAR COMMUNICATIONS

Attack Types	Target components	Proposed Defense Strategies	Contributions	Limitations
Network Intrusion attacks (identity theft, Sybil, replay, …) Eavesdropping (man in the middle, data manipulation, …),	Communication Protocols	1) Collaborative vehicular authentication [115], [125]. 2) Public key-based vehicular authentication [116], [118], [121], [126]. 3) Group signature-based vehicular authentication [117], [122]. 4) Identity-based vehicular authentication [119], [120]. 5) Multifactor-based vehicular authentication [123], [124]. 6) Blockchain-based vehicular authentication [127].	1) The collaborative authentication is decentralized and reduces the risk of single-point vulnerabilities. As multiple entities validate a vehicle's credentials, this technique can help balance the load of authentication tasks and identify malicious nodes more effectively. 2) The public key authentication uses strong encryption algorithms suitable for large-scale networks like vehicular networks, to ensure the integrity and authenticity of exchanged messages. 3) The group signature cancels the need for individual authentications, reducing communication delays. It also allows vehicles to authenticate themselves without revealing their specific identity, which enhances user privacy. 4) The identity-based authentication uses simple structures rather than complex and heavy cryptographic algorithms, which reduces computation and communication delays, making it suitable for time-sensitive applications. Also, it is easier to integrate identity-based methods across different vehicular networks to meet specific security requirements. 5) Combining multiple authentication factors (e.g., password/PIN, biometric data, cryptographic keys) makes it significantly harder for attackers to compromise the system. 6) The blockchain provides robust cryptographic security and immutability, making it difficult for attackers to alter authentication records. 7) The ML models gain from the high	1) The challenge with collaborative authentication is that it involves multiple messages exchanged among vehicles or nodes, leading to increased network traffic, and therefore introducing communication delays. Also, this method can lead to security and privacy breaches because of the exchange of vehicles' identity that can be intercepted and manipulated by attackers to have access to the network. 2) The public key operations (e.g., encryption, decryption, and signature verification) are computationally intensive, which can introduce delays and may affect the real-time requirements of vehicular networks. Also, this method can put the security of the entire network at risk, when the private keys or the certificate authority that it relies on, are compromised. 3) Generating and verifying group signatures can be computationally intensive, especially in a large number of group participants, which can impact real-time applications and reduce the network performance in real-world vehicular environments. 4) While promising, identity-based methods are less commonly deployed compared to certificate-based systems, limiting their effectiveness in verification and interoperability with existing infrastructures. 5) The authentication based on factors like biometrics can frequently fail due to mismeasurements or environmental conditions (e.g., dirt affecting fingerprint scanners). Also, authenticating multiple factors can take additional computation time, which may affect real-time applications like collision avoidance or emergency communications. 6) Blockchain transactions can require significant time to be validated and added to the ledger, which may not meet the real-time requirements of vehicular networks. Also, the increase in the number of vehicles and authentication transactions can lead to blockchain bloat, requiring more storage and computational resources.

Network saturation (DoS and DDoS)		7) ML-based intrusion detection [128], [134], [135], [138], [139].	accuracy of attack detection capable of analyzing large volumes of data to detect complex attacks with the ability to adapt to evolving attack patterns by retraining with updated datasets. This can monitor vehicular communications on a real-time basis.	7) The challenge with ML models is that their training and deployment require significant computational resources, which may become intensive for vehicles in case of the implementation of multiple ML algorithms. Also, there is a lack of experiments to determine the effectiveness of the ML models on realistic vehicular networks
		8) Game theory-based intrusion detection [130], [136], [137].	8) The game theory models allow the system to predict and counteract attacker strategies effectively by making the cost of an attack higher than its potential benefit. A well-designed game-theoretic approach can enhance the precision of attack detection when the attacker's behaviour patterns are incorporated.	8) The main challenge with game theory-based security methods is that they require detailed knowledge of attacker and defender behaviours to create a realistic game-theoretic model. Due to this dependency on accurate input data, such as network traffic patterns and known attack strategies, the game theory model can easily fail to detect an attack effectively if the attacker uses strategies outside the modelled game.
		9) Data protection based on cryptographic algorithms [129], [131], [132].	9) Cryptographic algorithms provide robust protection against unauthorized access and ensure data confidentiality throughout its lifecycle, even when transmitted over insecure channels. They can be implemented across various systems including vehicular networks.	9) Cryptographic algorithms can be resource-intensive, leading to delays in real-time systems like vehicular networks. Choosing a stronger encryption algorithm often involves a trade-off between security and performance, particularly in resource-constrained environments like IoT-based systems.
		Open challenges: Due to the high mobility of vehicles, they are permanently vulnerable to attacks through interactions with potentially compromised entities. Therefore, strong, efficient, and lightweight security protocols are still needed to defend vehicular networks from different intrusion, eavesdropping, and saturation attacks.		

V. Proposed Security Measures

To achieve good security and privacy requirements for vehicular applications while gaining from low computation and efficient implementation, we propose the use of powerful and innovative techniques. First, the cryptographic hashing algorithm is used for identity-based authentication, which achieves complex mathematical operations with faster computation. Secondly, homomorphic encryption is to be used to protect sensitive data communication between network entities for enhanced privacy and confidentiality. Finally, to build a Machine Learning based intrusion detection system using the transfer learning technique for multiple attack detection capability with efficient implementation.

A. Authentication Protocol

The proposed authentication protocol includes vehicles and their respective users, RSUs, and TA as illustrated in Fig. 10. First, each vehicle should be used by a legitimate user who is authenticated before the vehicular authentication in the network. The user authentication phase will protect user information to preserve privacy and avoid malicious traceability. Secondly, every vehicle and RSU is registered by the TA before being allowed in the vehicular network. In this registration phase, TA protects the privacy of vehicles and RSUs and provides secure authentication parameters for them. Third, vehicles are authenticated by RSUs to be admitted in their respective communication ranges, and vehicles authenticate each other to communicate among themselves. The authentication phase is based on mutual authentication where entities can identify each other and securely agree on a communication key. In this phase, each entity uses a pseudonym identity and other parameters received from TA during the registrations phase which avoid sharing the real identity and therefore preserve their privacy in the network. Finally, the authenticated or legitimate entities can securely participate in vehicular communications.

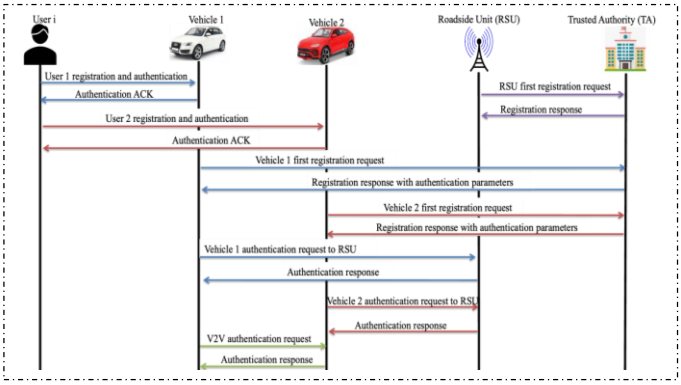


Fig. 10. Vehicular network authentication.

B. Privacy and Confidentiality

The proposed privacy and confidentiality mechanism for secure communication between vehicles, is based on homomorphic encryption as illustrated in Fig. 11. The homomorphic encryption offers the possibility to perform complex computations and data analysis on encrypted information without the need to decrypt them before [142], [143]. This represents a powerful solution for maintaining confidentiality and privacy during the transmission and processing of sensitive data. In the context of vehicular networks, each node can be a potential malicious node trying to collect private information. The homomorphic encryption can prevent unauthorized access to valuable information and therefore reduce the risk of data manipulation and breaches. Also, it enables traffic data analysis by the traffic authority and transportation companies without compromising the privacy of vehicular users and passengers.

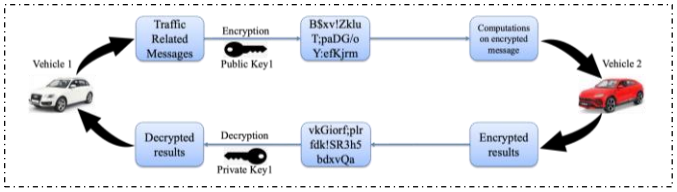


Fig. 11. Vehicular homomorphic encryption.

C. Intrusion Detection System

The proposed intrusion detection system for monitoring vehicular communications is based on transfer learning techniques as illustrated in Fig. 12. The transfer learning makes it possible to train a machine learning model using different datasets while gaining knowledge from all of them [144]. In this process, the model is trained with a starting dataset then the pre-trained model is trained again with a new dataset. Instead of implementing different models in the same device to detect individual types of attacks, transfer learning allows the accumulation of knowledge in a single model to save time and resources which is therefore suitable for vehicular applications.

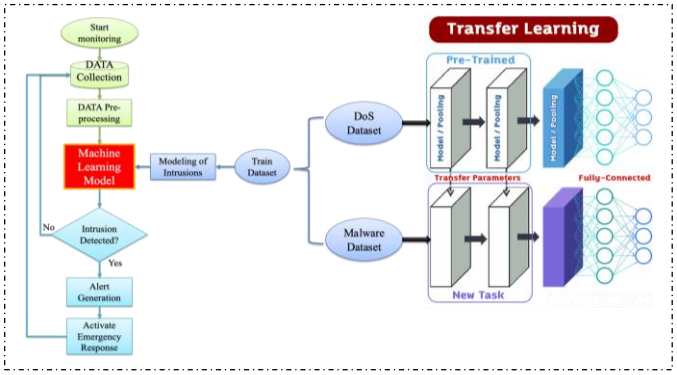


Fig. 12. Intrusion detection system based on transfer learning.

VI. Conclusion

Future transportation is expected to improve the quality of living by providing more safer and reliable mobility. While the introduction of autonomous vehicles has been presented to achieve this goal, it is also opening a new space for cyberattacks. Therefore, the cybersecurity concerns in the transportation area have raised interest from researchers and security experts to investigate and propose security measures for a trusted deployment. This paper reviewed the state of the art of cybersecurity issues defence strategies for AVs based on existing experiments and discussed methods in the literature. The review is organized by grouping attack methods and proposed defence techniques according to the target AV components. Based on this review, three major attack scenarios against AVs have been identified: 1) the attacker can target a component to interrupt its operations; 2) the attacker can target a component to have control over its operations without interrupting it; 3) the attacker can observe exchanged information without interrupting or controlling a component's operations. In response to the attacks, different defence approaches were proposed, which can also be categorised into three aspects including authentication, data protection, and intrusion detection. The authentication consists of identity verification and communication establishment to ensure that only trusted and legitimate entities are interacting. Data protection ensures that data transmitted between legitimate entities are trusted and secured from third parties. And, intrusion detection focuses on monitoring the interaction environment of legitimate entities to detect suspicious activities. The existing defence strategies were discussed to highlight their benefits in securing autonomous vehicles and also to show their limitations in satisfying critical requirements of vehicular networks, such as real-time and resource constraint applications, which can motivate further investigations. Furthermore, this paper presents some research directions that can be used to develop robust, efficient, and lightweight security measures, and therefore, contribute to building a trustworthy autonomous transportation ecosystem.

Acknowledgment

This work has been supported by the "Partnership for Skills in Applied Sciences, Engineering and Technology - Regional Scholarship and Innovation Fund (PASET-RSIF) through the African Centre of Excellence in Internet of Things (ACEIoT)".

References

[1] P. Suresh, J. V. Daniel, and R. H. Aswathy, "A state of the art review on the Internet of Things (IoT) History , Technology and fields of deployment," in *International Conference on Science, Engineering and Management Research (ICSEMR 2014)*, IEEE, 2014.

[2] P. Liu, "Internet of Thing Based Vehicular Network System and Application," in *Advances in Intelligent Systems Research*, 2018, pp. 298–302.

[3] F. Zhu, Y. Lv, Y. Chen, X. Wang, G. Xiong, and F. Y. Wang, "Parallel Transportation Systems: Toward IoT-Enabled Smart Urban Traffic Control and Management," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 10, pp. 4063–4071, 2020, doi: 10.1109/TITS.2019.2934991.

[4] SAE International, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor J3016_202104." Accessed: Dec. 14, 2023. [Online]. Available: https://www.sae.org/standards/content/j3016_202104/

[5] J. Wang, J. Liu, and N. Kato, "Networking and Communications in Autonomous Driving: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1243–1274, Apr. 2019, doi: 10.1109/COMST.2018.2888904.

[6] S. Muthuramalingam, A. Bharathi, S. Rakesh kumar, N. Gayathri, R. Sathiyaraj, and B. Balamurugan, "IoT Based Intelligent Transportation System (iot-its) for Global Perspective: A Case Study," *Internet of Things and Big Data Analytics for Smart Generation*. Springer Nature Switzerland AG 2019, pp. 279–300, 2019. doi: 10.1007/978-3-030-04203-5_13.

[7] A. O. Al Zaabi, C. Y. Yeun, and E. Damiani, "Autonomous Vehicle Security: Conceptual Model," in *2019 IEEE Transportation Electrification Conference and Expo, Asia-Pacific (ITEC Asia-Pacific)*, IEEE, May 2019, pp. 1–5. doi: 10.1109/ITEC-AP.2019.8903691.

[8] B. K. Ren, Q. Wang, C. Wang, Z. Qin, and X. Lin, "The Security of Autonomous Driving : Threats , Defenses , and Future Directions," *Proceedings of the IEEE*, pp. 1–16, 2019, doi: 10.1109/JPROC.2019.2948775.

[9] K. Kim, J. S. Kim, S. Jeong, J.-H. Park, and H. K. Kim, "Cybersecurity for autonomous vehicles: Review of attacks and defense," *Comput Secur*, vol. 103, p. 102150, Apr. 2021, doi: 10.1016/j.cose.2020.102150.

[10] X. Sun, F. R. Yu, and P. Zhang, "A Survey on Cyber-Security of Connected and Autonomous Vehicles (CAVs)," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 6240–6259, Jul. 2022, doi: 10.1109/TITS.2021.3085297.

[11] Y. Takefuji, "Connected Vehicle Security Vulnerabilities [Commentary]," *IEEE Technology and Society Magazine*, vol. 37, no. 1, pp. 15–18, Mar. 2018, doi: 10.1109/MTS.2018.2795093.

[12] K. Rudd, "Security of Autonomous Systems Employing Embedded Computing and Sensors," pp. 80–86, 2013.

[13] M. Kalmeshwar and K. S. Nandini Prasad, "Development of On-Board Diagnostics for Car and it's Integration with Android Mobile," in *2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*, IEEE, Dec. 2017, pp. 1–6. doi: 10.1109/CSITSS.2017.8447540.

[14] A. Rizwan *et al.*, "Simulation of IoT-based Vehicular Ad Hoc Networks (VANETs) for Smart Traffic Management Systems," *Wirel Commun Mob Comput*, vol. 2022, pp. 1–11, May 2022, doi: 10.1155/2022/3378558.

[15] M. M. A. Muslam, "Enhancing Security in Vehicle-to-Vehicle Communication: A Comprehensive Review of Protocols and Techniques," *Vehicles*, vol. 6, no. 1, pp. 450–467, Feb. 2024, doi: 10.3390/vehicles6010020.

[16] S. Adnan Yusuf, A. Khan, and R. Souissi, "Vehicle-to-everything (V2X) in the autonomous vehicles domain – A technical review of communication, sensor, and AI technologies for road user safety," *Transp Res Interdiscip Perspect*, vol. 23, p. 100980, Jan. 2024, doi: 10.1016/j.trip.2023.100980.

[17] A. Chattopadhyay and K. Lam, "Security of Autonomous Vehicle as a Cyber-Physical System," 2017.

[18] M. N. Ahangar, Q. Z. Ahmed, F. A. Khan, and M. Hafeez, "A Survey of Autonomous Vehicles: Enabling Communication Technologies and Challenges," *Sensors*, vol. 21, no. 3, p. 706, Jan. 2021, doi: 10.3390/s21030706.

[19] K. Abboud, H. A. Omar, and W. Zhuang, "Interworking of DSRC and Cellular Network Technologies for V2X Communications: A Survey," *IEEE Trans Veh Technol*, vol. 65, no. 12, pp. 9457–9470, Dec. 2016, doi: 10.1109/TVT.2016.2591558.

[20] Z. Wang, H. Wei, J. Wang, X. Zeng, and Y. Chang, "Security Issues and Solutions for Connected and Autonomous Vehicles in a Sustainable City: A Survey," *Sustainability*, vol. 14, no. 19, p. 12409, Sep. 2022, doi: 10.3390/su141912409.

[21] M. Pham and K. Xiong, "A survey on security attacks and defense techniques for connected and autonomous vehicles," *Comput Secur*, vol. 109, p. 102269, Oct. 2021, doi: 10.1016/j.cose.2021.102269.

[22] B. R. Mudhivarthi, P. Thakur, and G. Singh, "Aspects of Cyber Security in Autonomous and Connected Vehicles," *Applied Sciences*, vol. 13, no. 5, p. 3014, Feb. 2023, doi: 10.3390/app13053014.

[23] B. Sheehan, F. Murphy, M. Mullins, and C. Ryan, "Connected and autonomous vehicles: A cyber-risk classification framework," *Transp Res Part A Policy Pract*, vol. 124, pp. 523–536, Jun. 2019, doi: 10.1016/j.tra.2018.06.033.

[24] A. Singandhupe and H. M. La, "A Review of SLAM Techniques and Security in Autonomous Driving," in *2019 Third IEEE International Conference on Robotic Computing (IRC)*, IEEE, Feb. 2019, pp. 602–607. doi: 10.1109/IRC.2019.00122.

[25] A. M. Wyglinski, X. Huang, T. Padir, L. Lai, T. R. Eisenbarth, and K. Venkatasubramanian, "Security of Autonomous Systems Employing Embedded Computing and Sensors," *IEEE Micro*, vol. 33, no. 1, pp. 80–86, Jan. 2013, doi: 10.1109/MM.2013.18.

[26] J. Cui and B. Zhang, "VeRA: A Simplified Security Risk Analysis Method for Autonomous Vehicles," *IEEE Trans Veh Technol*, vol. 69, no. 10, pp. 10494–10505, Oct. 2020, doi: 10.1109/TVT.2020.3009165.

[27] S. Parkinson, P. Ward, K. Wilson, and J. Miller, "Cyber Threats Facing Autonomous and Connected Vehicles: Future Challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 11, pp. 2898–2915, Nov. 2017, doi: 10.1109/TITS.2017.2665968.

[28] A. Nanda, D. Puthal, J. J. P. C. Rodrigues, and S. A. Kozlov, "Internet of Autonomous Vehicles Communications Security: Overview, Issues, and Directions," *IEEE Wirel Commun*, vol. 26, no. 4, pp. 60–65, Aug. 2019, doi: 10.1109/MWC.2019.1800503.

[29] Q. Xiao, X. Pan, Y. Lu, M. Zhang, J. Dai, and M. Yang, "Exorcising "'Wraith"': Protecting LiDAR-based Object Detector in Automated Driving System from Appearing Attacks," *Proceedings of the 32nd USENIX Conference on Security Symposium*, pp. 2939–2956, Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.09731

[30] K. Ren, Q. Wang, C. Wang, Z. Qin, and X. Lin, "The Security of Autonomous Driving: Threats, Defenses, and Future Directions," *Proceedings of the IEEE*, vol. 108, no. 2, pp. 357–372, Feb. 2020, doi: 10.1109/JPROC.2019.2948775.

[31] H. M. Furqan, M. S. J. Solaija, H. Turkmen, and H. Arslan, "Wireless Communication, Sensing, and REM: A Security Perspective," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 287–321, 2021, doi: 10.1109/OJCOMS.2021.3054066.

[32] S. Tout, M. Abualkibash, and P. Patil, "Emerging Research in the Security of Modern and Autonomous Vehicles," in *2018 IEEE International Conference on Electro/Information Technology (EIT)*, IEEE, May 2018, pp. 0543–0547. doi: 10.1109/EIT.2018.8500204.

[33] A. Ferdowsi, U. Challita, W. Saad, and N. B. Mandayam, "Robust Deep Reinforcement Learning for Security and Safety in Autonomous Vehicle Systems," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, Nov. 2018, pp. 307–312. doi: 10.1109/ITSC.2018.8569635.

[34] L. Zhang, X. Yan, and D. Ma, "A Binarized Neural Network Approach to Accelerate in-Vehicle Network Intrusion Detection," *IEEE Access*, vol. 10, pp. 123505–123520, 2022, doi: 10.1109/ACCESS.2022.3208091.

[35] D. Uhlir, P. Sedlacek, and J. Hosek, "Practial overview of commercial connected cars systems in Europe," in *2017 9th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, Munich: IEEE, Nov. 2017, pp. 436–444. doi: 10.1109/ICUMT.2017.8255178.

[36] S. Zamfir and R. Drosescu, "Automotive Black Box and Development Platform Used for Traffic Risks Evaluation and Mitigation," in *The 30th SIAR International Congress of Automotive and Transport Engineering*, Cham: Springer International Publishing, 2020, pp. 426–438. doi: 10.1007/978-3-030-32564-0_50.

[37] J. Kang, D. Lin, E. Bertino, and O. Tonguz, "From Autonomous Vehicles to Vehicular Clouds: Challenges of Management, Security and Dependability," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, IEEE, Jul. 2019, pp. 1730–1741. doi: 10.1109/ICDCS.2019.00172.

[38] S. McCall, C. Yucel, and V. Katos, "Education in Cyber Physical Systems Security: The Case of Connected Autonomous Vehicles," in *2021 IEEE Global Engineering Education Conference (EDUCON)*, IEEE, Apr. 2021, pp. 1379–1385. doi: 10.1109/EDUCON46332.2021.9454086.

[39] A. A. Mehta *et al.*, "Securing the Future: A Comprehensive Review of Security Challenges and Solutions in Advanced Driver Assistance Systems," *IEEE Access*, vol. 12, pp. 643–678, 2024, doi: 10.1109/ACCESS.2023.3347200.

[40] Upstream Security Ltd, "H1'2023: AUTOMOTIVE CYBER TREND REPORT," 2023. Accessed: Dec. 08, 2023. [Online]. Available: https://upstream.auto/reports/h1-2023-automotive-cyber-trend-report/

[41] Upstream Security Ltd, "GLOBAL AUTOMOTIVE CYBERSECURITY REPORT: AUTOMOTIVE CYBER THREAT LANDSCAPE IN LIGHT OF NEW REGULATIONS," 2022. Accessed: Dec. 08, 2023. [Online]. Available: https://upstream.auto/2022report/

[42] B. G. B. Stottelaar, "PRACTICAL CYBER-ATTACKS ON AUTONOMOUS VEHICLES," University of Twente, 2015.

[43] J. Petit, B. Stottelaar, M. Feiri, and F. Kargl, "Remote Attacks on Automated Vehicles Sensors: Experiments on Camera and LiDAR," in *Black Hat Europe*, 2015, p. 995. Accessed: May 12, 2024. [Online]. Available: blackhat.com

[44] H. Shin, D. Kim, Y. Kwon, and Y. Kim, "Illusion and Dazzle: Adversarial Optical Channel Exploits Against Lidars for Automotive Applications," in *Cryptographic Hardware and Embedded Systems – CHES 2017*, W. Fischer and N. Homma, Eds., in Lecture Notes in Computer Science. , Cham: Springer International Publishing, 2017, pp. 445–467. doi: 10.1007/978-3-319-66787-4_22.

[45] Q. Wang *et al.*, "Pseudorandom modulation quantum secured lidar," *Optik (Stuttg)*, vol. 126, no. 22, pp. 3344–3348, Nov. 2015, doi: 10.1016/j.ijleo.2015.07.048.

[46] D. Davidson, H. Wu, and R. Jellinek, "Controlling UAVs with Sensor Input Spoofing Attacks," in *10th USENIX Workshop on Offensive Technologies (WOOT '16)*, Austin: USENIX, Aug. 2016, pp. 1–11.

[47] R. Matsumura, T. Sugawara, and K. Sakiyama, "A Secure LiDAR with AES-Based Side-Channel Fingerprinting," in *2018 Sixth International Symposium on Computing and Networking Workshops (CANDARW)*, Takayama: IEEE, Nov. 2018, pp. 479–482. doi: 10.1109/CANDARW.2018.00092.

[48] Y. Cao *et al.*, "Adversarial Sensor Attack on LiDAR-based Perception in Autonomous Driving," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, New York, NY, USA: ACM, Nov. 2019, pp. 2267–2281. doi: 10.1145/3319535.3339815.

[49] K. M. A. Alheeti, A. Alzahrani, and D. Al Dosary, "LiDAR Spoofing Attack Detection in Autonomous Vehicles," in *2022 IEEE International Conference on Consumer Electronics (ICCE)*, IEEE, Jan. 2022, pp. 1–2. doi: 10.1109/ICCE53296.2022.9730540.

[50] H. Zhang, Z. Li, S. Cheng, and A. Clark, "Cooperative Perception for Safe Control of Autonomous Vehicles under LiDAR Spoofing Attacks," in *Proceedings Inaugural International Symposium on Vehicle Security & Privacy*, Reston, VA: Internet Society, 2023. doi: 10.14722/vehiclesec.2023.23066.

[51] R. Chauhan, "A Platform for False Data Injection in Frequency Modulated Continuous Wave Radar," Utah State University, 2014. [Online]. Available: https://digitalcommons.usu.edu/etd/3964

[52] R. Komissarov and A. Wool, "Spoofing Attacks Against Vehicular FMCW Radar," in *Proceedings of the 5th Workshop on Attacks and Solutions in Hardware Security*, New York, NY, USA: ACM, Nov. 2021, pp. 91–97. doi: 10.1145/3474376.3487283.

[53] A. Lazaro, A. Porcel, M. Lazaro, R. Villarino, and D. Girbau, "Spoofing Attacks on FMCW Radars with Low-Cost Backscatter Tags," *Sensors*, vol. 22, no. 6, p. 2145, Mar. 2022, doi: 10.3390/s22062145.

[54] Walter E. Buehler, Roger M. Whitson, and Michael J. Lewis, "AIRBORNE RADAR JAMMING SYSTEM," US00883 0112B1, Sep. 09, 2014

[55] Y. Shoukry, P. Martin, Y. Yona, S. Diggavi, and M. Srivastava, "PyCRA: Physical Challenge-Response Authentication For Active Sensors Under Spoofing Attacks," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, New York, NY, USA: ACM, Oct. 2015, pp. 1004–1015. doi: 10.1145/2810103.2813679.

[56] P. Kapoor, A. Vora, and K.-D. Kang, "Detecting and Mitigating Spoofing Attack Against an Automotive Radar," in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, Chicago: IEEE, Aug. 2018, pp. 1–6. doi: 10.1109/VTCFall.2018.8690734.

[57] S. Cohen, E. Levy, A. Shaked, T. Cohen, Y. Elovici, and A. Shabtai, "RadArnomaly: Protecting Radar Systems from Data Manipulation Attacks," *Sensors*, vol. 22, no. 11, p. 4259, Jun. 2022, doi: 10.3390/s22114259.

[58] N. O. Tippenhauer, C. Pöpper, K. B. Rasmussen, and S. Capkun, "On the requirements for successful GPS spoofing attacks," in *Proceedings of the 18th ACM conference on Computer and communications security*, New York, NY, USA: ACM, Oct. 2011, pp. 75–86. doi: 10.1145/2046707.2046719.

[59] A. Helfrick, "Question: Alternate position, navigation timing, APNT? Answer: ELORAN," in *2014 IEEE/AIAA 33rd Digital Avionics Systems Conference (DASC)*, Colorado: IEEE, Oct. 2014, pp. 1–9. doi: 10.1109/DASC.2014.6979452.

[60] K. Zeng *et al.*, "All Your GPS Are Belong To Us: Towards Stealthy Manipulation of Road Navigation Systems," in *Proceedings of the 27th USENIX Security Symposium*, Baltimore: USENIX, Aug. 2018, pp. 1527–1544. Accessed: Jan. 07, 2025. [Online]. Available: https://www.usenix.org/conference/usenixsecurity18/presentation/zeng

[61] S. Narain, A. Ranganathan, and G. Noubir, "Security of GPS/INS Based On-road Location Tracking Systems," in *2019 IEEE Symposium on Security and Privacy (SP)*, San Francisco: IEEE, May 2019, pp. 587–601. doi: 10.1109/SP.2019.00068.

[62] Q. Meng, L.-T. Hsu, B. Xu, X. Luo, and A. El-Mowafy, "A GPS Spoofing Generator Using an Open Sourced Vector Tracking-Based Receiver," *Sensors*, vol. 19, no. 18, p. 3993, Sep. 2019, doi: 10.3390/s19183993.

[63] Paul Y. Montgomery, Todd E. Humphreys, and Brent M. Ledvina, "Receiver-Autonomous Spoofing Detection: Experimental Results of a Multi-Antenna Receiver Defense against a Portable Civil GPS Spoofer," in *Proceedings of the Institute of Navigation, National Technical Meeting*, Anaheim: Institute of Navigation, Jan. 2010, pp. 124–130. doi: 10.15781/T2GB1Z038.

[64] A. Purwar, D. Joshi, and V. K. Chaubey, "GPS signal jamming and anti-jamming strategy — A theoretical analysis," in *2016 IEEE Annual India Conference (INDICON)*, Bangalore: IEEE, Dec. 2016, pp. 1–6. doi: 10.1109/INDICON.2016.7838933.

[65] B. W. O'Hanlon, M. L. Psiaki, J. A. Bhatti, D. P. Shepard, and T. E. Humphreys, "Real-Time GPS Spoofing Detection via Correlation of Encrypted Signals," *Navigation*, vol. 60, no. 4, pp. 267–278, Dec. 2013, doi: 10.1002/navi.44.

[66] Y. Yang and J. Xu, "GNSS receiver autonomous integrity monitoring (RAIM) algorithm based on robust estimation," *Geod Geodyn*, vol. 7, no. 2, pp. 117–123, Mar. 2016, doi: 10.1016/j.geog.2016.04.004.

[67] Q. MENG, J. LIU, Q. ZENG, S. FENG, and R. XU, "Impact of one satellite outage on ARAIM depleted constellation configurations," *Chinese Journal of Aeronautics*, vol. 32, no. 4, pp. 967–977, Apr. 2019, doi: 10.1016/j.cja.2019.01.004.

[68] Q. Meng, J. Liu, Q. Zeng, S. Feng, and R. Xu, "Improved ARAIM fault modes determination scheme based on feedback structure with probability accumulation," *GPS Solutions*, vol. 23, no. 1, p. 16, Jan. 2019, doi: 10.1007/s10291-018-0809-8.

[69] M. Foruhandeh, A. Z. Mohammed, G. Kildow, P. Berges, and R. Gerdes, "Spotr: GPS spoofing detection via device fingerprinting," in *Proceedings of the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, New York, NY, USA: ACM, Jul. 2020, pp. 242–253. doi: 10.1145/3395351.3399353.

[70] S. Liu *et al.*, "Stars can tell: A robust method to defend against GPS spoofing attacks using off-the-shelf chipset," in *Proceedings of the 30th USENIX Security Symposium*, 2021.

[71] M. Ahmad and Y. Wang, "A Low-Cost Approach to Securing Commercial GPS Receivers Against Spoofing Attacks," in *Lecture Notes in Control and Information Sciences*, vol. 489, 2022, pp. 149–175. doi: 10.1007/978-3-030-83236-0_6.

[72] F. Wang, Y. Hong, and X. Ban, "Infrastructure-Enabled GPS Spoofing Detection and Correction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 12, pp. 13878–13892, Dec. 2023, doi: 10.1109/TITS.2023.3298785.

[73] M. Shabbir, M. Kamal, Z. Ullah, and M. M. Khan, "Securing Autonomous Vehicles Against GPS Spoofing Attacks: A Deep Learning Approach," *IEEE Access*, vol. 11, pp. 105513–105526, 2023, doi: 10.1109/ACCESS.2023.3319514.

[74] K. S. Jasim, K. M. Ali Alheeti, and A. K. A. Najem Alaloosy, "Intelligent Detection System for Spoofing and Jamming Attacks in UAVs," 2023, pp. 97–110. doi: 10.1007/978-3-031-21101-0_8.

[75] C. Yan, W. Xu, and J. Liu, "Can You Trust Autonomous Vehicles : Contactless Attacks against Sensors of Self-driving Vehicle," *ACM SIGARCH Computer Architecture News*, pp. 1–13, 2016, doi: 10.1145/1235.

[76] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," *ICLR 2015*, pp. 1–11, Dec. 2014, [Online]. Available: http://arxiv.org/abs/1412.6572

[77] J. Kos, I. Fischer, and D. Song, "Adversarial Examples for Generative Models," in *2018 IEEE Security and Privacy Workshops (SPW)*, IEEE, May 2018, pp. 36–42. doi: 10.1109/SPW.2018.00014.

[78] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal Adversarial Perturbations," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu: IEEE, Jul. 2017, pp. 86–94. doi: 10.1109/CVPR.2017.17.

[79] C. Sitawarin, A. N. Bhagoji, A. Mosenia, M. Chiang, and P. Mittal, "DARTS: Deceiving Autonomous Cars with Toxic Signs," *ACM CCS 2018*, pp. 1–18, Feb. 2018, [Online]. Available: http://arxiv.org/abs/1802.06430

[80] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial Patch," *31st Conference on Neural Information Processing Systems (NIPS 2017)*, pp. 1–6, Dec. 2017, [Online]. Available: http://arxiv.org/abs/1712.09665

[81] K. Eykholt *et al.*, "Robust Physical-World Attacks on Deep Learning Visual Classification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 1625–1634. doi: 10.1109/CVPR.2018.00175.

[82] K. B. Kelarestaghi, M. Foruhandeh, K. Heaslip, and R. Gerdes, "Intelligent Transportation System Security: Impact-Oriented Risk Assessment of in-Vehicle Networks," *IEEE Intelligent Transportation Systems Magazine*, vol. 13, no. 2, pp. 91–104, Jun. 2021, doi: 10.1109/MITS.2018.2889714.

[83] W. Wang, Y. Yao, X. Liu, X. Li, P. Hao, and T. Zhu, "I Can See the Light: Attacks on Autonomous Vehicles Using Invisible Lights," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, New York, NY, USA: ACM, Nov. 2021, pp. 1930–1944. doi: 10.1145/3460120.3484766.

[84] [84] Sharath Yadav and A. Ansari, "Autonomous Vehicles Camera Blinding Attack Detection Using Sequence Modelling and Predictive Analytics," in *SAE Technical Paper 2020-01-0719*, Apr. 2020. doi: 10.4271/2020-01-0719.

[85] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering Adversarial Images using Input Transformations," *International Conference on Learning Representations*, pp. 1–12, Oct. 2017, [Online]. Available: http://arxiv.org/abs/1711.00117

[86] V. Zantedeschi, M.-I. Nicolae, and A. Rawat, "Efficient Defenses Against Adversarial Attacks," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, New York, NY, USA: ACM, Nov. 2017, pp. 39–49. doi: 10.1145/3128572.3140449.

[87] G. K. Dziugaite, Z. Ghahramani, and D. M. Roy, "A study of the effect of JPG compression on adversarial images," *International Society for Bayesian Analysis (ISBA 2016) World Meeting*, pp. 1–8, Aug. 2016, [Online]. Available: http://arxiv.org/abs/1608.00853

[88] W. Xu, D. Evans, and Y. Qi, "Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks," in *Proceedings 2018 Network and Distributed System Security Symposium*, Reston, VA: Internet Society, Feb. 2018, pp. 1–15. doi: 10.14722/ndss.2018.23198.

[89] W. Jiang, H. Li, S. Liu, X. Luo, and R. Lu, "Poisoning and Evasion Attacks Against Deep Learning Algorithms in Autonomous Vehicles," *IEEE Trans Veh Technol*, vol. 69, no. 4, pp. 4439–4449, Apr. 2020, doi: 10.1109/TVT.2020.2977378.

[90] C. Szegedy *et al.*, "Intriguing properties of neural networks," *ArXiv*, pp. 1–10, Dec. 2013, [Online]. Available: http://arxiv.org/abs/1312.6199

[91] T. Miyato, S. Maeda, M. Koyama, K. Nakae, and S. Ishii, "Distributional Smoothing with Virtual Adversarial Training," *International Conference on Learning Representations*, pp. 1–12, Jul. 2015, [Online]. Available: http://arxiv.org/abs/1507.00677

[92] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow, "THERMOMETER ENCODING: ONE HOT WAY TO RESIST ADVERSARIAL EXAMPLES," in *International Conference on Learning Representations*, 2018, pp. 1–22.

[93] Q. He, X. Meng, and R. Qu, "Towards a Severity Assessment Method for Potential Cyber Attacks to Connected and Autonomous Vehicles," *J Adv Transp*, vol. 2020, pp. 1–15, Sep. 2020, doi: 10.1155/2020/6873273.

[94] V. L. L. Thing and J. Wu, "Autonomous Vehicle Security: A Taxonomy of Attacks and Defences," in 2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), IEEE, Dec. 2016, pp. 164–170. doi: 10.1109/iThings-GreenCom-CPSCom-SmartData.2016.52.

[95] C.-W. Lin and A. Sangiovanni-Vincentelli, "Cyber-Security for the Controller Area Network (CAN) Communication Protocol," in *2012 International Conference on Cyber Security*, Washington : IEEE, Dec. 2012, pp. 1–7. doi: 10.1109/CyberSecurity.2012.7.

[96] K. Koscher *et al.*, "Experimental Security Analysis of a Modern Automobile," *2010 IEEE Symposium on Security and Privacy Experimental*, vol. 34, pp. 447–462, 2010, doi: 10.1109/SP.2010.34.

[97] S. Woo, H. J. Jo, and D. H. Lee, "A Practical Wireless Attack on the Connected Car and Security Protocol for In-Vehicle CAN," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 1–14, Apr. 2014, doi: 10.1109/TITS.2014.2351612.

[98] S. Checkoway *et al.*, "Comprehensive Experimental Analyses of Automotive Attack Surfaces," in *20th USENIX security symposium*, 2011, pp. 447–462.

[99] A. Boudguiga, J. Letailleur, R. Sirdey, and W. Klaudel, "Enhancing CAN Security by Means of Lightweight Stream-Ciphers and Protocols," in *SAFECOMP 2019 Workshops, LNCS 11699*, 2019, pp. 235–250. doi: 10.1007/978-3-030-26250-1_19.

[100] S. Nie, L. Liu, and Y. Du, "FREE-FALL : HACKING TESLA FROM WIRELESS TO CAN BUS," *Keen Security Lab of Tencent*, pp. 1–16, 2017.

[101] T. Keen Security Lab, "Experimental Security Research of Tesla Autopilot," Mar. 2019. Accessed: Jan. 28, 2024. [Online]. Available: https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf

[102] D. S. Fowler, J. Bryans, M. Cheah, and P. Wooderson, "A Method for Constructing Automotive Cybersecurity Tests , a CAN Fuzz Testing Example," in *2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, IEEE, 2019, pp. 1–8. doi: 10.1109/QRS-C.2019.00015.

[103] M. Muter and N. Asaj, "Entropy-based anomaly detection for in-vehicle networks," in *2011 IEEE Intelligent Vehicles Symposium (IV)*, Baden-Baden: IEEE, Jun. 2011, pp. 1110–1115. doi: 10.1109/IVS.2011.5940552.

[104] T. Matsumoto, M. Hata, M. Tanabe, K. Yoshioka, and K. Oishi, "A Method of Preventing Unauthorized Data Transmission in Controller Area Network," in *2012 IEEE 75th Vehicular Technology Conference (VTC Spring)*, Yokohama: IEEE, May 2012, pp. 1–5. doi: 10.1109/VETECS.2012.6240294.

[105] M. Gmiden, M. H. Gmiden, and H. Trabelsi, "An intrusion detection method for securing in-vehicle CAN bus," in *2016 17th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA)*, Sousse,: IEEE, Dec. 2016, pp. 176–180. doi: 10.1109/STA.2016.7952095.

[106] Kyong-Tak Cho and Kang G. Shin, "Fingerprinting Electronic Control Units for Vehicle Intrusion Detection," in *25th USENIX Security Symposium*, Austin: USENIX Association, Aug. 2016, pp. 910–927.

[107] C. Valasek and Charlie Miller, "A Survey of Remote Automotive Attack Surfaces," Jul. 2014.

[108] A. S. Siddiqui, Y. Gui, J. Plusquellic, and F. Saqib, "Secure communication over CANBus," in *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Boston: IEEE, Aug. 2017, pp. 1264–1267. doi: 10.1109/MWSCAS.2017.8053160.

[109] Z. Tyree, R. A. Bridges, F. L. Combs, and M. R. Moore, "Exploiting the Shape of CAN Data for In-Vehicle Intrusion Detection," in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, Chicago: IEEE, Aug. 2018, pp. 1–5. doi: 10.1109/VTCFall.2018.8690644.

[110] D. Basavaraj and S. Tayeb, "Towards a Lightweight Intrusion Detection Framework for In-Vehicle Networks," *Journal of Sensor and Actuator Networks*, vol. 11, no. 1, pp. 1–20, 2022, doi: 10.3390/jsan11010006.

[111] S. Adly, A. Moro, S. Hammad, and S. A. Maged, "Prevention of Controller Area Network (CAN) Attacks on Electric Autonomous Vehicles," *Applied Sciences (Switzerland)*, vol. 13, no. 16, pp. 1–23, 2023, doi: 10.3390/app13169374.

[112] F. W. Alsaade and M. H. Al-Adhaileh, "Cyber Attack Detection for Self-Driving Vehicle Networks Using Deep Autoencoder Algorithms," *Sensors*, vol. 23, no. 8, pp. 1–26, 2023, doi: 10.3390/s23084086.

[113] M. Amoozadeh *et al.*, "Security vulnerabilities of connected vehicle streams and their impact on cooperative driving," *IEEE Communications Magazine*, vol. 53, no. 6, pp. 126–132, Jun. 2015, doi: 10.1109/MCOM.2015.7120028.

[114] N. Ekedebe, W. Yu, H. Song, and C. Lu, "On a simulation study of cyber attacks on vehicle-to-infrastructure communication (V2I) in

Intelligent Transportation System (ITS)," in *Mobile Multimedia/Image Processing, Security, and Applications 2015*, S. S. Agaian, S. A. Jassim, and E. Y. Du, Eds., Baltimore,: SPIE, May 2015, p. 94970B. doi: 10.1117/12.2177465.

[115]J. Grover, M. S. Gaur, V. Laxmi, and N. K. Prajapati, "A sybil attack detection approach using neighboring vehicles in VANET," in *Proceedings of the 4th international conference on Security of information and networks*, New York, NY, USA: ACM, Nov. 2011, pp. 151–158. doi: 10.1145/2070425.2070450.

[116]W. Whyte, A. Weimerskirch, V. Kumar, and T. Hehn, "A security credential management system for V2V communications," in *2013 IEEE Vehicular Networking Conference*, Boston: IEEE, Dec. 2013, pp. 1–8. doi: 10.1109/VNC.2013.6737583.

[117]M. Alimohammadi and A. A. Pouyan, "Sybil attack detection using a low cost short group signature in VANET," in *2015 12th International Iranian Society of Cryptology Conference on Information Security and Cryptology (ISCISC)*, Rasht: IEEE, Sep. 2015, pp. 23–28. doi: 10.1109/ISCISC.2015.7387893.

[118]J. Li, H. Lu, and M. Guizani, "ACPN: A Novel Authentication Framework with Conditional Privacy-Preservation and Non-Repudiation for VANETs," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 4, pp. 938–948, Apr. 2015, doi: 10.1109/TPDS.2014.2308215.

[119]S. Taha, M. Alhassany, and X. Shen, "Lightweight Handover Authentication Scheme for 5G-Based V2X Communications," in *2018 IEEE Global Communications Conference (GLOBECOM)*, IEEE, Dec. 2018, pp. 1–6. doi: 10.1109/GLOCOM.2018.8648020.

[120]C. Xu, X. Huang, M. Ma, and H. Bao, "A Secure and Efficient Message Authentication Scheme for Vehicular Networks based on LTE-V," *KSII Transactions on Internet and Information Systems*, vol. 12, no. 6, Jun. 2018, doi: 10.3837/tiis.2018.06.022.

[121]S. Tangade, S. S. Manvi, and P. Lorenz, "Trust Management Scheme Based on Hybrid Cryptography for Secure Communications in VANETs," *IEEE Trans Veh Technol*, vol. 9545, pp. 1–12, 2020, doi: 10.1109/TVT.2020.2981127.

[122]H. Jiang, L. Hua, and L. Wahab, "SAES: A self-checking authentication scheme with higher efficiency and security for VANET," *Peer Peer Netw Appl*, vol. 14, no. 2, pp. 528–540, Mar. 2021, doi: 10.1007/s12083-020-00997-0.

[123]J. Miao, Z. Wang, X. Ning, N. Xiao, W. Cai, and R. Liu, "Practical and secure multifactor authentication protocol for autonomous vehicles in 5G," *John Wiley & Sons, Ltd*, pp. 1–18, 2022, doi: 10.1002/spe.3087.

[124]R. Ma, J. Cao, D. Feng, H. Li, X. Li, and Y. Xu, "A robust authentication scheme for remote diagnosis and maintenance in 5G V2N," *Journal of Network and Computer Applications*, vol. 198, p. 103281, Feb. 2022, doi: 10.1016/j.jnca.2021.103281.

[125]H. P. Hyunhee Park, "Edge Based Lightweight Authentication Architecture Using Deep Learning for Vehicular Networks," *Journal of Internet Technology*, vol. 23, no. 1, pp. 195–202, Jan. 2022, doi: 10.53106/160792642022012301020.

[126]Q. Li, "A V2V Identity Authentication and Key Agreement Scheme Based on Identity-Based Cryptograph," *Future Internet*, vol. 15, no. 1, p. 25, Jan. 2023, doi: 10.3390/fi15010025.

[127]J. Noh, Y. Kwon, J. Son, and S. Cho, "Blockchain-Based One-Time Authentication for Secure V2X Communication Against Insiders and Authority Compromise Attacks," *IEEE Internet Things J*, vol. 10, no. 7, pp. 6235–6248, Apr. 2023, doi: 10.1109/JIOT.2022.3224465.

[128]X. Feng, X. Wang, H. Liu, H. Yang, and L. Wang, "A Privacy-Preserving Aggregation Scheme With Continuous Authentication for Federated Learning in VANETs," *IEEE Trans Veh Technol*, vol. 73, no. 7, pp. 9465–9477, Jul. 2024, doi: 10.1109/TVT.2024.3369942.

[129]L. Zhang, "OTIBAAGKA : A New Security Tool for Cryptographic Mix-Zone Establishment in Vehicular Ad Hoc Networks," *IEEE Transactions on Information Forensics and Security*, pp. 1–13, 2017, doi: 10.1109/TIFS.2017.2730479.

[130]A. Riahi Sfar, Y. Challal, P. Moyal, and E. Natalizio, "A Game Theoretic Approach for Privacy Preserving Model in IoT-Based Transportation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 12, pp. 4405–4414, 2019, doi: 10.1109/TITS.2018.2885054.

[131]S. A. A. Hakeem, M. A. A. El-gawad, and H. Kim, "Comparative Experiments of V2X Security Protocol Based on Hash Chain Cryptography," *MDPI Sensors*, vol. 5719, no. 20, pp. 2–23, 2020.

[132]C. Zhao, N. Guo, T. Gao, X. Deng, and J. Qi, "PEPA: Paillier cryptosystem-based efficient privacy-preserving authentication scheme for VANETs," *Journal of Systems Architecture*, vol. 138, pp. 1–10, 2023, doi: 10.1016/j.sysarc.2023.102855.

[133]U. Khan, S. Agrawal, and S. Silakari, "Detection of Malicious Nodes (DMN) in vehicular ad-hoc networks," *Procedia Comput Sci*, vol. 46, pp. 965–972, 2014, doi: 10.1016/j.procs.2015.01.006.

[134]H. Sedjelmaci and S. M. Senouci, "An accurate and efficient collaborative intrusion detection framework to secure vehicular networks," *Computers and Electrical Engineering*, vol. 43, pp. 33–47, 2015, doi: 10.1016/j.compeleceng.2015.02.018.

[135]M. J. Kang and J. W. Kang, "Intrusion Detection System using Deep Neural Network for In-Vehicle Network Security," *PLoS One*, vol. 11, no. 6, pp. 1–17, 2016, doi: 10.1371/journal.pone.0155781.

[136]S. M. Sangve, R. Bhati, and V. N.Gavali, "Intrusion Detection System for Detecting Rogue Nodes in Vehicular Ad-hoc Network," in *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, Pune, India: IEEE, 2017, pp. 127–131.

[137]A. Anwar, T. Halabi, and M. Zulkernine, "A coalitional security game against data integrity attacks in autonomous vehicle networks," *Vehicular Communications*, vol. 37, pp. 1–10, 2022, doi: 10.1016/j.vehcom.2022.100517.

[138]P. K. Singh, S. Kumar Jha, S. K. Nandi, and S. Nandi, "ML-Based Approach to Detect DDoS Attack in V2I Communication Under SDN Architecture," in *TENCON 2018 - 2018 IEEE Region 10 Conference*, Jeju: IEEE, Oct. 2018, pp. 0144–0149. doi: 10.1109/TENCON.2018.8650452.

[139]G. O. Anyanwu, C. I. Nwakanma, J.-M. Lee, and D.-S. Kim, "RBF-SVM kernel-based model for detecting DDoS attacks in SDN integrated vehicular network," *Ad Hoc Networks*, vol. 140, p. 103026, Mar. 2023, doi: 10.1016/j.adhoc.2022.103026.

[140]M. Zhao, D. Qin, R. Guo, and G. Xu, "Efficient Protection Mechanism Based on Self-Adaptive Decision for Communication Networks of Autonomous Vehicles," *Mobile Information Systems*, vol. 2020, pp. 1–9, Jun. 2020, doi: 10.1155/2020/2168086.

[141]A. Al-Sabaawi, K. Al-Dulaimi, E. Foo, and M. Alazab, "Addressing Malware Attacks on Connected and Autonomous Vehicles: Recent Techniques and Challenges," in *Malware Analysis Using Artificial Intelligence and Deep Learning*, Cham: Springer International Publishing, 2021, pp. 97–119. doi: 10.1007/978-3-030-62582-5_4.

[142]A. Acar, H. Aksu, A. S. Uluagac, and M. Conti, "A Survey on Homomorphic Encryption Schemes," *ACM Comput Surv*, vol. 51, no. 4, pp. 1–35, Jul. 2019, doi: 10.1145/3214303.

[143]X. Sun, F. R. Yu, P. Zhang, W. Xie, and X. Peng, "A Survey on Secure Computation Based on Homomorphic Encryption in Vehicular Ad Hoc Networks," *Sensors*, vol. 20, no. 15, p. 4253, Jul. 2020, doi: 10.3390/s20154253.

[144]F. Zhuang *et al.*, "A Comprehensive Survey on Transfer Learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021, doi: 10.1109/JPROC.2020.3004555.

Handling Imbalanced Data in Medical Records Using Entropy with Minkowski Distance

Lastri Widya Astuti[1], Ermatita[2*], Dian Palupi Rini[3]
Doctoral Program in Engineering Science, Universitas Sriwijaya, Palembang, Indonesia[1]
Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia[2, 3]
Faculty of Computer & Science, Universitas Indo Global Mandiri, Palembang, Indonesia[1]

***Abstract*—Medical records are essential for disease detection to help establish a diagnosis. Many issues with imbalanced classification are discovered in many cases of early disease detection and diagnosis using machine learning methods, resulting in decreased accuracy values due to imbalanced data distribution caused by the number of positive patients with less disease than normal individuals. To improve the accuracy of the results, a classification architectural model is proposed through a modified oversampling method (SMOTE) using Minkowski distance and adding entropy as a weight value estimation to figure out the number of samples to be made. The feature selection procedure will adopt the hybrid Particle Swarm Optimisation Grey-Wolf Optimisation approach (PSO GWO). Dataset selection evaluated high, medium, and low data dimensions based on the number of features and the total number of dataset samples. The six classification algorithms were compared using datasets involving diabetes, heart, and breast cancer. The final classification results indicated an average accuracy of 74% for diabetes, 83% for heart, and 96% for breast cancer. The proposed approach successfully solves imbalances in medical record data, outperforming Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), and Random Forest classification approaches.**

***Keywords*—Medical record; imbalanced data; classification; distance; entropy**

I. INTRODUCTION

Developing professional and quality health services requires a knowledge base to provide health services for patients based on evidence of the latest medical actions obtained from medical records of every action performed on patients [1]. Medical records are considered a key factor for improving the quality and safety of health services [2]. The use of medical records as documentation of patient care is increasingly being used to understand the clinical manifestations of various diseases, as well as the relationship between multiple diseases [3] [4]. Different knowledge bases have been engineered to extract useful information from medical records to assist the learning process for health professionals in making diagnoses [5] [6].

Medical record data stores many critical patient attributes from several medical action records and patient health records that describe a health condition [7] [8]. In many cases of early disease detection and diagnosis using machine learning methods, the problem of unbalanced classification is a case that is often encountered where the distribution of negative patients is greater than that of positive patients with the disease, causing an imbalance [9]. Several previous studies have shown that imbalanced data causes a decrease in accuracy values due to unequal class distribution, causing minority classes to be grouped into the majority class [10].

Several improvement methods are proposed to balance the amount of data in different classes through integration methods or improving learning methods by adding samples to the minority class (oversampling) or by removing samples from the majority class (undersampling) [11]. Unbalanced algorithms have three learning approach methods: data level, algorithm level, and hybrid methods [12]. The data level method improves classification performance by changing the distribution of data sets based on classes to balance the data. In contrast, the algorithm level method modifies the classification algorithm to adapt to unbalanced data structures, while the hybrid method combines the data level method and the algorithm level method in an integrated manner. Simultaneously, it is necessary to achieve the best performance [5] [13].

Oversampling techniques have been proven effective in resolving data imbalance problems. The Synthetic Minority Oversampling Technique (SMOTE) is the most popular technique for overcoming data imbalance [14]. The SMOTE method interpolates several data points from the minority class by considering neighbouring point data to produce new samples, but the interpolation process is ineffective for high-dimensional data [15] [16]. SMOTE uses Euclidean distance to calculate the neighbour distance for each minority class instance. They often increase overlap because the distance between each other is almost uniform and converges to the same value for all the cases, thereby reducing the accuracy value during the classification process [17]. Several previous studies have made improvements to the SMOTE method, including adding weights [18], learning rate [19], or combining two algorithms to improve performance and accuracy [20]. This research proposes modifications to SMOTE by using Minkowski distance and adding entropy as a weight value calculation to calculate the number of samples to be made. Minkowski distance is a measurement metric in normed vector spaces built on the distance spectrum and is considered a generalization of Euclidean and Manhattan distance. Minkowski distance can also describe the geometric structure of majority and minority class data [21]. The plan being developed for ensemble models will integrate the balanced data and then select the best features to be utilized by the classification method to achieve the highest level of accuracy. All stages will be carried out in a structured manner following the proposed design, and the research contribution will be determined so that the direction is clear and measurable.

The following are the contributions to this research:

1) This research presents improvements to the SMOTE method to overcome data imbalance by adding entropy as a weight to generate the number of data instances and changing the distance learning method to overcome overfitting.

2) The feature selection process in this research presents a hybrid architecture that combines Particle Swarm Optimization (PSO) with Grey Wolf Optimization (GWO). Combining two algorithms for feature selection can reduce the number of attributes to increase accuracy and computing time.

3) This research compares several classification methods, such as Decision Tree, K-NN, Logistic Regression, Naïve Bayes, Random Forest, and SVM, to see the best performance of the six methods used in disease classification research experiments.

4) Ensemble learning techniques improve medical record data classification results in this research. This research aims to optimize the accuracy of classification results by using the advantages of various model comparisons to overcome the complexity of data imbalance and reduce data dimensions. That offers a perspective for improving the accuracy of pattern recognition results in medical record data.

II. Literature Review

A. Imbalance Data

Class imbalance is a condition where the number of *instances of* the majority class is greater than the number *of cases of the* minority class. The difference in the number of samples in each class in the classification is known as an unbalanced data set [22]. Extension *ensemble* for unbalanced learning considers changes in the pattern probability function during the training phase, where training on unbalanced data tends to overtrain the majority class, while training for the minority class becomes difficult to predict due to undertraining [23][24]. Data imbalance can be divided into two types, namely relative imbalance and absolute imbalance, while based on the cause of the problem, the imbalance can be divided into two: intrinsic and extrinsic. Intrinsic indicates an inherent property of the data. For example, the probability of equipment failing is much lower than standard running equipment, or the number of people with cancer is less than that of healthy people. Extrinsic means other or external factors that cause data imbalance. For example, sporadic interruptions occur when balanced data is transmitted to the database [25].

B. Feature Selection

Dimensionality reduction is done by selecting features to increase accuracy by eliminating irrelevant features and assigning selected features to make the data easier to understand [26]. In the feature determination process, there are three approaches used to select features:

1) Filter method, selecting features based on feature relevance using ranking criteria, where the lowest calculated value will be removed. The filter method does not depend on the classification algorithm used in modelling so that it can be generalized and impacts accelerating computing time.

2) The wrapper method is part of a supervised algorithm where features are selected based on the relationship between features by comparing the resulting subset of features. The training and testing process is based on predicting the representativeness of each feature, which causes the computing time to become more complex.

3) The embedded method performs feature selection by looking for an optimal subset of features and including them in the training process of the classifier, thereby reducing computing time.

C. Classification

Utilization of machine learning helping to enforce fast, accurate diagnosis and detection is one of the efforts made in the health sector. Algorithm development machine learning, used to predict and diagnose disease based on medical record data, can help reduce diagnosis errors and independent pre-diagnosis learning materials. Several studies are related to solving or detecting diseases through classification methods, including the K-NN method, which is used to classify medical health data related to diseases and drugs [27] [28]. Method Naïve Bayes and Random Forest, which is used to diagnose and detect various diseases [29] [30], the SVM method is used to classify medical record data in the form of images [31] [32]. The utilization of machine learning also touches on digital data-based disease classification data processing, such as the utilization of telehealth and IoT using the decision tree method [33]. Combining two machine learning methods or modifying a technique has been developed to improve medical record data classification results and diagnose diseases [34]. The classification results of medical record data in previous research show variations in accuracy. Several factors influence the level of accuracy, including high-dimensional data and data imbalance.

The model proposed in this research is to overcome data imbalance in medical record data, where in several previous studies, many modifications or additional functions have been made to overcome the imbalance problem by adding entropy and changing the distance calculation to make it more flexible for data with different dimensions. Meanwhile, the feature selection process uses a metaheuristic method that focuses on finding optimal solutions flexibly. PSOGWO is used to optimize subsets to get features that provide the best performance on the model. This model was tested using six algorithms: decision tree, logistic regression, naïve Bayes, KNN, random forest, and SVM to test the ensemble combination that produces the most optimal level of accuracy.

III. Methodology

The methodology used in the research contains comprehensive stages of each built model process. The stages for overcoming the data imbalance problem begin with preparing a medical record dataset, data preprocessing, feature selection, classification, and evaluation methods. The research stages are shown in Fig. 1.

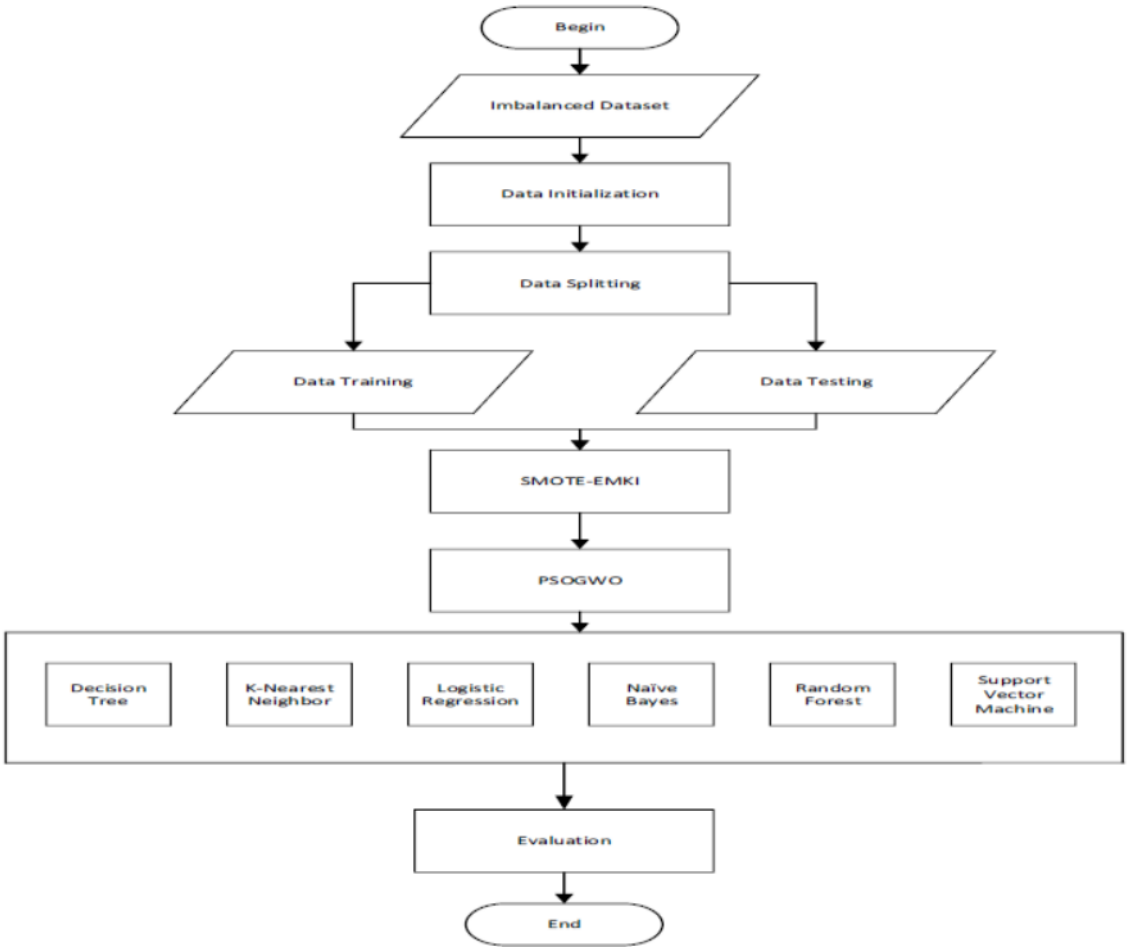


Fig. 1. The proposed architecture.

A. Dataset

The dataset used in this research is medical record data for diabetes, heart disease, and breast cancer, accessed in the UC repository *machine learning* and Kaggle. The selection of medical record data is based on the prevalence rate of sufferers, which increases yearly. Detailed data consists of the number of features, number of example data, number of minor data, number of classes, and Imbalanced Ratio (IR), described in Table I.

TABLE I. DESCRIPTION OF MEDICAL RECORDS DATASET

Dataset	Feature	Instance	Minority	Class	IR
Diabetes	8	768	268	2	1.86
Heart	14	1025	499	2	1.054
Breast Cancer	32	569	212	2	1.68

B. Data Imbalance Method

The approach used in this research is to modify the SMOTE algorithm by adding an entropy value at the beginning of the data balancing process and changing the distance calculation method from Euclidean Distance to Minkowski with the advantage of distance adjustment capabilities. The development stages of this model start from:

Step 1: Calculate the entropy weight for each majority and minority class. Entropy is used to measure how balanced the class distribution is in the dataset, where the entropy calculation for each class is carried out before the distance calculation. The entropy equation used is:

$$Entropy\ (S) = \ \sum_{i=1}^{C} p_{i\ log_2}(p_i) \qquad (1)$$

High entropy values indicate a more significant imbalance, while low entropy values indicate a more balanced class distribution.

Step 2: Determine the number of synthetic samples. The number of synthetic samples that need to be added is calculated based on the entropy proportion of the majority and minority

classes. If the entropy of the minority class is higher, the class imbalance is more significant, so many synthetic samples need to be added.

Step 3: Calculate the distance according to the given weight. After calculating the entropy value, normalization is carried out before proceeding to the distance calculation. Where the previous distance calculation using Euclidean distance was changed using Minkowski to calculate the closest K value from the minority class with the equation:

$$D_{(p,q)} = \left(\sum_{i=1}^{n} |p_i - q_i|^r\right)^{1/r} \quad (2)$$

Step 4: Synthetic interpolation creates synthetic samples based on linear interpolation between minority data points and their nearest neighbours.

One of the problems in SMOTE is that the synthetic samples produced have a high level of similarity to existing samples, so they do not adequately represent the diversity of minority classes. Adding entropy in the SMOTE technique is intended to improve the quality of synthetic samples and reduce the possibility of bias. Entropy is used to help select sample points and interpolation, where the results can increase the representativeness and variation of synthetic samples. Entropy positioning is also a consideration in this research; entropy is placed before applying the SMOTE method with consideration of the entropy value used to determine the number of synthetic samples needed. The entropy value also helps assess relevant features and eliminates the possibility of redundant minority classes. Placing entropy before the SMOTE process helps increase data diversity, accelerates convergence, and reduces the potential for overfitting.

A part from adding entropy, the proposed model also changes the distance calculation technique from Euclidean Distance to Minkowski Distance. This change aims to provide flexibility in measuring the distance between data points for data with different dimensions, thereby increasing the quality of synthetic samples. Changing the distance calculation method to Minkowski distance also helps determine the nearest neighbour selection of parameters to produce synthetic samples. In addition, the parameters can be adjusted to obtain distance measurements that better suit the characteristics of the data or specific features so that when the parameters are chosen appropriately, the diversity of the resulting synthetic samples can be increased. The SMOTE algorithm has been modified by adding entropy and replacing the distance approach with Minkowski, now called EMKI SMOTE. Furthermore, the results of improving this method will be used in the classification process, which was previously preceded by a feature selection process.

C. Algorithm PSO GWO

One of the population-based metaheuristic optimization techniques is Particle Swarm Optimization (PSO), where the initial inspiration for this method was the social behavior of flocks of birds or schools of fish when foraging for food [35] [36]. In this technique, a number of particles are allowed to move in a multidimensional search space, where an initial population is generated randomly in the search domain [37] [38]. All particles in the moving swarm update their positions using Eq. (3) and Eq. (4) at each iteration. The entire change history of the best position information of each particle in the cluster is also updated and saved.

$$X^{-i}{}_{n+1} = X_n^{-1} + V_{n+1}^{-1}, \quad (3)$$

$$V^{-i}{}_{n+1} = \omega V^{-1}{}_n + c_1 r_1 \left(\overline{p}^i{}_n - \overline{x}^i{}_n\right) \quad (4)$$

The particle symbols used in swarm for optimization are as follows:

N: The iteration steps carried out

r1 dan r2: Value representing a random number in the range [0, 1]

ω : Parameter weight inertia

c1 dan c2 : Coefficients representing optimization parameters

x : Position vector

v : Velocity vector

$\overline{x}^i$: The best position that the i particle has achieved

$\overline{p}^g$: The best position representation available on the *swarm*.

In the PSO algorithm, the position and velocity of the particles are determined randomly in the search space to find the best value. The goal of this operation is to avoid local minima. The search is continued until optimal results are achieved or based on achieving a predetermined maximum number of iterations. The iteration process is carried out to obtain the best solution through the following equation:

$$x_i^{k+1} = \begin{Bmatrix} x_i new^{i \iint (x_i new) \leq \int (x_i)} \\ x_i otherwise \end{Bmatrix} \quad (5)$$

The grey wolf leadership hierarchy at the top of the food chain is the rationale for the GWO algorithm. The grey wolf group is divided into four categories, alpha, beta, delta and omega. The alpha wolf represents the highest hierarchy, which provides the best solution for the group, while the second and third hierarchies are occupied by the beta and delta wolves, respectively. The omega wolf's role in the pack as the best solution candidate in the pack if needed. The stages in hunting prey by groups of grey wolves are divided into three parts, namely: the first stage is carrying out reconnaissance on the prey to be hunted, the second stage is stopping the movement of the prey by circling and confining it so that it makes it easier to attack and prey on the victim as the final stage. Eq. (6) and Eq. (7) provide a mathematical model for surrounding the victim.

$$D = \left|C \times X_p(t) - X(t)\right| \quad (6)$$

$$X(t+1) = X_p(t) - A \times D \quad (7)$$

Here, t denotes the number of instantaneous iterations, X_p is the prey position, X is the grey wolf's location, and A and C

are the vector coefficients. The coefficients A and C are calculated as shown in Eq. (8) and Eq. (9).

$$A = a \times (2 \times r_1 - 1) \quad (8)$$

$$C = 2 \times r_2 \quad (9)$$

The number of drops linearly from two to zero as the number of iterations decreases. r_1 and r_2 are random numbers picked uniformly from [0,1].

Alpha wolves direct grey wolves to the location of their prey. Alpha wolves occasionally receive assistance from beta and delta wolves. The GWO method assumes that the alpha wolf is the best option, with beta and delta wolves coming in second and third. As a result, the other wolves in the population migrated following the positions of these three wolves. The formulated mathematically in Eq. (10), (11), and (12):

$$D_\alpha = |C_1 \times X_\alpha - X(t)| \quad (10)$$

$$D_\beta = |C_2 \times X_\beta - X(t)| \quad (11)$$

$$D_\delta = |C_3 \times X_\delta - X(t)| \quad (12)$$

Values X_α, X_β, and X_δ each represent the three best wolves in each iteration.

$$X_1 = |X_\alpha - a_1 D_\alpha| \quad (13)$$

$$X_2 = |X_\beta - a_2 D_\beta| \quad (14)$$

$$X_3 = |X_\delta - a_3 D_\delta| \quad (15)$$

$$X_P(t+1) = \frac{X_1 + X_2 + X_3}{3} \quad (16)$$

Here, the new prey position is expressed as Xp(t + 1) as the mean of the positions of the three best wolves in the population. The grey wolf completes the hunt by attacking its prey. To attack, they must be close enough to their prey. When Eq. (5) is checked, A takes varying values from [− 2a, 2a] while A takes decreasing values from 2 to 0. When the |A| value exceeds or equals 1, the existing hunt is abandoned to find a better solution. The grey wolf is forced to attack the prey if the prey is close enough to a value less than 1. This approach prevents wolves from getting stuck at local minima. The search is complete when the GWO algorithm reaches the desired number of iterations.

D. Evaluation Metric

An evaluation of model performance is needed to see the effectiveness of the proposed ensemble model. Model performance is measured using accuracy, precision, recall, and F1 score metrics based on actual and predicted results from the classification process, where the basis for calculating metrics can be defined as follows: True Positive (TP) notation: the amount of data labelled positive and classified by the model labelled positive; True Negative (TN) is the amount of data labelled negative and classified as negative; False Positive (FP) is the amount of actual data labelled negative and predicted positive; False Negative (FN) is the amount of actual data labelled positive and predicted negative.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (17)$$

$$Precision = \frac{TP}{TP+FP} \quad (18)$$

$$Recall = \frac{TP}{TP+FN} \quad (19)$$

$$F1\ score = 2\,x\,\frac{Precision\ x\ Recall}{Precision+Recall} \quad (20)$$

IV. Result and Discussion

The designed EMKI SMOTE model causes the process of creating synthetic samples in the minority class to become more dynamic and diverse. The modification results in Table II show a significant improvement in overcoming data imbalance. The imbalance ratio (IR) in the Table compares the ratios before and after the modification was applied.

TABLE II. Imbalance Ratio

Dataset	Imbalance Ratio	
	SMOTE	EMKI SMOTE
Diabetes	1: 1.86	1: 1.47
Heart	1: 1.05	1: 1.02
Breast Cancer	1: 1.68	1: 1.43

Applying the EMKI SMOTE method to the model allows for the balance of data and the production of diverse synthetic samples so that the minority class data is more representative. Adding entropy and the Minkowski method for calculating distance can also eliminate redundant sample selection, reducing the risk of overfitting. After completing the data imbalance stage, the model will continue with the feature selection stage.

Particle Swarm Optimization - Grey Wolf Optimization (PSO-GWO) feature selection is a hybrid method that takes advantage of the benefits of both optimization algorithms to pick the most relevant subset of features in a dataset. PSO (Particle Swarm Optimization) is an algorithm inspired by the behaviour of flocks of birds or fish, in which each particle represents a solution and interacts with the others to discover the optimal position. Meanwhile, GWO (Grey Wolf Optimization) is an algorithm inspired by grey wolf hunting behaviours that uses a highly efficient search mechanism that divides roles among alpha, beta, and delta wolves. In feature selection, PSO identifies an initial solution, and GWO enhances the search for the solution by steering it in a more efficient path. The combination of these two algorithms aids in overcoming the dataset's complexity and feature redundancy, resulting in a more efficient model with improved performance. The PSO-GWO method can avoid traps in local search and speed up feature selection while maintaining model fidelity.

Feature selection in this study uses a combined algorithm between Particle Swarm Optimization and Grey Wolf Optimization (PSO-GWO) to produce the best feature selection process from the tested medical record datasets. The selection of features using this method will select attributes as representatives for each dataset. Fig. 2 shows the results of feature selection before using the feature selection method, after using the PSO GWO method, and the results of feature selection after the dataset is balanced, showing the results of feature selection with the optimal number of features without losing information for the classification process.

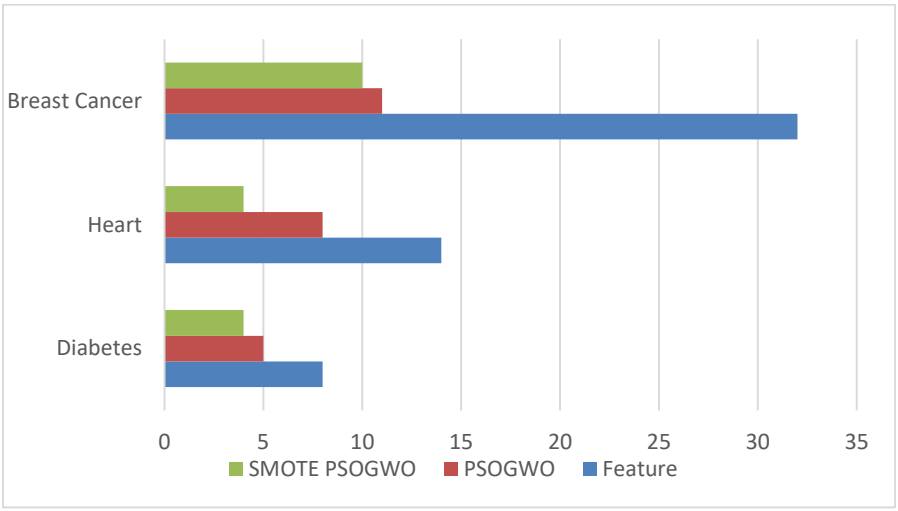


Fig. 2. Feature selection results using the PSO GWO.

The classification process divides the dataset into two parts: training data and testing data. The data is divided proportionally using the holdout method. Training data helps the model recognize patterns or relationships between features and labels while testing data measures the accuracy or significance of the model in classifying data that has not been introduced previously.

The experiment used medical record data for diabetes, heart disease, and breast cancer accessed from the UCI Machine Learning Repository. The results of feature selection with balanced data then go through the classification stage. The classification process uses a Decision Tree, KNN, Logistic Regression, Naïve Bayes, Random Forest, and SVM methods. The classification method comparison process is intended to strengthen evidence regarding the performance of the proposed model. Fig. 3 shows a confusion matrix graph for testing data on diabetes.

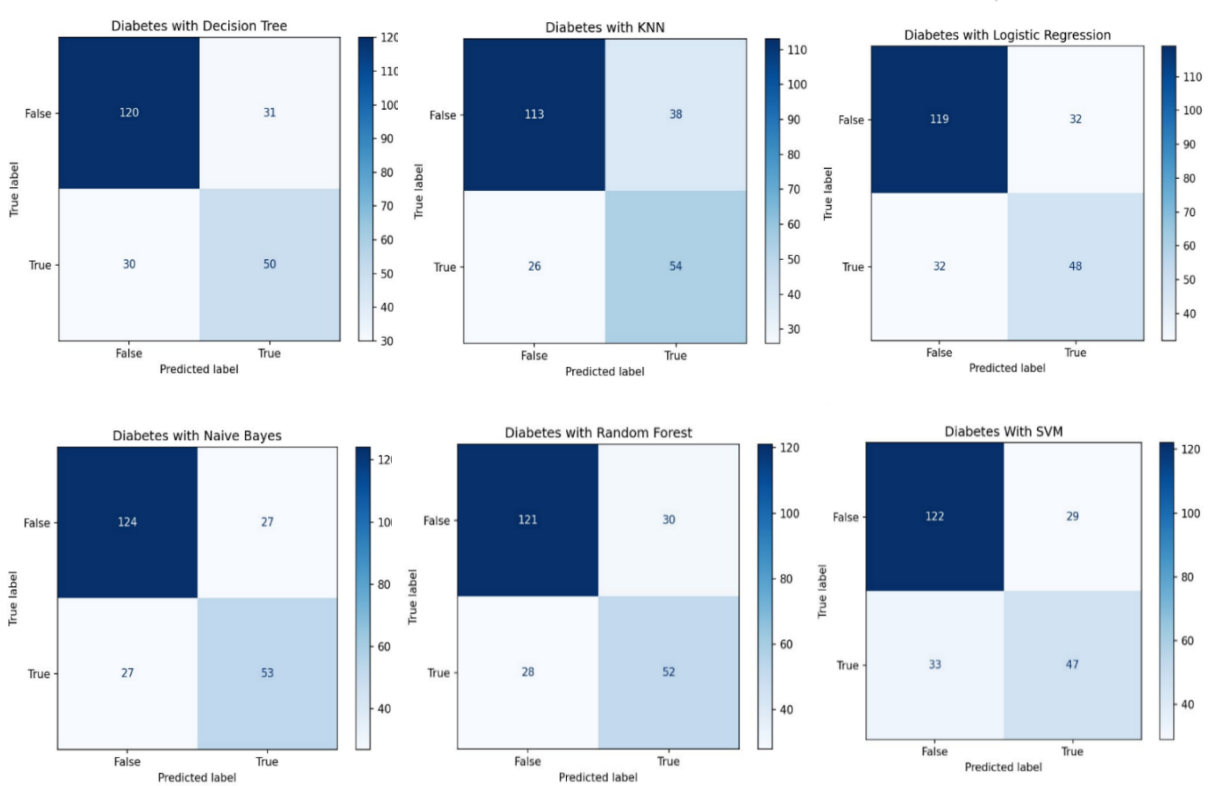


Fig. 3. Confusion matrix for the diabetes dataset.

The test results show that the diabetes dataset, when tested using six classification methods, shows an average Accuracy value of 0.7411, with an average value for precision of 0.79355, while an average value for recall is 0.8026 and an average F1 score of 0.7984. The methods that provide the best performance results are the Naïve Bayes and Random Forest methods, while the lowest performance results in experiments using the Decision Tree method. Details of the experimental results using each classification method are shown in Table III.

TABLE III. EVALUATION MATRIX DIABETES

Dataset	Classifier	Accuracy	Precision	Recall	F1 Score
Diabetes	Decision Tree	0.7359	0.7947	0.7947	0.7946
	K Nearest Neighbor	0.7229	0.7483	0.8129	0.7792
	Logistic Regression	0.7229	0.7880	0.7880	0.7887
	Naïve Bayes	0.7662	0.8211	0.8211	0.8210
	Random Forest	0.7489	0.8013	0.8120	0.8066
	Support Vector Machine	0.7316	0.8079	0.7870	0.8008

The second test was carried out on the heart disease dataset, where the classification process was given the same treatment for each stage as in the diabetes dataset. The confusion matrix of the results of heart disease data testing is shown in Fig. 4. The experimental results show increased accuracy, precision, recall, and F1 score after the medical data was balanced using EMKI SMOTE. Analysis of the experimental results also shows that the PSO GWO method used in the feature selection process combined with EMKI SMOTE produces several relevant features to use during the classification process.

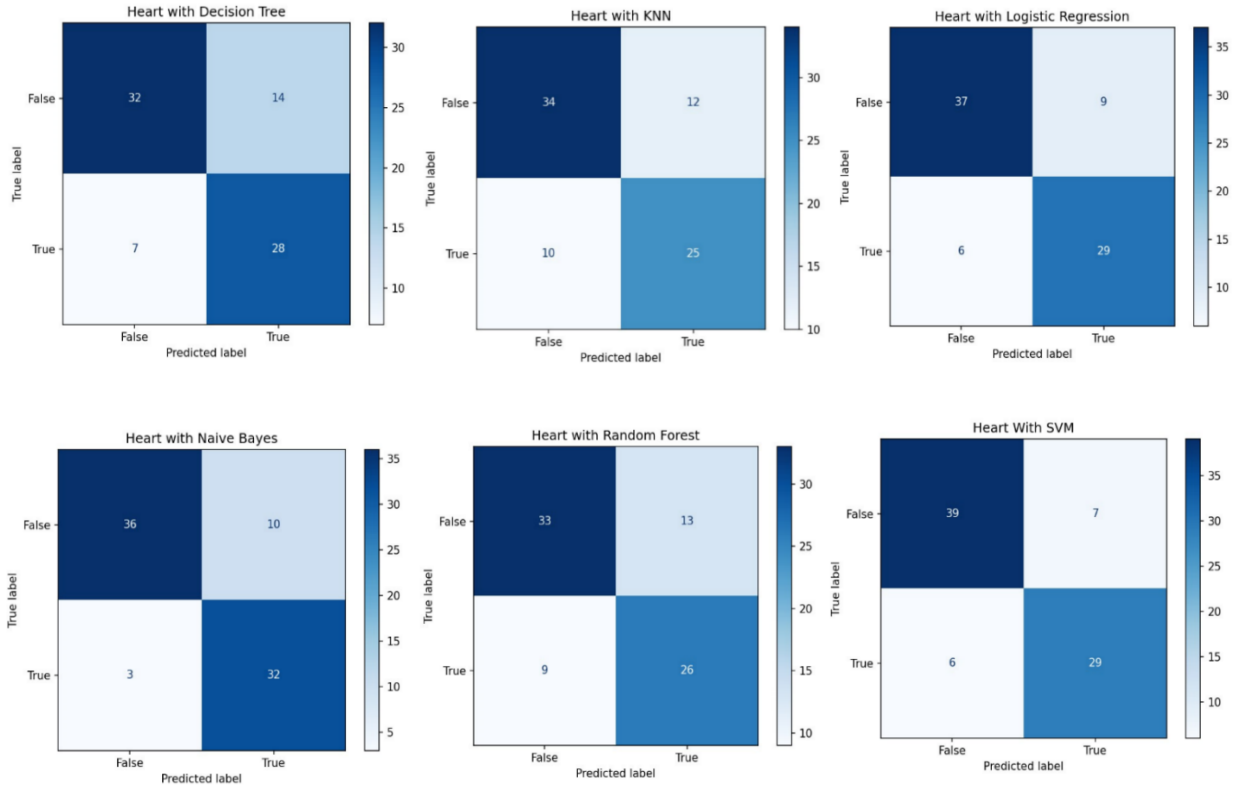


Fig. 4. Confusion matrix for heart disease data.

Experiments using the heart disease dataset showed an average accuracy of 0.7818, precision of 0.7644, recall of 0.8381, and F1 Score of 0.7989. Meanwhile, the best-combined method is EMKI SMOTE, PSO GWO combined with SVM, Naïve Bayes, and Logistic Regression, which is shown as a whole in Table IV.

The third experiment was carried out on the breast cancer dataset, with 32 features. The experimental results of the model also show that the feature selection results are significant and relevant, thus influencing the improvement of experimental results. The Minority class value is lower than the majority, with an imbalance ratio of 1.68, after being corrected with the proposed model, indicating good and even experimental results for the six classification methods.

TABLE IV. EVALUATION MATRIX HEART

Dataset	Classifier	Accuracy	Precision	Recall	F1 Score
Heart	Decision Tree	0.7407	0.6956	0.8205	0.7527
	K Nearest Neighbor	0.7283	0.7391	0.7727	0.7555
	Logistic Regression	0.8148	0.8043	0.8604	0.8313
	Naïve Bayes	0.8395	0.7826	0.9230	0.8469
	Random Forest	0.7283	0.7173	0.7857	0.7499
	Support Vector Machine	0.8395	0.8478	0.8667	0.8571

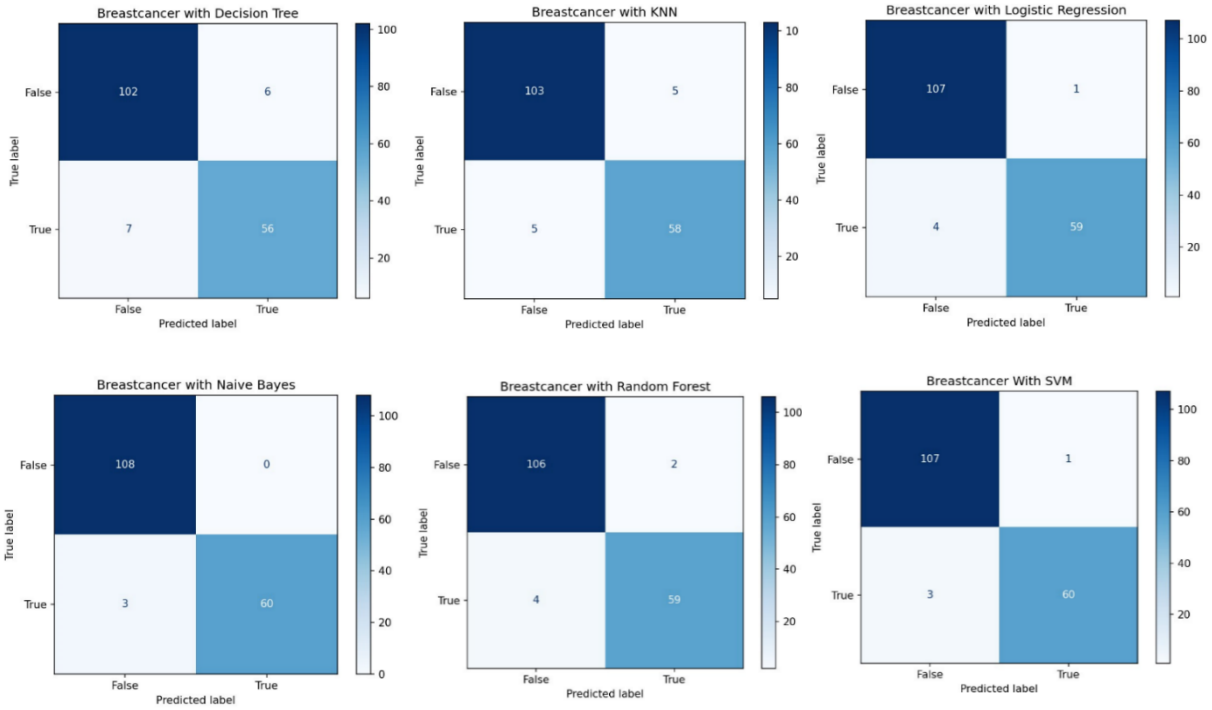


Fig. 5. Confusion matrix for breast cancer.

The application of the EMKI SMOTE model combined with PSO GWO produces averages for accuracy, precision, recall, and F1 Score in the range of 0.9684 with the best results in the Naïve Bayes, SVM, and Logistic Regression classification methods as shown in Table V.

The application of the EMKI SMOTE model combined with the PSO GWO method on medical record data was due to the discovery of a lot of unbalanced data, with the prevalence of patients in normal conditions being higher than patients suffering from disease. The experiment also selected datasets by considering high, medium, and low data dimensions based on the number of features and the total number of dataset samples. Analysis of experimental results from various aspects shows a significant increase in results, and computing time calculations also indicate that the training and testing process is faster.

TABLE V. EVALUATION MATRIX BREAST CANCER

Dataset	Classifier	Accuracy	Precision	Recall	F1 Score
Breast Cancer	Decision Tree	0.9239	0.9444	0.9357	0.9399
	K Nearest Neighbor	0.9415	0.9537	0.9537	0.9536
	Logistic Regression	0.9707	0.9907	0.9639	0.9770
	Naïve Bayes	0.9824	1.00	0.9729	0.9863
	Random Forest	0.9649	0.9814	0.9636	0.9723
	Support Vector Machine	0.9766	0.9907	0.9727	0.9816

V. CONCLUSION

Imbalanced class data has become an essential problem in medical diagnosis, but using classification algorithms alone cannot meet the classification needs effectively. This research proposes a combined method by modifying the distance calculation using Minkowski and adding entropy to determine the value of the number of samples created. The proposed calculation model is hereafter called EMKI SMOTE. The EMKI SMOTE oversampling approach is designed to handle imbalances between majority and minority classes in data so that it can produce relevant features in high-dimensional data. Next, the balanced data will be ensembled using the PSO GWO method in the feature selection process. The combined results of the data balance method and feature selection will then be used in the training and testing process using the classification method. The research results show increased accuracy, especially in the Naïve Bayes method, Logistic Regression, and SVM using diabetes, heart, and breast cancer datasets. The further development of this research model is to obtain constant balanced data, namely, by determining the interval between the majority and minority sample numbers. In addition to adding parameters to the dataset model configuration, it is possible to balance the data, which causes an increase in accuracy. Meanwhile, testing large amounts of data and high dimensions by making several substitutions in the feature selection method is also considered to increase the accuracy of classification results and help in the problem of early disease detection.

ACKNOWLEDGMENT

We are grateful to everyone who helped with this study, especially the promoter and co-promoter who advised completing it. In addition, I want to convey my heartfelt gratitude to the Indo Global Mandiri University for allowing me to improve personally.

REFERENCES

[1] Y. Huang, W. Bian, and Y. Han, "Effect of knowledge acquisition on gravida's anxiety during COVID-19," Sex. Reprod. Healthc., vol. 30, no. 639, p. 100667, 2021, doi: 10.1016/j.srhc.2021.100667.

[2] A. De Benedictis, E. Lettieri, L. Gastaldi, C. Masella, A. Urgu, and D. Tartaglini, "Electronic medical records implementation in hospital: An empirical investigation of individual and organizational determinants," PLoS One, vol. 15, no. 6, pp. 1–12, 2020, doi: 10.1371/journal.pone.0234108.

[3] A. Caccamisi, L. Jørgensen, H. Dalianis, and M. Rosenlund, "Natural language processing and machine learning to enable automatic extraction and classification of patients' smoking status from electronic medical records," Ups. J. Med. Sci., vol. 125, no. 4, pp. 316–324, 2020, doi: 10.1080/03009734.2020.1792010.

[4] H. L. Lin, D. C. Wu, S. M. Cheng, C. J. Chen, M. C. Wang, and C. A. Cheng, "Association between Electronic Medical Records and Healthcare Quality," Med. (United States), vol. 99, no. 31, 2020, doi: 10.1097/MD.0000000000021182.

[5] M. Khushi et al., "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," IEEE Access, vol. 9, pp. 109960–109975, 2021, doi: 10.1109/ACCESS.2021.3102399.

[6] H. Goodrum, K. Roberts, and E. V. Bernstam, "Automatic classification of scanned electronic health record documents," Int. J. Med. Inform., vol. 144, no. September, p. 104302, 2020, doi: 10.1016/j.ijmedinf.2020.104302.

[7] X. Li, H. Wang, H. He, J. Du, J. Chen, and J. Wu, "Intelligent diagnosis with Chinese electronic medical records based on convolutional neural networks," BMC Bioinformatics, vol. 20, no. 1, pp. 1–12, 2019, doi: 10.1186/s12859-019-2617-8.

[8] J. Xu, X. Xi, J. Chen, V. S. Sheng, J. Ma, and Z. Cui, "A Survey of Deep Learning for Electronic Health Records," Appl. Sci., vol. 12, no. 22, pp. 1–24, 2022, doi: 10.3390/app122211709.

[9] W. Zhang, X. Li, X. D. Jia, H. Ma, Z. Luo, and X. Li, "Machinery fault diagnosis with imbalanced data using deep generative adversarial networks," Meas. J. Int. Meas. Confed., vol. 152, 2020, doi: 10.1016/j.measurement.2019.107377.

[10] M. Hayaty, S. Muthmainah, and S. M. Ghufran, "Random and Synthetic Over-Sampling Approach to Resolve Data Imbalance in Classification," Int. J. Artif. Intell. Res., vol. 4, no. 2, p. 86, 2021, doi: 10.29099/ijair.v4i2.152.

[11] X. W. Liang, A. P. Jiang, T. Li, Y. Y. Xue, and G. T. Wang, "Knowledge-Based Systems LR-SMOTE — An improved unbalanced data set oversampling based on K-means and SVM," vol. 196, 2020, doi: 10.1016/j.knosys.2020.105845.

[12] Q. Dai, J. Liu, and Y. Liu, "Multi-granularity Relabeled Under-sampling Algorithm for Imbalanced Data," 2022.

[13] C. Kaope and Y. Pristyanto, "The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance," MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput., vol. 22, no. 2, pp. 227–238, 2023, doi: 10.30812/matrik.v22i2.2515.

[14] H. A. Gameng, B. B. Gerardo, and R. P. Medina, "Modified Adaptive Synthetic SMOTE to Improve Classification Performance in Imbalanced Datasets," ICETAS 2019 - 2019 6th IEEE Int. Conf. Eng. Technol. Appl. Sci., pp. 19–23, 2019, doi: 10.1109/ICETAS48360.2019.9117287.

[15] S. Maldonado, J. López, and C. Vairetti, "An alternative SMOTE oversampling strategy for high-dimensional datasets," Appl. Soft Comput. J., 2018, doi: 10.1016/j.asoc.2018.12.024.

[16] A. S. Hussein, T. Li, C. W. Yohannese, and K. Bashir, "A-SMOTE: A new preprocessing approach for highly imbalanced datasets by improving SMOTE," Int. J. Comput. Intell. Syst., vol. 12, no. 2, pp. 1412–1422, 2019, doi: 10.2991/ijcis.d.191114.002.

[17] E. A. Pambudi and A. G. Fahrezi, "Forecasting Brown Sugar Production Using k-NN Minkowski Distance and Z-Score Normalization," vol. 5, no. 2, pp. 580–589, 2023, doi: 10.51519/journalisi.v5i2.485.

[18] S. Maldonado, C. Vairetti, A. Fernandez, and F. Herrera, "FW-SMOTE : A feature-weighted oversampling approach for imbalanced classification," vol. 124, 2022, doi: 10.1016/j.patcog.2021.108511.

[19] Y. Wang, Y. Wei, H. Yang, J. Li, Y. Zhou, and Q. Wu, "Utilizing imbalanced electronic health records to predict acute kidney injury by ensemble learning and time series model," BMC Med. Inform. Decis. Mak., vol. 20, no. 1, pp. 1–13, 2020, doi: 10.1186/s12911-020-01245-4.

[20] N. G. Ramadhan, "Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus," Sci. J. Informatics, vol. 8, no. 2, pp. 276–282, 2021, doi: 10.15294/sji.v8i2.32484.

[21] C. Fu, "Granular Classification for Imbalanced Datasets : A Minkowski Distance-Based Method," 2021.

[22] K. Roy et al., "An Enhanced Machine Learning Framework for Type 2 Diabetes Classification Using Imbalanced Data with Missing Values," Complexity, vol. 2021, 2021, doi: 10.1155/2021/9953314.

[23] D. Fahrudy and S. 'uyun, "Classification of Student Graduation by Naïve Bayes Method by Comparing between Random Oversampling and Feature Selections of Information Gain and Forward Selection," Int. J. Informatics Vis., vol. 6, no. 4, pp. 798–808, 2022, doi: 10.30630/joiv.6.4.982.

[24] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, "Classification of Imbalanced Data:Review of Methods and Applications," IOP Conf. Ser. Mater. Sci. Eng., vol. 1099, no. 1, p. 012077, 2021, doi: 10.1088/1757-899x/1099/1/012077.

[25] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," J. Big Data, vol. 5, no. 1, 2018, doi: 10.1186/s40537-018-0151-6.

[26] A. G. Hussien, D. Oliva, E. H. Houssein, A. A. Juan, and X. Yu, "Binary whale optimization algorithm for dimensionality reduction," Mathematics, vol. 8, no. 10, pp. 1–24, 2020, doi: 10.3390/math8101821.

[27] W. Xing and Y. Bei, "Medical Health Big Data Classification Based on KNN Classification Algorithm," IEEE Access, vol. 8, pp. 28808–28819, 2020, doi: 10.1109/ACCESS.2019.2955754.

[28] F. Aldi, I. Nozomi, and S. Soeheri, "Comparison of Drug Type Classification Performance Using KNN Algorithm," SinkrOn, vol. 7, no. 3, pp. 1028–1034, 2022, doi: 10.33395/sinkron.v7i3.11487.

[29] T. Widiyaningtyas, I. A. E. Zaeni, and N. Jamilah, "Diagnosis of fever symptoms using naive bayes algorithm," ACM Int. Conf. Proceeding Ser., pp. 23–28, 2020, doi: 10.1145/3427423.3427426.

[30] J. E. Aurelia, Z. Rustam, I. Wirasati, S. Hartini, and G. S. Saragih, "Hepatitis classification using support vector machines and random forest," IAES Int. J. Artif. Intell., vol. 10, no. 2, pp. 446–451, 2021, doi: 10.11591/IJAI.V10.I2.PP446-451.

[31] J. Latif, C. Xiao, S. Tu, S. U. Rehman, A. Imran, and A. Bilal, "Implementation and Use of Disease Diagnosis Systems for Electronic Medical Records Based on Machine Learning: A Complete Review," IEEE Access, vol. 8, pp. 150489–150513, 2020, doi: 10.1109/ACCESS.2020.3016782.

[32] A. Y. Saleh, C. K. Chin, V. Penshie, and H. R. H. Al-Absi, "Lung cancer medical images classification using hybrid cnn-svm," Int. J. Adv. Intell. Informatics, vol. 7, no. 2, pp. 151–162, 2021, doi: 10.26555/ijain.v7i2.317.

[33] C. C. Chern, Y. J. Chen, and B. Hsiao, "Decision tree-based classifier in providing telehealth service," BMC Med. Inform. Decis. Mak., vol. 19, no. 1, pp. 1–15, 2019, doi: 10.1186/s12911-019-0825-9.

[34] C. Iwendi et al., "COVID-19 patient health prediction using boosted random forest algorithm," Front. Public Heal., vol. 8, no. July, pp. 1–9, 2020, doi: 10.3389/fpubh.2020.00357.

[35] Q. Al-Tashi, S. J. Abdul Kadir, H. M. Rais, S. Mirjalili, and H. Alhussian, "Binary Optimization Using Hybrid Grey Wolf Optimization for Feature Selection," IEEE Access, vol. 7, no. c, pp. 39496–39508, 2019, doi: 10.1109/ACCESS.2019.2906757.

[36] M. A. M. Shaheen, H. M. Hasanien, and A. Alkuhayli, "A novel hybrid GWO-PSO optimization technique for optimal reactive power dispatch problem solution," Ain Shams Eng. J., vol. 12, no. 1, pp. 621–630, 2021, doi: 10.1016/j.asej.2020.07.011.

[37] S. Prithi and S. Sumathi, "Automata Based Hybrid PSO–GWO Algorithm for Secured Energy Efficient Optimal Routing in Wireless Sensor Network," Wirel. Pers. Commun., vol. 117, no. 2, pp. 545–559, 2021, doi: 10.1007/s11277-020-07882-2.

[38] Z. H. A. Al-Tameemi, T. T. Lie, G. Foo, and F. Blaabjerg, "Optimal Coordinated Control of DC Microgrid Based on Hybrid PSO–GWO Algorithm," Electricity, vol. 3, no. 3, pp. 346–364, 2022, doi: 10.3390/electricity3030019

IoMT-Enabled Noninvasive Lungs Disease Detection and Classification Using Deep Learning-Based Analysis of Lungs Sounds

Muhammad Sajid[1], Wareesa Sharif[2], Ghulam Gilanie[3], Maryam Mazher[4],
Khurshid Iqbal[5], Muhammad Afzaal Akhtar[6], Muhammad Muddassar[7], Abdul Rehman[8]
Faculty of Computing, The Islamia University of Bahawalpur, Bahawalpur, Pakistan[1, 2, 3, 6, 8]
Department of Computer Science, COMSATS University Islamabad, Islamabad, Pakistan[4]
Department of Computer Science, Virtual University of Pakistan, Islamabad, Pakistan[5, 7]

***Abstract*—Noninvasive and accurate methods for diagnosing respiratory diseases are essential to improving healthcare consequences. The Internet of Medical Things (IoMT) is critical in driving developments in this field. This work presents an IoMT-enabled approach for lung disease detection and classification, using deep learning techniques to analyze lung sounds. The proposed approach uses three datasets: the Respiratory Sound, the Coronahack Respiratory Sound, and the Coswara Sound. Traditional machine learning models, including the Extra Tree Classifier and AdaBoost Classifier, are used to benchmark performance. The Extra Tree Classifier achieved 94.12%, 95.23%, and 94.21% across the datasets, while the AdaBoost Classifier showed improvements with 95.42%, 96.33%, and 94.76%. The proposed deep neural network (DNN) achieves accuracies of 98.92%, 99.33%, and 99.36% for the same datasets. This study explores the transformative potential of the Internet of Medical Things (IoMT) in augmenting diagnostic precision and advancing the field of respiratory healthcare.**

***Keywords*—Deep learning; respiratory sound; coronahack respiratory sound and coswara sound; IoMT**

I. Introduction

According to epidemiological statistics on respiratory disorders published by the World Health Organization (WHO), 210 million people worldwide suffer from chronic obstructive pulmonary disease (COPD), and 30 million people have asthma. Studies show that between 15 and 25 million people in India have asthma [1]. Physicians often use the noninvasive, low-cost lung auscultatory technique to assess the state of the lungs [2]. The noises the lungs make while air passes through them during breathing are known as lung sounds [3].

It is critical for recognizing lung diseases because it provides precise lung function data. Aberrant and accidental lung sounds are the two general classifications for lung sounds [4]. The Vesicular, Bronchial, Broncho-Vesicular, and Tracheal lung sounds are common [5]. Accidental lung sounds can be classified as constant or intermittent, depending on their duration and persistence [6]. Early identification and close observation of pneumonia are essential for adequate medical care [7]. Lung inspection is a standard clinical procedure for diagnosing respiratory disorders [8]. It involves hearing the sounds of an individual's lungs with a stethoscope. Usually, these noises are classified as either abnormal or normal [9]. The frequent unusual noises audible over characteristic lung sounds are crackles, wheezes, and squawks; they commonly exist in a lung condition [10].

Common lung sounds and cyclical patterns represent air passage during breathing. Pulmonary illnesses characterized by persistent, incomplete reversible airflow obstruction and normal breathing [11]. Auscultation using a listening instrument is only a qualitative diagnostic tool, even though it provides direct information [12]. However, the results of auscultation evaluation are inadequate due to several factors, such as inter and intra-observer inconsistency, bias errors in distinguishing fine sound structures, and frequency reduction [13]. Lung sound-based diagnosis is accurate and free of subjectivity errors due to the application of computer-based automated approaches and developments in lung sound recording techniques [14]. Computer-based lung sound assessment allows for a comprehensive assessment of lung sound features through visual representations, recording evaluations, suppression of noise contaminants, and evaluation of changes in lung sound action [15]. The sounds generated by air passing through the tracheobronchial tree are sounds produced by the respiratory system [16].

A. Common Lungs Disease

Asthma is a chronic respiratory condition characterized by inflammation and narrowing of the airways, leading to recurring episodes of wheezing, shortness of breath, chest tightness, and coughing [17]. It affects people of all ages but often starts in childhood and can persist into adulthood[18]. New opportunities for the early identification and categorization of lung diseases related to asthma are created by the combination of deep learning algorithms and heart sound analysis [19]. These algorithms can identify intricate patterns in cardiac sound data, which makes it possible to create precise and effective diagnostic models [20].

The respiratory ailment known as bronchitis is typified by discomfort in the bronchial tubes, the airways that supply oxygen to the lungs [21]. Two primary types of bronchitis can be distinguished: acute and chronic. Pneumonia is a dangerous respiratory system infection that affects the lungs [22]. Numerous pathogens, such as bacteria, viruses, fungi, or parasites, cause it. All ages are susceptible to pneumonia, but

young children, aged people, and those with compromised immune systems are most at risk.

The symptoms of Chronic Obstructive Pulmonary Disease (COPD) frequently include wheezes and reduced breath sounds, which indicate airway narrowing and blockage. These auditory characteristics are analyzed by deep learning models to consistently identify COPD trends in IoMT-enabled, noninvasive detection. These devices capture precise auditory cues, allowing for early and accurate COPD monitoring and classification, facilitating proactive treatment and intervention.

Crackles in lung sounds, a sign of respiratory disorders like pneumonia or bronchitis, are detected using IoMT-enabled, noninvasive lung disease detection. Deep learning models analyze these acoustic patterns, providing accurate classification and automated assessment of potential lung abnormalities. Wheezes, high-pitched sounds during exhalation, indicate airway obstructions in conditions like asthma or COPD. IoMT-enabled lung disease detection uses deep learning algorithms to improve diagnostic accuracy, distinguishing between obstructive and restrictive respiratory disorders and facilitating timely medical interventions.

B. Research Objectives and Motivation

1) Develop a noninvasive system for detecting and classifying lung diseases using lung sound data collected via IoMT (Internet of Medical Things) devices.

2) Implement deep learning techniques, including an Extra Tree Classifier, an AdaBoost Classifier, and a Deep Neural Network, to classify lung disease from sound data.

3) Analyze the effectiveness of enabling monitoring, correct diagnosis, and efficient data processing for remote healthcare applications.

Motivation: Respiratory health is safety-critical for human life, as effective diagnosis and treatment are essential to prevent severe consequences. Traditional methods are invasive and inaccessible, so there is a need for advancements in this domain. This work includes developing a noninvasive IoMT-enabled using deep learning for accurate and accessible lung disease detection, ensuring diagnosis, and supporting remote healthcare solutions.

II. Literature Review

Neural network model, lowering data leakage and memory utilization. CNN-LSTM layers, self-attention layers, dropout, Fully Linked (FC), and softmax layers comprise the model using the ICBHI 2017 dataset. The purpose of hyperparameter tuning is to reduce training failure. Self-attention is an independent layer that works with LSTM and CNN models [23]. According to experimental data, the suggested CNN+LSTM+Selfattention model performs better overall in terms of accuracy score than the CNN+LSTM+Hybrid CNN+LSTM, CNN+LSTM+Simple Attention, and CNN+GRU+Selfattention models. With a score of 57.02% for the initial train-test split, the model produces more dependable results.

A DNN is developed to diagnose interstitial lung diseases (ILD) in patients with connective tissue diseases (CTD), and preprocessing methods are evaluated on various lung sound data sets. The DNN offers remarkable accuracy on high-resolution CT scans, with an F1-score and an F2-score of 97% [24]. Since screening for ILD in patients with chronic autoimmune disorders is still a work in progress, this technique serves as an enabler for the early, safe, accurate, and affordable identification of CTD-ILD.

Augmentation techniques to resolve the imbalanced dataset problem. The model, which has two LSTM layers, five convolutional blocks, and no augmentation, achieves a remarkable F1 score of 0.9887 in 91 s per training epoch. Misclassifications accounted for just 3.05% of COVID-19 data and mostly happened in typical instances [25]. While the standard class showed recall and an F1 score, the pneumonia class showed exceptional precision. Deep Residual Network (DRN) uses a fractional water cycle swarm optimizer (Fr-WCSO-based DRN) to identify lung disorders from respiratory sound waves. The Fr-WCSO is a novel design that combines the Water Cycle Algorithm and Competitive Swarm Optimizer with Fractional Calculus and Water Cycle Swarm Optimizer (WCSO). To reduce overfitting problems, the system preprocesses respiratory input sound signals, identifies relevant features, and augments data [26]. DRN training and feature selection are then carried out using the Fr-WCSO algorithm.

Hybrid Interpretable Strategies with Ensemble Techniques (HISET) for respiratory sound classification. The first approach uses a GSSR technique, and the second uses a novel Realm Revamping Sparse Representation Classification (RR-SRC) technique, the third uses Distance Metric dependent Variational Mode Decomposition (DM-VMD) with Extreme Learning Machine (ELM) classification process, the fourth uses Harris Hawks Optimization with Scaling Factor based Pliable Differential Evolution (SFPDE), and the fifth uses Gray Wolf Optimization based Support Vector Classification (GWO-SVC) and Grasshopper Optimization Algorithm (GOA) based Sparse Autoencoder for dimensionality reduction techniques [27]. The ICBHI dataset is used to analyze the results, and the best results are obtained for the 2-class classification when Manhattan distance-based VMD-ELM is used. This method reported an accuracy of 95.39% for the 3-class classification, 90.61% for the 3-class classification, and 89.27% for the 4-class classification. The classification of pulmonary sounds obtained from patients with connective tissue illnesses using deep learning techniques [28].

Features such as Wavelet Entropy (WE) and wavelet packet energy (WPE) are extracted from the LS. Various classifiers, including Support Vector Machine (SVM), Decision Tree (DT), k-nearest Neighbor (KNN), and Discriminant Analysis (DA), are employed to classify healthy, COPD, and asthma cases using WE and WPE features. The proposed algorithm achieves a notable classification accuracy of 99.3% with the Decision Tree (DT) classifier, effectively distinguishing between healthy individuals and those with asthma or COPD based on LS[29]. Future work will validate this algorithm with real-time LS data from asthmatic and COPD patients.

Section I introduces the research work, focusing on using IoMT and deep learning for noninvasive lung disease diagnosis. Section II reviews the existing literature,

highlighting the limitations of conventional methods and the potential of IoMT technologies. Section III discusses the background studies, describing foundational principles and relevant advancements in IoMT and lung sound analysis. Section IV outlines the materials and methods, including data acquisition and deep learning techniques. Section V presents the results and discussions, analyzing findings. Section VI concludes the study, summarizing contributions, implications, and recommendations for future work.

III. Background Studies

A. Internet of Medical Things

The Internet of Medical Things (IoMT) refers to a network of interconnected medical devices, software applications, and healthcare systems designed to collect, transmit, and analyze patient data in real time. IoMT advanced technologies, i.e., sensors, wearable devices, remote monitoring tools, and cloud computing, to provide continuous healthcare solutions. These systems enable personalized patient care, early disease detection, and effective chronic disease management by continuously tracking vital signs and other health parameters.

IoMT enhances patient outcomes by enabling remote consultations, reducing hospital visits, and facilitating proactive treatment through real-time alerts. It also streamlines healthcare workflows by integrating data from diverse sources, improving clinical decision-making. However, IoMT faces challenges, including data security, interoperability, and compliance with regulatory standards. Despite these hurdles, IoMT represents a transformative advancement in modern healthcare, driving a shift toward precision medicine and empowering patients to actively engage in health management.

B. Extra Tree Classifier

One method of group decision-tree education is called the Extra Trees Classifier. When splitting a tree node, the Extra Trees classifier strongly randomizes the choice of features and reduces points, producing an unpruned collection of decision trees and trees. Extra trees function by combining the output of several de-correlated decision trees into a forest, from which they derive the classification result determined by the bulk of the voting technique. Compared to conventional decision trees or even random forests, this model adds more randomness by constructing multiple decision trees using arbitrary portions of characteristics and splitting at random points. This volatility reduces overfitting and improves the model's ability to generalize. Features taken from the audio recordings, such as time-domain, frequency-domain, and time-frequency domain characteristics, are fed into the Extra Tree classifier in the context of lung sound analysis. The model uses the rich and varied feature set to differentiate between various lung sounds, including wheezes, crackles, and other pathological sounds connected to illnesses like pneumonia, COPD, or asthma, as well as typical breathing.

C. AdaBoost Classifier

The AdaBoost methodology is a method for improving the performance of a model by combining weak classifiers. It involves extracting relevant features from audio recordings, such as time-domain and frequency-domain features, and training a weak classifier, typically a decision tree with a single split. Classifier kj can express an opinion, denoted by kj(xi) when their proposal is considered a training example for classifier acquisition for a given input model xi. Taking into account the issue of splitting the learning vector gathering into two classes, kj(xi) only accepts two values, such as 0 or 1, respectively, as shown in equation Sign C(xi), the sign of the linear mixture of the weighted total of the sub-classifiers' opinions, determines the classifier K's ultimate decision.

$$C(x_i) = a_1 k_1(x_i) + a_2 k_2(x_i) + a_l k_l(x_i)$$

Where weights are denoted bya_1, a_2,..., a_l And sub-classifiers by k_1, k_2,..., k_l. To generate a set of subpar learners, the adaboost technique keeps track of weights across instruction data and continually adjusts them after each weak learning cycle. The weights of training instances that the current weak learner incorrectly classifies will be increased, while the weights of training instances that are correctly classified will be decreased.

D. Deep Neural Network

The components of deep learning networks developed in the two-stage model will be clarified in the following part on artificial neural networks. Artificial Neural Networks (ANNs) are dynamic models that can adapt their internal architectures to meet specific functional requirements, making them ideal for managing nonlinear type issues. The essential parts of an ANN are the links and nodes that make up it, each with an output and input for interaction with other nodes or the surroundings. Each neuron in the network applies an activation function to introduce non-linearity through weighted connections. A labeled dataset is used to train the network, which uses a loss function to minimize the error between predicted and actual class labels during training by adjusting the weights of the connections through backpropagation. The learning process is one of the core characteristics of ANN, as they can understand the connections that define the data by adapting its connection to the information structure that makes up its surroundings. Neurons can be arranged in any topological configuration based on the kind and volume of input data. The feed towards construction is used in designing the most widely used ANN, with an input layer typically consisting of a particular amount of neurons paired. The data is sent to the secret layer or layers operating within the ANN and the output layer are created specifically to address the issue and provide the solution. Each neuron in the layer below is linked to every other neuron, with a fixed number of inputs and weights. Measurements are crucial for operating the deep neural network as they can be learned parameters.

$$C(x_i) = a_1 k_1(x_i) + a_2 k_2(x_i) + \cdots + a_l k_l(x_i)$$

Weight values are randomly initialized to be near zero but not zero before acquiring starts. The values of the data are modified to new information during learning, and this modification will aid in determining the significance of inputs. The activation function translates the weighted average from one neuron into the afterwards neuron's stimulation. Numerous mechanisms for activation are described in this research. Two factors influence the selection of rectified linear activation units in hidden layers in this work: (1) their ease of computation; and (2) the possibility of deep neural network

optimization because of their linear behaviour. After receiving input from hidden layer #2, the network's output layer transforms it into a binary (zero = unhealthy or one = healthy). The following equation is a representation of the sigmoid activation function:

$$\check{y} = \frac{1}{1 + e^{-z\prime}}$$

Where ў the neuron's results and z is the hidden layer #2 outputs.The average error was determined for each sample using the loss function with cross entropy. Here is a representation of the cross-entropy loss function in equation.

$$H\ (y, \check{y}) = -\sum_{i=1}^{n} y_i - \mathrm{Log}(\check{y}_i)$$

Where ў the network's output and y is is the real value. After every single propagation forward, the neural network searches for a set of heavy objects that minimizes the variance among the expected and actual values.

IV. Material and Methods

A. Dataset Descriptions

The audio files used in the thesis came from three distinct data sets. A variety of numbers of audio files from different datasets are included to create a balanced dataset. Table I Describes three Datasets used in this work.

1) Respiratory sound: The Respiratory Sound is a collection of 920 audio recordings from two research teams in different countries. Samples gathered at the Hospital Infante D. Pedro in Aveiro, Portugal, and the ESSUA Respiratory Research and Rehabilitation Laboratory by the School of Health Sciences, University of Aveiro research team. The second research team, from the Universities of Coimbra and Aristotle University of Thessaloniki, gathered respiratory sounds at the Papanikolaou General Hospital in Thessaloniki and the General Hospital of Imathia in Greece. Most of the database consists of audio, with samples from two hospitals in Portugal and Greece. The researchers analyzed the recordings using various instruments, including stethoscopes and microphones. They found 761 recordings suitable for evaluation, and 761 audio files were added to the model dataset without additional requirements [30]. The database includes 6898 breathing cycles from 126 patients, with 1864 having crackles, 886 having wheezes, and 506 having both.

2) Coronahack respiratory sound: The Coronahack Respiratory Sound includes respiration sound files from both COVID-19-affected and non-affected users. The file Corona-Hack-Respiratory-Sound-Metadata.csv includes additional disturbances and demographic data about the user. Audio recordings of patients with asthma or pneumonia were included in the dataset. Respiratory sound recordings from people with asthma or pneumonia were carefully selected from records that did not indicate probable COVID-19. The dataset contained these recordings [31].

3) Coswara sound: Coswara aims to develop a cost-effective method for diagnosing COVID-19 using speech, cough, and breath sounds. The study focuses on respiratory distress, a common symptom of the illness, and measures disease biomarkers in the acoustics of these noises. The project collects voice samples from healthy and sick individuals, examining nine categories of breathing, coughing, vowel phonation, and counting. Age, gender, location, current health status, and co-morbidities are collected. The dataset includes audio recordings from patients who have not contracted COVID-19 or recovered, tagged as "Asthma" or "Pneumonia." Thirty-eight healthy voice files with respiratory sound for at least 10 seconds were included in the dataset. The study includes 38 healthy, 58 asthmatic, and nine pneumonia voice recordings under specific conditions. The Coswara dataset is valuable for understanding and diagnosing COVID-19[32].

B. Preprocessing

Preprocessing is essential for IoMT-enabled noninvasive lung disease classification and detection. Fig. 1 displays the proposed methodology for this work Preprocessing procedures are necessary to prepare the data for successful categorization since lung sound recordings frequently contain noise and unpredictability due to patient movements and natural influences. Special values in proportional variables are crucial for maintaining data integrity and optimizing model performance in lung disease detection and classification from lung sound analysis. 'Nan' values, which occur when numerator and denominator are zero, are removed in the first preprocessing step. Normalization or scaling techniques can be used to handle these special cases. When the denominator is zero, special values arise, such as positive infinity (+∞) or negative infinity (-∞), as shown in Eq. (1). These extreme values can significantly impact deep learning model performance during training.

$$X = \begin{cases} \frac{N}{D} & if\ D \neq 0 \\ +\infty & \text{if } N > 0 \text{ and } D = 0 \\ -\infty & \text{if } N < 0 \text{ and } D = 0 \end{cases} \quad (1)$$

TABLE I. Three Datasets used in this Work

Dataset Name	Audio Samples	Patients	Key Features	Source
Respiratory Sound	920 recordings, 761 used	126 patients, 6898 breathing cycles (1864 crackles, 886 wheezes, 506 both)	Breathing sounds recorded using stethoscopes and microphones	[30]
Coronahack Respiratory Sound	Varied, includes both COVID-19 and non-COVID-19 patients	Includes asthmatic, pneumonia, and COVID-19-negative patients	Demographic data, respiratory conditions, and sound disturbances included	[31]
Coswara Sound	38 healthy, 58 asthmatic, 9 pneumonia	Data collected includes age, gender, health status, and co-morbidities	Focus on speech, cough, and breath sounds for diagnosing respiratory distress	[32]

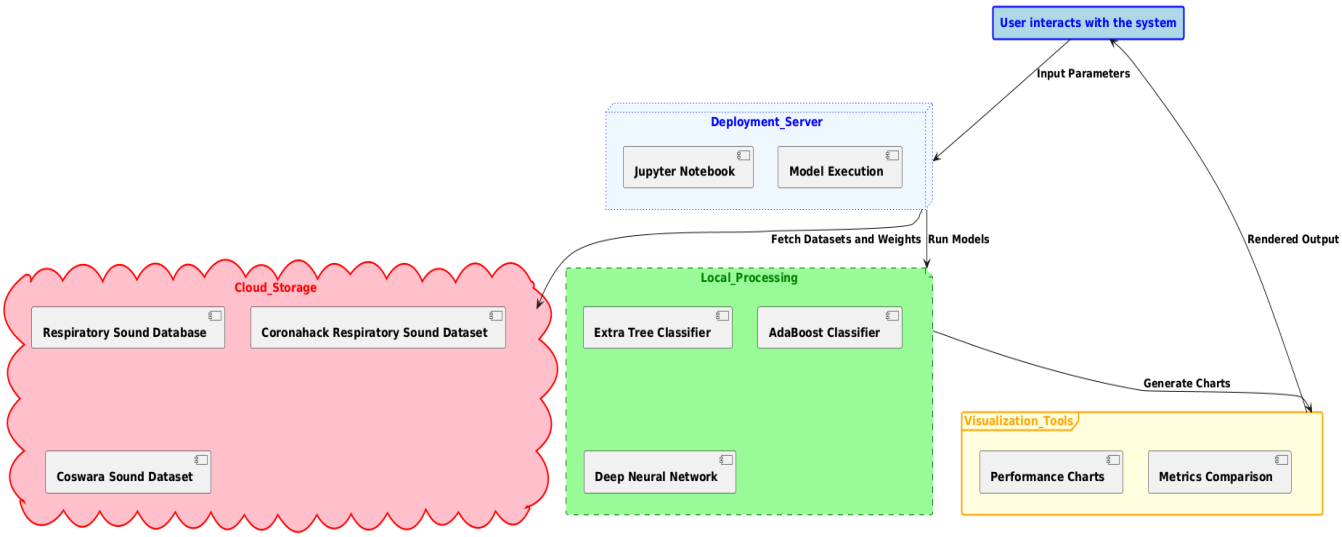


Fig. 1. Proposed methodology for this work.

The Weight of Evidence (woe) measures how well an organizing technique can discriminate between positive and negative results, or between 1 and 0. This method works for any issue where the binary variable is the target, even though it was initially created to create a predictive model for assessing credit default risk in the finance and credit sectors. The amount of evidence that either confirms or disproves a hypothesis is measured by its weight. The Weight of Evidence is determined as follows in Eq. (2):

$$WoE = \ln\left(\frac{Distribution\ of\ 1's}{Distribution\ of\ 0's}\right) * 100 \qquad (2)$$

Feature elimination is crucial for optimizing model performance by removing unnecessary, redundant, or noisy features from the dataset using statistical testing, correlation analysis, and machine learning algorithms. Feature elimination techniques such as Recursive Feature Elimination (RFE) remove irrelevant or redundant features, reducing overfitting and computational load. The model's performance metric is calculated without feature removal, and the impact of eliminating features is assessed. The model's performance metric J is calculated without feature x_j, and the impact of elimination x_j is shown in Eq. (3).

$$J_{-x_j} = J\left(X\left\{x_j\right\}\right) \qquad (3)$$

C. Feature Extraction

An individual measurable functionality or characteristic that defines a phenomenon is called a feature in machine learning. Useful algorithmic methods for classification rely on selecting independent, making distinctions, and useful characteristics. In this work, 1522 features from different categories were created for each sound recording as follows:

1) Time domain features: Three distinct groups are created from the signal's time series features: the audio recording's 0–1 s, 0–6 s, and 0–10 s segments. The relevant signal's increasing average series and accumulative series were calculated. The number of data points in the initial collection always equals the total moving average's term count, as shown in Eq. (4). C_k Is defined recursively as follows, where x1, x2,..., xn are the related respiration sound time series and C1, C2,..., Cn are the accumulated average with weights a series.

$$C_k = \frac{(x_k + (k-1) * x_{k-1})}{k} \qquad (4)$$

2) Spectral feature: It is essential to differentiate between typical and abnormal respiratory conditions by capturing the frequency domain characteristics of lung sounds through features.

a) Time-frequency spectrogram statistical features: Mel-spectrogram, MFCC, Short-Term Fourier Transform, and Chroma In the earlier part, each respiration sound was subjected to a Short Term Fourier Transform to transform it into a time-frequency spectroscopy and extract features akin to those found in a spirometer. After computations, a long list of factors is created, which includes the statistical properties of the time-frequency spectra obtained for the 0–1, 0–6, and 0–10 s periods of each audio recording.

b) Power spectrogram statistical features: Power spectrogram statistical features provide insights into power distribution across different frequencies over time, which is essential for identifying and distinguishing various respiratory conditions, is described in Eq. (5).

$$f(x) = c_o + c_1 * x \qquad (5)$$

Where c_1 Is the coefficient of the corresponding column and c_o is the constant term.

D. Feature Elimination Process

Feature elimination is crucial for optimizing model performance by removing unnecessary, redundant, or noisy features from the dataset using statistical testing, correlation analysis, and machine learning algorithms.

1) GINI Elimination: The Receiver Operating Characteristic (ROC) curve is a crucial tool in signal detection

theory, used alongside the Neyman-Pearson method to visualize a classifier's efficacy. It is used for assessing and comparing the overall efficacy of testing or diagnostic procedures. The AUC index, a summary of the ROC curve, is often used in this assessment. The process is summarized as follows:

a) The training and test sets of the data set are split 80/20.

b) The single-variate regression procedure was applied to each variable's training set to determine the AUC of each variable.

c) Using the AUC values as a guide, the Gini coefficient for each variable was determined using the formula below Eq. (6):

$$GINI = (2 * AUC - 1) * 100 \quad (6)$$

E. Machine Learning Models

This section includes the details of two-stage machine/deep learning models and the working principles of applied algorithms. The Extra Tree Classifier and Ada Boost Classifier techniques with the most effective binary categorization were chosen as modeling algorithms using Python's open-source "pycaret" library.

Weight values are randomly initialized to be near zero but not zero before acquiring starts. The values of the data are modified to new information during learning, and this modification will aid in determining the significance of inputs. The activation function translates the weighted average from one neuron into the afterwards neuron's stimulation. Numerous mechanisms for activation are described in this research. Two factors influence the selection of rectified linear activation units in hidden layers in this work: (1) their ease of computation; and (2) the possibility of deep neural network optimization because of their linear behaviour. After receiving input from hidden layer #2, the network's output layer transforms it into a binary (zero = unhealthy or one = healthy). The following Eq. (7) is a representation of the sigmoid activation function:

$$\check{y} = \frac{1}{1+ e^{-z\prime}} \quad (7)$$

Where ŷ the neuron's results and z is the hidden layer #2 outputs. The average error was determined for each sample using the loss function with cross entropy. Here is a representation of the cross-entropy loss function in Eq. (8):

$$H\ (y, \check{y}) = -\sum_{i=1}^{n} y_i - \mathrm{Log}(\check{y}_i) \quad (8)$$

Where ŷ the network's output and y is the real value. After every single propagation forward, the neural network searches for a set of heavy objects that minimizes the variance between the expected and actual values.

Dropout is a regularization technique for deep neural networks that helps lessen learning when nerve cells are interconnected. It suggests that a subset of randomly chosen neurons from a particular layer may be removed during learning. As a result, during a specific forward or backward pass, the results of the eliminated nerve cells disappear. In the present study, every iteration saw the removal of 0.1% of the neurons in the relevant layer from the input and hidden layers.

F. Evaluation Measures

We utilize various assessment measures to evaluate the effectiveness of the proposed models. These measures provide insight into the models' accuracy, predictive power, and generalization capability. The percentage of accurately categorized cases out of all instances is known as accuracy, as shown in Eq. (9).

$$\text{Accuracy} = \frac{Number\ of\ correct\ prediction}{Total\ numer\ of\ prediction} * 100 \quad (9)$$

Other evaluation measures adopted for assessing the proposed models are specificity, sensitivity, and F1 score, shown in Eq. (10) and Eq. (11). The specificity and sensitivity formula are as follows:

$$Specificity = \frac{True\ positive}{True\ positive+False\ negative} \quad (10)$$

$$Sensitivity = \frac{True\ negative}{True\ negative+False\ positive} \quad (11)$$

True positives (TP) are positive in the test set and correctly labeled as positive by the classifier. True negatives (TN) are negative in the test set and correctly labeled as negative by the classifier. False positives (FP) are negative in the test set but incorrectly labeled as positive by the classifier. False negatives (FN) are positive in the test set but incorrectly labeled as negative by the classifier. Eq. (12) shows that the F1 score is the harmonic mean of precision and recall, providing a combined measure of precision and recall.

$$F1\ score = 2 * \frac{Precision*Recall}{Precision\ +Recall} \quad (12)$$

Where precision and recall are calculated Eq. (13) and Eq. (14), respectively.

$$Precision = \frac{True\ positive}{True\ positive+False\ positive} \quad (13)$$

$$Recall = \frac{True\ positive}{True\ positive+False\ negative} \quad (14)$$

V. Discussion and Results

Spectrogram (top) and onset strength analysis (bottom) of lung sound data, critical for IoMT-enabled noninvasive lung disease detection. The spectrogram visualizes frequency (Hz) over time, with color intensity indicating sound energy. It highlights distinct acoustic patterns associated with respiratory cycles, facilitating feature extraction for disease classification. The onset strength graph below shows temporal variations in sound intensity, with detected onsets marked by red dashed lines, capturing significant events like wheezes or crackles. These features, analyzed using deep learning, improve the precision of lung sound classification, aiding in early and accurate detection of pulmonary diseases. Fig. 2 represents Spectro-Temporal Analysis for IoMT-Based Lung Disease Detection.

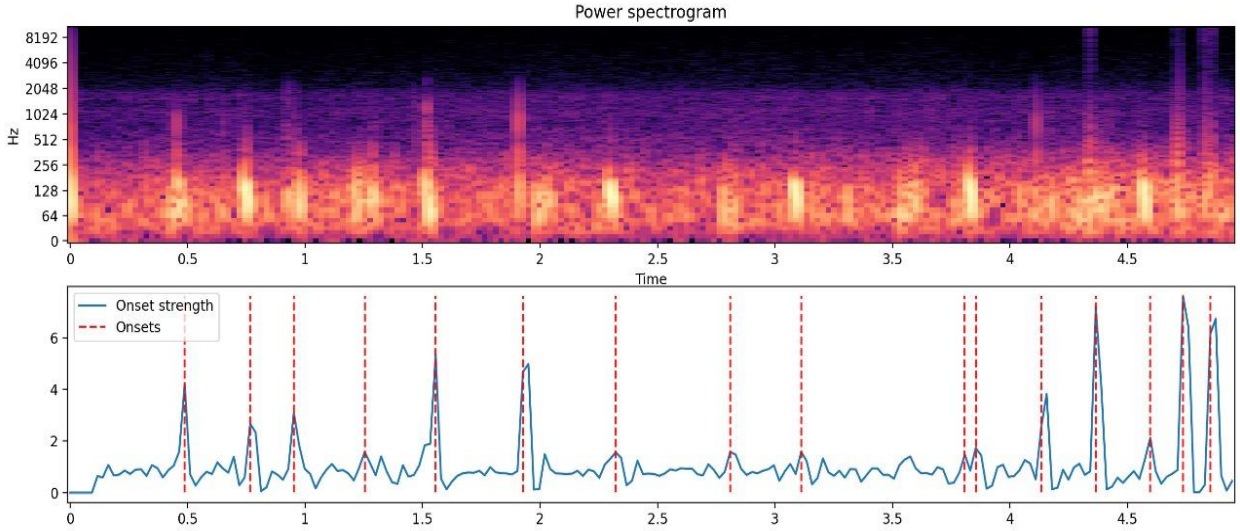


Fig. 2. Spectro-Temporal analysis for IoMT-Based lung disease detection.

Onset detection and energy analysis for lung sounds, crucial for IoMT-based noninvasive respiratory disease classification. The top graph visualizes onset strength with raw onsets (blue peaks) and backtracked onsets (red vertical lines). These onsets correspond to significant acoustic events, such as wheezes or crackles, indicative of lung abnormalities. By combining onset detection and energy analysis, these features enhance the ability of deep learning algorithms to accurately classify lung diseases, enabling early diagnosis through efficient feature extraction and temporal event mapping. This dual-layer analysis improves robustness in detecting subtle patterns in lung sounds, aiding real-time and remote healthcare applications. Fig. 3 shows Onset and Energy Analysis for Lung Sound Classification.

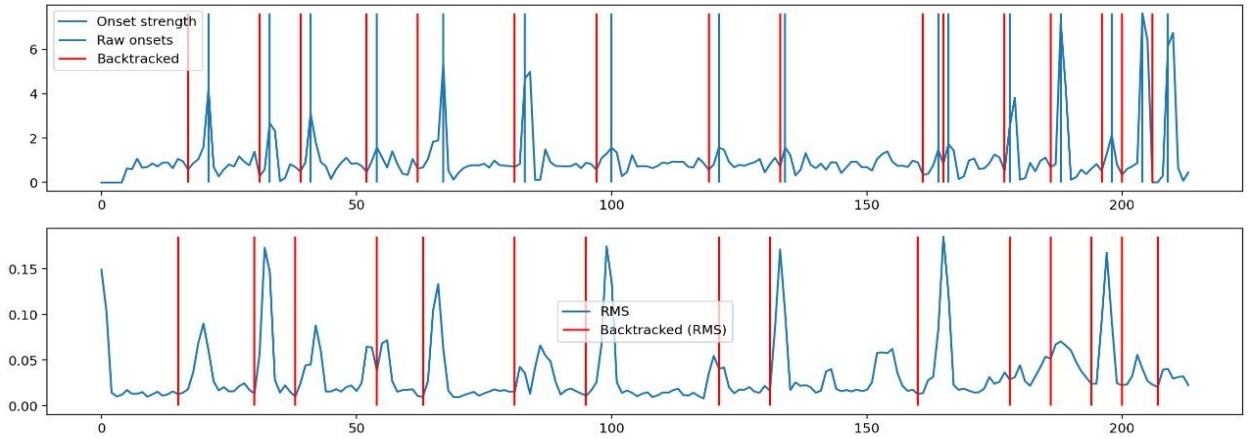


Fig. 3. Onset and energy analysis for lung sound classification.

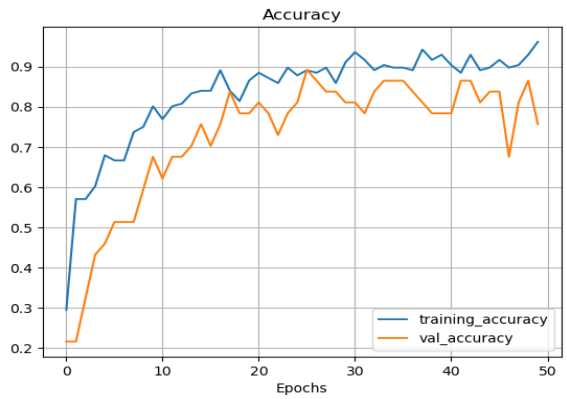


Fig. 4. Training history.

The Fig. 4 shows the training and validation accuracy trends over 50 epochs in a machine learning model. The x-axis represents epochs, and the y-axis represents accuracy values ranging from 0 to 1. The training accuracy (blue line) steadily improves, indicating the model's ability to fit the training data, with slight fluctuations towards the later epochs. The validation accuracy (orange line) increases initially, stabilizes, and exhibits minor oscillations, reflecting the model's generalization performance. The gap between training and validation accuracy suggests potential overfitting, as training accuracy surpasses validation accuracy in later epochs. This trend presents the need for optimization or regularization techniques.

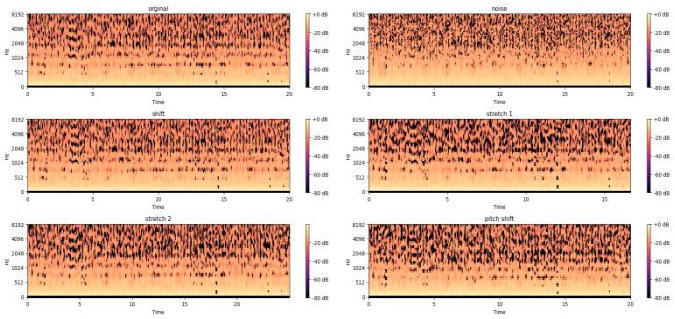


Fig. 5. Unaltered audio spectrogram.

The Fig. 5 show spectrograms visualizing feature extraction from an audio signal under different augmentations. The first plot (original) shows the unaltered spectrogram. Subsequent plots depict the effects of augmentations: noise addition, time shift, time stretching (two variations), and pitch shifting. These transformations simulate variability in audio datasets to improve machine learning model generalizability. The model can robustly learn features under varying conditions by augmenting the original signal, critical in tasks like speech recognition or environmental sound classification.

The results presents analysis of three machine learning models—AdaBoost Classifier, Extra Tree Classifier, and Deep Neural Network—evaluated on three datasets: the Respiratory Sound, Coronahack Respiratory Sound, and Coswara Sound. Each model's performance is measured using precision, recall, F1-score, and overall accuracy across diseases like crackles, wheezes, COVID-19, asthma, pneumonia, and healthy cases.

Table II, III and IV shows the performance of the models. Accuracy ranged from 94.12% to 95.23%, with consistent precision, recall, and F1-scores for all diseases, indicating robust yet moderate effectiveness. Table II displays the AdaBoost Classifier's assessment, which improved accuracy (95.42%–96.33%) and balanced precision-recall for detecting COVID-19 and other diseases, suggesting its superior predictive reliability compared to the ensemble approach. Table III evaluates a Deep Neural Network, showcasing the highest performance metrics, with accuracy surpassing 98% across all datasets. The network's precision, recall, and F1 scores consistently reached 0.99 for most diseases, demonstrating its efficacy in detecting subtle respiratory anomalies.

TABLE II. PERFORMANCE MEASURES OF EXTRA TREE CLASSIFIER

Dataset used for experiments	Diseases	Precision	Recall	F1-score	Overall Accuracy
Respiratory Sound	Crackles	0.95	0.96	0.95	**94.12%**
	Wheezes	0.93	0.94	0.96	
Coronahack respiratory sound	COVID-19	0.94	0.95	0.93	**95.23%**
	Healthy	0.96	0.93	0.92	
Coswara sound	Asthma	0.92	0.94	0.94	**94.21%**
	Pneumonia	0.94	0.93	0.94	
	Healthy	0.92	0.94	0.94	

TABLE III. PERFORMANCE MEASURES OF ADA BOOST CLASSIFIER

Dataset used for experiments	Diseases	Precision	Recall	F1-score	Accuracy
Respiratory Sound	Crackles	0.96	0.94	0.95	95.42%
	Wheezes	0.94	0.96	0.95	
Coronahack respiratory sound	COVID-19	0.97	0.96	0.97	96.33%
	Healthy	0.95	0.96	0.96	
Coswara sound	Asthma	0.96	0.95	0.94	94.76%
	Pneumonia	0.97	0.96	0.95	
	Healthy	0.95	0.95	0.96	

TABLE IV. PERFORMANCE MEASURES OF DEEP NEURAL NETWORK

Dataset used for experiments	Diseases	Precision	Recall	F1-score	Accuracy
Respiratory Sound	Crackles	0.99	0.98	0.98	98.92%
	Wheezes	0.99	0.99	0.99	
Coronahack respiratory sound	COVID-19	0.98	0.99	0.99	99.33%
	Healthy	0.99	0.99	0.99	
Coswara sound	Asthma	0.99	0.99	0.99	99.36%
	Pneumonia	0.98	0.99	0.99	
	Healthy	0.97	0.98	0.99	

VI. Conclusion and Future Work

This work presents an IoMT-based noninvasive approach to lungs disease detection and classification. The work uses an IoMT-enabled, noninvasive approach for lung disease detection and classification using Respiratory Sound, Coronahack Respiratory Sound, and Coswara Sound. Using machine learning models such as the Extra Trees classifier and AdaBoost classifier alongside a proposed deep learning model, this approach achieved impressive accuracy levels across various respiratory conditions. . The DNN achieves accuracy across all datasets, with 98.92% for the Respiratory Sound, 99.33% for the Coronahack Respiratory Sound, and 99.36% for the Coswara Sound. These results highlight the potential of deep learning models to support reliable and accurate respiratory health assessment in IoMT applications. Future work will enhance model robustness to handle diverse datasets and real-world variations and optimize the model for low-power IoMT devices to facilitate clinical deployment. Future work will optimize the proposed model for real-world applications, explore additional features such as multi-modal data for improved accuracy, and conduct large-scale evaluations across diverse network environments to assess generalizability.

Acknowledgment

The author would like to acknowledge the support of The Islamia University of Bahawalpur Pakistan, for providing the resources and guidance during the Ph.D. journey.

References

[1] Altan, G., Kutlu, Y., & Gökçen, A. (2020). Chronic obstructive pulmonary disease severity analysis using deep learning on multi-channel lung sounds. *Turkish Journal of Electrical Engineering and Computer Sciences, 28*(5), 2979-2996.

[2] Altan, G., Kutlu, Y., & Allahverdi, N. (2019). Deep learning on computerized analysis of chronic obstructive pulmonary disease. *IEEE Journal of Biomedical and Health Informatics, 24*(5), 1344-1350.

[3] Korenbaum, V., & Shiryaev, A. (2020). Features of sound conduction in human lungs in the 80–1000 Hz and 10–19 KHz frequency ranges. *Acoustical Physics, 66*, 548-558.

[4] Lee, S. H., Kim, Y.-S., Yeo, M.-K., Mahmood, M., Zavanelli, N., Chung, C., ... & Yeo, W.-H. (2022). Fully portable continuous real-time auscultation with a soft wearable stethoscope designed for automated disease diagnosis. *Science Advances, 8*(21), eabo5867.

[5] Tariq, Z., Shah, S. K., & Lee, Y. (2019). Lung disease classification using deep convolutional neural network. Paper presented at the *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*.

[6] Au, Y. K., Muqeem, T., Fauveau, V. J., Cardenas, J. A., Geris, B. S., Hassen, G. W., & Glass, M. (2022). Continuous monitoring versus intermittent auscultation of wheezes in patients presenting with acute respiratory distress. *The Journal of Emergency Medicine, 63*(4), 582-591.

[7] Tong, L., Huang, S., Zheng, C., Zhang, Y., & Chen, Z. (2022). Refractory Mycoplasma pneumoniae pneumonia in children: Early recognition and management. *Journal of Clinical Medicine, 11*(10), 2824.

[8] Mukhopadhyay, S., Mehrad, M., Dammert, P., Arrossi, A. V., Sarda, R., Brenner, D. S., ... & Ghobrial, M. (2020). Lung biopsy findings in severe pulmonary illness associated with E-cigarette use (vaping): A report of eight cases. *American Journal of Clinical Pathology, 153*(1), 30-39.

[9] Rajkumar, S., Sathesh, K., & Goyal, N. K. (2020). Neural network-based design and evaluation of performance metrics using adaptive line enhancer with adaptive algorithms for auscultation analysis. *Neural Computing and Applications, 32*, 15131-15153.

[10] Sathesh, K., Rajkumar, S., & Goyal, N. K. (2020). Least Mean Square (LMS)-based neural design and metric evaluation for auscultation signal separation. *Biomedical Signal Processing and Control, 59*, 101784.

[11] Sanchez-Perez, J. A., Berkebile, J. A., Nevius, B. N., Ozmen, G. C., Nichols, C. J., Ganti, V. G., ... & Wright, D. W. (2022). A wearable multimodal sensing system for tracking changes in pulmonary fluid status, lung sounds, and respiratory markers. *Sensors, 22*(3), 1130.

[12] Ferreira, H. d. M. G. (2021). Pulmonary auscultation using mobile devices: Feasibility study in respiratory diseases. *Universidade do Porto (Portugal)*.

[13] Razvadauskas, H., Vaičiukynas, E., Buškus, K., Arlauskas, L., Nowaczyk, S., Sadauskas, S., & Naudžiūnas, A. (2024). Exploring classical machine learning for identification of pathological lung auscultations. *Computers in Biology and Medicine, 168*, 107784.

[14] Sfayyih, A. H., Sulaiman, N., & Sabry, A. H. (2023). A review on lung disease recognition by acoustic signal analysis with deep learning networks. *Journal of Big Data, 10*(1), 101.

[15] Hsu, F.-S., Huang, S.-R., Huang, C.-W., Huang, C.-J., Cheng, Y.-R., Chen, C.-C., ... & Lai, Y.-C. (2021). Benchmarking of eight recurrent neural network variants for breath phase and adventitious sound detection on a self-developed open-access lung sound database—HF_Lung_V1. *PLOS One, 16*(7), e0254134.

[16] Waddingham, P. H., Mangual, J. O., Orini, M., Badie, N., Muthumala, A., Sporton, S., & Chow, A. W. (2023). Electrocardiographic imaging demonstrates electrical synchrony improvement by dynamic atrioventricular delays in patients with left bundle branch block and preserved atrioventricular conduction. *Europace, 25*(2), 536-545.

[17] Koefoed, H. J. L., Zwitserloot, A. M., Vonk, J. M., & Koppelman, G. H. (2021). Asthma, bronchial hyperresponsiveness, allergy, and lung function development until early adulthood: A systematic literature review. *Pediatric Allergy and Immunology, 32*(6), 1238-1254.

[18] Chang, A. B., Oppenheimer, J. J., Irwin, R. S., Adams, T. M., Altman, K. W., Azoulay, E., ... & Boulet, L.-P. (2020). Managing chronic cough as a symptom in children and management algorithms: CHEST guideline and expert panel report. *Chest, 158*(1), 303-329.

[19] Srivastava, A., Jain, S., Miranda, R., Patil, S., Pandya, S., & Kotecha, K. (2021). Deep learning-based respiratory sound analysis for detection of chronic obstructive pulmonary disease. *PeerJ Computer Science, 7*, e369.

[20] Kumar, N., & Kumar, D. (2021). Machine learning-based heart disease diagnosis using noninvasive methods: A review. Paper presented at the *Journal of Physics: Conference Series*.

[21] Deshmukh, R., Bandyopadhyay, N., Abed, S. N., Bandopadhyay, S., Pal, Y., & Deb, P. K. (2020). Strategies for pulmonary delivery of drugs. In *Drug Delivery Systems* (pp. 85-129): Elsevier.

[22] Oliva, J., & Terrier, O. (2021). Viral and bacterial co-infections in the lungs: Dangerous liaisons. *Viruses, 13*(9), 1725.

[23] Bhushan, P., Fahad, M. S., Agrawal, S., Kamesh, K. S. D., Tripathi, P., Mishra, P., ... Deepak, A. (2024). A Self-Attention Based Hybrid CNN-LSTM Architecture for Respiratory Sound Classification. *GMSARN International Journal, 18*(1), 54-61.

[24] Fava, A., Dianat, B., Bertacchini, A., Manfredi, A., Sebastiani, M., Modena, M., & Pancaldi, F. (2024). Preprocessing techniques to enhance the classification of lung sounds based on deep learning. *Biomedical Signal Processing and Control, 92*, 106009.

[25] Fachrel, J., Pravitasari, A. A., Yulita, I. N., Ardhisasmita, M. N., & Indrayatna, F. (2023). Enhancing an imbalanced lung disease x-ray image classification with the CNN-LSTM model. *Applied Sciences, 13*(14), 8227.

[26] Dar, J. A., Srivastava, K. K., & Mishra, A. (2023). Lung anomaly detection from respiratory sound database (sound signals). *Computers in Biology and Medicine, 164*, 107311.

[27] Prabhakar, S. K., & Won, D.-O. (2023). HISET: Hybrid interpretable strategies with ensemble techniques for respiratory sound classification. *Heliyon, 9*(8).

[28] Dianat, B., La Torraca, P., Manfredi, A., Cassone, G., Vacchi, C., Sebastiani, M., & Pancaldi, F. (2023). Classification of pulmonary sounds through deep learning for the diagnosis of interstitial lung diseases secondary to connective tissue diseases. *Computers in Biology and Medicine, 160*, 106928.

[29] Haider, N. S., & Behera, A. (2022). Computerized lung sound based classification of asthma and chronic obstructive pulmonary disease (COPD). *Biocybernetics and Biomedical Engineering, 42*(1), 42-59.

[30] Vbookshelf. (2017). *Respiratory sound database* [Dataset]. Kaggle. Retrievedfrom https://www.kaggle.com/datasets/vbookshelf/respiratory-sound-database

[31] Praveengovi. (2020). *Coronahack respiratory sound dataset* [Dataset]. Kaggle.Retrievedfrom https://www.kaggle.com/datasets/praveengovi/coronahack-respiratory-sound-dataset\

[32] Janashreeananthan. (2021). *Coswara* [Dataset]. Kaggle. Retrieved from https://www.kaggle.com/datasets/janashreeananthan/coswara

Readmission Risk Prediction After Total Hip Arthroplasty Using Machine Learning and Hyperparameter Optimized with Bayesian Optimization

Intan Yuniar Purbasari[1], Athanasius Priharyoto Bayuseno[2], R. Rizal Isnanto[3], Tri Indah Winarni[4]
Doctoral Program of Information System, School of Postgraduate Studies, Diponegoro University, Semarang, Indonesia[1]
Department of Informatics-Faculty of Computer Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya, Indonesia[1]
Department of Mechanical Engineering-Faculty of Engineering, Diponegoro University, Semarang, Indonesia[2]
Department of Computer Engineering-Faculty of Engineering, Diponegoro University, Semarang, Indonesia[3]
Department of Anatomy-Faculty of Medicine, Diponegoro University, Semarang, Indonesia[4]
UNDIP Biomechanics Engineering & Research Centre (UBM-ERC), Universitas Diponegoro, Semarang, Indonesia[4]

***Abstract*—Machine learning techniques are increasingly used in orthopaedic surgery to assess risks such as length of stay, complications, infections, and mortality, offering an alternative to traditional methods. However, model performance varies depending on private institutional data, and optimizing hyperparameters for better predictions remains a challenge. This study incorporates automatic hyperparameter tuning to improve readmission prediction in orthopaedics using a public medical dataset. Bayesian Optimization was applied to optimize hyperparameters for seven machine learning algorithms—Extreme Gradient Boosting, Stochastic Gradient Boosting, Random Forest, Support Vector Machine, Decision Tree, Neural Network, and Elastic-net Penalized Logistic Regression—predicting readmission risk after Total Hip Arthroplasty (THA). Data from the MIMIC-IV database, including 1,153 THA patients, was used. Model performance was evaluated using Precision, Recall, and AUC-ROC, comparing optimized algorithms to those without hyperparameter tuning from previous studies. The optimized Extreme Gradient Boosting algorithm achieved the highest AUC-ROC of 0.996, while other models also showed improved accuracy, precision, and recall. This research successfully developed and validated optimized machine learning models using Bayesian Optimization, enhancing readmission prediction following THA based on patient demographics and preoperative diagnosis. The results demonstrate superior performance compared to prior studies that either lacked hyperparameter optimization or relied on exhaustive search methods.**

***Keywords*—Total hip arthroplasty; orthopaedic surgery; Bayesian Optimization; machine learning algorithm; hyperparameter optimization**

I. Introduction

Total Hip Arthroplasty (THA) is a surgery to replace the entire hip joint with an artificial joint, called an implant, and is typically performed on patients with severe osteoarthritis and having a prevalent rate of 1.2% [1], [2], [3]. THA is an effective option for reducing pain and improving hip joint function in patients with severe symptomatic osteoarthritis, with a success rate of more than 85% and clinical outcomes lasting for 15 to 25 years [4], [5], [6], [7]. The demand for THA procedures has significantly increased and is expected to continue to rise until 2040 [8], [9], [10]. However, like all surgery procedures, THA comes with risks concerning a patient's condition for infection, dislocation, and mortality and may require unplanned hospital readmission for a corrective procedure or even a revision surgery. Dislocation, complication (both implant and non-implant-associated complication), including infection and mental status, are identified as the reasons for readmission after THA [11], [12]. Meanwhile, 30-day readmission rate of 5% to 9.5% and 5% to 10% for 90-day readmission have also been reported [13], [14], [15], [16], [17], [18].

Bundled payments in healthcare have been established in several countries, such as United States, The Netherlands, Sweden, United Kingdom, and Indonesia, as part of a transition toward value-based treatment. This payment structure makes healthcare providers receive incentives for coordinating treatments, preventing complications and failures, and reducing unnecessary or duplicate tests and treatments, including unplanned hospital readmission [19], [20], [21], [22]. Therefore, a readmission risk prediction system for THA surgery is valuable in not only improving cost management, but also in improving patient care quality. It serves as a tool in the cooperative decision-making process between patients and doctors regarding the final decision to undertake the elective THA surgery, as well as in providing pre- and post-operative treatment tailored to the patient's condition [23], [24], [25]. Based on the predicted risks, in which is very individuals for each patient, doctors may design a customized treatment procedure such as medication prescription, diets, and/or supervised exercise before and after the surgery. The readmission risk prediction system is also a source of information for the hospital to manage the distribution of its resources, such as specialty doctors, operating rooms, inpatient

rooms, and medical equipment [26], [27]. With the rise of medical expenses, implementing effective and efficient procedures across all channels puts hospital management to the test while maintaining the quality of care, as often mirrored by the concept of patient-centered care [28].

Both classical and machine learning-based methods have already applied to predict the risk, including readmission risk, of THA surgery. Classical methods to predict the risk of THA surgery, as well as general surgery, are currently used, such as the Risk Assessment and Predictor Tool (RAPT) [29], Charlson Comorbidity Index (CCI) [30], and Elixhauser Comorbidity Index (ECI) [31], [32]. CCI and ECI measure risks associated with mortality, hospital length of stay, and hospital charges based on the overall severity of comorbidities. Instead, RAPT predicts discharge destination and length of stay for patients after hip or knee arthroplasty. With the rising complexity of patient electronic medical history data, a solution based on artificial intelligence (AI) and machine learning (ML) is critical to understanding each unique health profile of a patient and providing more accurate information related to risks to doctors and patients.

Machine learning (ML) algorithms have been effectively introduced and used in a variety of medical sectors such as predicting length of stay of lung cancer patients [33], elderly patients readmission risk prediction [34], and heart disease classification [35]; nonetheless, its technology is relatively new in the field of orthopaedic surgery [36]. Although classical approaches to calculate complication and mortality risks using patient demographic characteristics and comorbidities do exist, they are considered having weak discrimination, are not verified, and not general enough to accurately predict risk preoperatively and mostly are developed using multivariable regressions techniques [37]. According to the authors' preliminary literature review [38], various ML algorithms have been utilized to predict THA outcomes using AI/ML technology. Among them are Logistic Regression [39], [40], [41], Linear Support Vector Machine [39], [42], Linear Discriminant Analysis [39], Elastic Net Penalized Linear Regression [40], [43], Random Forest [40], [42], [43], Neural Networks [42], [43], Stochastic Gradient Boosting [42], [43], [44], and an automated machine learning tool developed for prognosis (dubbed AutoPrognosis) [45]. Specifically for predicting readmission risk, some machine learning tools used were a Random Forest approach with Natural Language Processing (NLP) feature [46], Linear Regression, Support Vector Machine, Random Forest to predict 90-day readmission [47], Logistic Regression [48], and Light Gradient Boosting Machine [49].

Research by Kuo et al. [50] aimed to predict periprosthetic joint infection by implementing two levels of ML architecture. Naïve-Bayes, eXtreme Gradient Boosting, Linear Regression, and Random Forest were put in the first level as base classifiers, while Support Vector Machine as the meta-classifier acted in the second level as the final decision maker based on the prediction result of the first level. Some patient-related variables used were demographic, biomedical, comorbidity, surgery-related variables and had a very good Area under Curve (AUC) score of 0.988. There was no additional information on hyperparameters fixed for the final classification algorithms in each level.

Research by Klemt [36] investigated the risk of failure of a revision THA involving similar variables to [50] using Neural Network, Elastic-net penalized logistic regression, and Random Forest and achieved the highest AUC score of 0.85 for Neural Network.

The investigation to predict complication and irregular surgery duration was performed in study [44] using eXtreme Gradient Boosting algorithm which incorporated 12 (twelve) variables related to patient demographic, implant type, and surgeon experience. The AUC scores were 0.64 for complication prediction and 0.89 for surgery duration. Based on the supplied codes in GitHub, hyperparameters search was performed using GridSearch method.

Nham et al. [39] conducted a study exploring various tree-based algorithms, neural networks, statistical and probabilistic approaches, and Support Vector Machine (SVM) to predict in-patient mortality, discharge disposition, and length of stay. The study incorporated patient-specific variables, such as demographic, diagnosis, and mortality risk, and also situational variables, such as operation location, hospital type, and number of hospital beds. The best three algorithms were different for each predicted outcome, but Linear Support Vector Machine (LSVM) and Decision List (DL) were the most frequent to appear in the best three. Again, there was no discussion on how to select the best hyperparameters for each algorithm.

A machine learning prediction of all-cause complications within two years of primary THA was introduced by Kunze [42] using preoperative variables, such as demographics, comorbidities, known allergies, and Modified Harris Hip Score. The study investigated several machine learning algorithms such as stochastic gradient boosting, random forest, support vector machine, neural network, and Elastic-net penalized logistic regression, including their hyperparameter settings which were tuned through a grid search approach. The best performing algorithm was Elastic-net penalized logistic regression with an AUC value of 0.93 on test set.

Research by Shah et al. [37] is one of the studies that applied Bayesian Optimization to optimize hyperparameters in an Automated Machine Learning (AutoML) environment. It consists of an ensemble of several machine learning pipelines, each fitted with its own hyperparameters, to make prediction on complications after THA procedure. It achieved AUC value of 0.732 and claimed to exceed the performance of other single-type machine learning algorithms such as Logistic Regression, XGBoost, Gradient Boosting, AdaBoost, and Random Forest.

Specifically for predicting readmission risk, research by Slezak et al. used and compared multiple machine learning techniques such as Light Gradient Boosting Machine (LGBM), Logistic Regression (LR), eXtreme Gradient Boosting (XGB), and Random Forests (RF) to predict risk for readmission after Total Joint Replacement surgery [49]. They included American Society of Anaesthesiologists Physical Status Classification (ASA) and modified Frailty Index (mFI-5) as the contributing

factors and achieved an AUC value of 0.672 with LGBM being the best algorithm for readmission prediction.

Unplanned 90-day readmission after THA was also predicted with LR [48] and the predictive performance of the model was evaluated using the validation dataset, resulting in an AUC value of 0.715. All variables were grouped into five categories, including demographics, perioperative factors, past surgical histories, comorbidities, and medication usages, where contributing factors for each category were identified.

Another prediction of 90-day readmission after TJA (Total Joint Arthroplasty) using comprehensive data from electronic medical records and patient-reported outcome measures was investigated in study [47] by utilizing a variety of machine learning algorithms including LR, LR with LASSO (Least Absolute Shrinkage and Selection Operator) penalty, SVM, and RF. They achieved quite an impressive AUC score of 0.862 with LR-LASSO being the best performing algorithm and selected nine significant variables including diabetes, prior use of corticosteroid medication, and number of follow-up visits to orthopaedic clinics as the significant risk factors.

Most of the machine learning algorithms applied require fine-tuning on a number of the hyperparameters (i.e., the pre-set algorithm parameters to control the learning process and determine the values of model parameters) to achieve the best result (i.e., highest accuracy and precision, lowest error, or highest discriminant value). Finding the best hyperparameters are a manual trial-and-error process, though there has been an increase in using automatic search algorithms (i.e., Hyperparameter Optimization) such as Grid Search, Random Search, Evolutionary-based Search, and Bayesian Optimization. Only a few earlier studies on predicting THA outcomes using machine learning have included automatic hyperparameter tuning with Grid Search [42] and the Bayesian Optimization methodology [37].

In addition, risk prediction of THA patients in earlier studies were based on data from respective authors' affiliations, thus might not be suitable for benchmarking the algorithms' performance and is difficult in terms of reproducibility to verify the results [51]. To the author's knowledge, no previous research related to predicting outcomes following Total Hip Arthroplasty surgery has used public datasets.

Machine learning techniques have proven to be effective in creating models to predict post-operative risks in THA patients, albeit there is no definitive technique best suited for all outcomes. Several previously cited articles briefly discussed how to determine the appropriate hyperparameters for each machine learning algorithm utilized using Grid Search approach, such as [42] and [44], which include systematically testing all possible hyperparameter combinations in the search space, but it is a time consuming and repetitive activity. With the rapidly increasing number and complexity of electronic medical records generated in this big data era, sweeping hyperparameters in the hyperparameter search space thoroughly as Grid Search does would take hours or days as the number of hyperparameters increases and finding the best ones remains a challenge. While Random Search provides better solution by randomly sampling points in the search domain, it is also less efficient for expensive evaluation functions [52]. Another more advanced method such as Bayesian Optimization capable of finding global optimization in a complex problem is gaining more recognition [37], [45]. Consequently, this study aims to provide a Bayesian Optimization approach to MLA in readmission risk predicting after THA operation. This research performed experiments on various machine learning algorithms commonly utilized in existing research with the optimization enhancement on a publicly available medical dataset MIMIC and evaluate the generalizability of the algorithms' claimed performance.

Hyperparameter optimization is particularly advantageous in improving the performance of machine learning algorithms and reducing tedious human effort to find the best setting [53]. It is also widely known that different hyperparameter settings work best for different datasets [54]. In the prediction of THA outcome context, one can expect a more accurate readmission risk prediction while delivering broader and easy-to-use machine learning tools to benefit non-technical expert users (i.e., doctors), incorporating hyperparameter optimization in the machine learning algorithms used. An automated hyperparameter optimization provides the search for the best hyperparameter configuration for each machine learning algorithm used automatically, thus improving accuracy and efficiency, as well as lowering error [55]. Meanwhile, automation makes it easier and simpler for non-technical experts to use various machine-learning technologies and gain more from the outcomes [53], [56]. The suggested system's contributions include the following:

1) Development of readmission risk prediction method: The study proposes to incorporate automatic hyperparameters tuning to the machine learning algorithm selected by users, providing a more accurate and precise prediction result while offering ease and simplicity in applying and customizing machine learning algorithms to non-technical expert users.

2) Exploration of hyperparameters optimization algorithm based on statistical approach: The study introduces a Bayesian Optimization method which is based on statistical calculation to find the best hyperparameters setting of the machine learning algorithm selected by users.

3) Performance evaluation on known public tabular dataset (Medical Information Mart for Intensive Care-IV) MIMIC-IV: To evaluate the generalizability of high performing machine learning algorithms acclaimed in previous studies, this study uses a known public medical dataset MIMIC-IV [57], which is based on real medical dataset.

4) Performance evaluation with robust metrics: to assess the effectiveness of the proposed system, the study introduces four evaluation metrics: Accuracy, Precision, Recall, and Area under the Curve Receiver Operating Characteristic (AUC-ROC) values. These metrics measure the algorithm's performance quality quantitatively.

This journal article is organized as follows: Section I introduces the background, objectives, and significance of the study; Section II details the methods, including data sources,

preprocessing, and machine learning techniques; Section III presents the results of the experiments; Section IV discusses the findings in comparison to existing research; and Section V concludes with key takeaways and future research directions.

II. METHODS

A. System Overview

The proposed readmission risk prediction system after THA operation furnishes decision support feature to both patients and healthcare professionals in the domain of peril prognosis. The system processes medical data from the electronic medical record dataset which integrates various relevant attributes from a number of tables. All related data from patients underwent THA procedure are extracted, including demographics, diagnosis, and postoperative procedures of medical rehabilitation, if any. Additionally, some data imputations of missing values from the attributes were performed and the readmission status of patients within a year after the THA procedure is set to be the target value of prediction. The system includes a varied array of classification algorithms in previous studies, including eXtreme Gradient Boosting (XGB), Stochastic Gradient Boosting (SGB), Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), Neural Network (NN), and Elastic-net Penalized Logistic Regression (EnPLR). The best hyperparameters of those algorithms are then selected using Bayesian Optimization (BO) technique to ensure maximum results of the classification algorithms in predicting the readmission risk. Evaluation metrics such as Precision, Recall, and Area under the Receiver Operating Characteristic (AUROC) curve are employed to measure the algorithms' performance quality. Fig. 1 explicates the system overview of the proposed readmission risk prediction system.

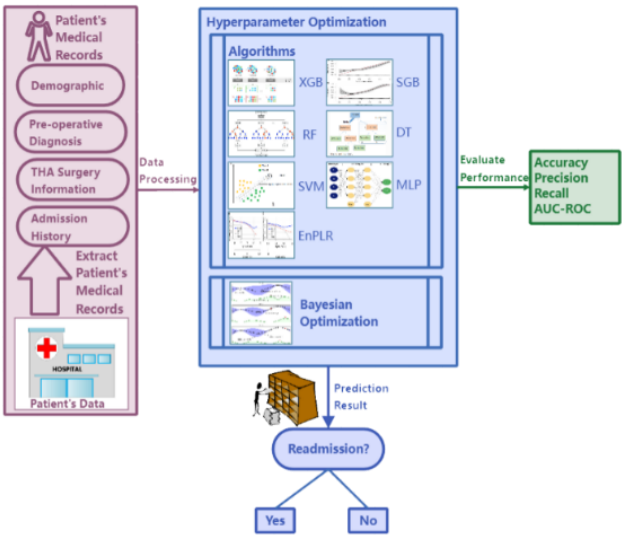


Fig. 1. System overview of optimized readmission risk prediction model.

The proposed method predicts the readmission risk of a THA patient based on one's demographics and preoperative diagnosis to help doctors plan preoperative treatments, aid patients in making informed decision on the surgery, and to assist hospitals managing their resources efficiently and properly with data-driven decision making. The system involved steps as follows:

1) Preprocess the patient dataset from MIMIC-IV (Medical Information Mart for Intensive Care version IV, which is a large de-identified public dataset of patients admitted to the emergency unit at the Beth Israel Deaconess Medical Center in Boston, MA). Duplicate data is identified and removed.

2) Identify the hyperparameters of previously utilized machine learning algorithms, including XGB, SGB, RF, SVM, DL, NN, and EnPLR, and perform the Bayesian Optimization technique to find the best hyperparameters for each algorithm.

3) Divide the dataset into 70% training and 30% testing set with stratified sampling.

4) Compare the model performance using evaluation metrics such as Precision, Recall, and AUROC.

B. Algorithm Introduction

1) eXtreme Gradient Boosting (XGB): XGB is a tree boosting machine learning algorithm whose capability of handling sparse data becomes one of its highlight features. The algorithm was introduced by Chen and Guestrin [58] and has won various machine learning competitions. It implements gradient boosting technique to perform additive optimization and incorporates a regularized model to prevent overfitting using L1 and L2 regularization methods. The algorithm is available in various popular languages such as Python, R, C, C++, and Java.

2) Stochastic Gradient Boosting (SGB): Introduced long before XGB, SGB has different approach to prevent overfitting in a general gradient boosting technique. It proposed a randomness into the algorithm by drawing a subsample, which is a fraction f of the size of the training set, to replace the full training data set at each iteration [59]. In scikit-learn, SGB is implemented with GradientBoostingClassifier by defining subsample hyperparameter < 1.

3) Random Forest (RF): RF works by building multiple decision trees and combines the output to give a single result. RF technique extends the bagging approach by combining bagging and feature randomness to generate an uncorrelated forest of decision trees [60]. Some of key benefits of RF are its capability to handle classification and regression tasks, and it is also easy to determine feature importance from the prediction result. This advantage can help to identify which features contribute the most to the prediction made.

4) Support Vector Machine (SVM): SVM has the ability to perform both linear and non-linear classification using a kernel trick. It works by constructing a hyperplane in a high dimension space and determine functional margin of data points (called support vectors) which define a separation boundary between classes [61].

5) Decision Tree (DT): In this study, DT is used as a replacement of Decision List (DL), since a Python's scikit-learn implementation of DL is not available. DT is a simpler

version of RF which consists of decision nodes, chance nodes, and end nodes. Following the path from root and making decisions at every branch according to the attribute value leading to an end node will create a rule that defines the temporal or causal relations among attributes.

6) Neural network: The Neural Network implemented in this study is Multi-Layer Perceptron (MLP), which is a supervised learning algorithm that maps a number of inputs to a number of outputs with a number of nodes in between (called hidden layers). The nodes in the hidden layers transform the values from the previous layer by adjusting the nodes' weights and propagate the result to the next layer. MLP implements a regularization technique (called L2) to avoid overfitting by punishing weights of large magnitudes.

7) Elastic-net Penalized Logistic Regression (EnPLR): EnPLR is a modified version of the classic Logistic Regression with a regularization method to overcome overfitting that linearly combines L1 (LASSO-Least Absolute Shrinkage and Selection Operator) and L2 (Ridge Regression). In scikit-learn, EnPLR is implemented with LogisticRegression model by setting hyperparameters solver=saga, penalty=elasticnet, and l1_ratio=0.5.

8) Hyperparameter optimization with Bayesian Optimization (BO): Hyperparameter optimization is a process to fine tuning the hyperparameter values of machine learning algorithms to obtain the best setting which yield the best result, in a classification or regression problem. The domains of hyperparameter may range from real values (i.e. learning rate), integer (i.e. number of layers), binary (i.e. with or without early stopping), to category (i.e. a selection of optimizer functions). Let $\mathcal{A}$ be a machine learning algorithm with N hyperparameters. The domain of n-th hyperparameter is denoted as Λ_n and the overall configuration hyperparameter space is denoted as $\Lambda = \Lambda_1 \, x \, \Lambda_2 \, x \, ... \, \Lambda_N$. A hyperparameter vector is denoted as $\lambda \in \Lambda$ and a machine learning algorithm with its instantiated hyperparameter λ is denoted as $\mathcal{A}_\lambda$. Given a dataset $\mathcal{D}$, the goal of hyperparameter optimization is to find Eq. (1) as in [53]:

$$\lambda^* = \underset{\lambda \in \mathcal{A}}{argmin} \, \mathbb{E}_{(D_{train}, D_{valid}) \sim \mathcal{D}} \boldsymbol{V}(\mathcal{L}, \mathcal{A}_\lambda, D_{train}, D_{valid}) \quad (1)$$

with $\boldsymbol{V}(\mathcal{L}, \mathcal{A}_\lambda, D_{train}, D_{valid})$ measures the loss value of the model generated by algorithm A with hyperparameter λ on the training data D_{train} and evaluated on validation data D_{valid}. $D \sim \mathcal{D}$ denoted the finite data which its expectation needs to be approximated.

Grid Search and Random Search are the two most basic and simple hyperparameter optimization methods. While Grid Search performs exhaustive search on the hyperparameter search space, Random Search [56] samples a hyperparameter configuration randomly until it satisfies a defined threshold value. On the other hand, Bayesian Optimization (BO) is a global optimization framework approach of a black box objective function with high complexity and undefined shape and offers better efficiency given limited evaluation budget [62], [63].

BO is used in hyperparameter optimization (HO) that constructs a probability model of an objective function and use it to select the most promising hyperparameters to be evaluated in the real objective function [64]. Utilizing a Bayesian approach, BO takes note of the result of the previous evaluation to build a probability model which maps hyperparameters onto a probability score of an objective function, *P(score|hyperparameters)*, called a surrogate function. Compared to the original function, the new function is easier to be evaluated and the Bayesian method works by finding the next hyperparameter having the best performance on the surrogate function to be evaluated on the real objective function. In a BO, a Gaussian Process (GP) is used as a surrogate function to produce a posterior distribution over the function values based on observed data.

Meanwhile, an acquisition function selects the next point for the surrogate function, with common options including probability of improvement, expected improvement, and upper confidence bounds. This study uses the Bayesian Optimization library *bayes-opt*, which defaults to upper confidence bounds to balance exploitation of high surrogate values and exploration of uncertain regions.

C. Datasets

The proposed study utilized public medical dataset from MIMIC-IV (Medical Information Mart for Intensive Care version IV) [62] from Beth Israel Deaconess Medical Centre in Boston, MA, United States. The dataset consists of 223,452 patients in its hospitalization hosp module and after a careful selection resulted in 1,153 unique THA patients' data. The target variable is readmission within one year of first THA procedure which is a Boolean value TRUE or FALSE.

A comprehensive description of characteristics of the patient data is provided in Table I. Most of the demographic-related variables are summarized as mean values, while the rest are written based on their respective numbers and percentage. The numbers in preoperative diagnosis variable do not sum up to the total number of patients as a patient may have more than one diagnosis.

D. Preprocessing

Data preprocessing involves identifying and removing duplicates, missing values, and erroneous information from the data. Proper techniques such as Label Encoding and One-Hot Encoding are applied to transform categorical data into a numerical format [63]. Additionally, a standard scaler is applied to ensure all attribute values are standardized. The overall preprocessing step is summarized in Fig. 2.

Data preprocessing involved extracting relevant data from the general dataset, identifying 1,153 unique THA patients with 55 variables (54 inputs and 1 output), as summarized in Table I. Patients had up to 39 preoperative diagnoses, totalling 1,881 distinct diagnoses coded in International Classification of Diseases ICD-9 and ICD-10.

In addition, a feature processing was applied. Missing values in weight, height, and Body Mass Index (BMI) were handled by filling with the mean value of the attribute column. All 1,881 diagnosis ICDs were assigned integer values from 1 to 1881 (a value of 9999 was given to empty cells), implant

materials variable in categorical type were assigned 1 to 7 (99 for empty cells), implant type in categorical were assigned values of 1 to 4 (9 for empty cells), and insurance type in categorical were assigned values of 1 to 3.

TABLE I. CHARACTERISTICS OF THA PATIENTS

Variables	All Patients (n=1,153)	
	Mean	(IQR)
Demographics		
Age (years)	65.66	16 (58-74)
Weight (pounds)	180.91	65.75 (146.25-212)
Height (feet)	43.67	67 (0-67)
Body Mass Index (kg/m^2)	28.98	8.05 (24.6-32.65)
Preoperative Blood pressure		
Systole	131.62	20 (120-140)
Diastole	74.88	15 (67-82)
	Number	**(%)**
Male Sex	506	43.88
Insurance type		
Medicare	505	43.80
Medicaid	35	3.03
Other	613	53.16
Preoperative diagnosis		
Osteoarthritis	1,023	8.9
Hypertension	579	5.04
Hyperlipidemia	440	3.83
Esophageal reflux	336	3.83
Nicotine/tobacco dependence	207	2.93
Anemia	338	2.94
Depressive disorder	111	0.97
Anxiety disorder	154	1.34
Sleep apnea	200	1.74
Diabetes	260	2.26
Asthma	155	1.35
Others	7,679	66.88
Implant material		
Synthetic	196	16.99
Metal on Polyethylene	85	7.37
Ceramic on Polyethylene	437	37.90
Ceramic	31	2.69
Ceramic on Ceramic	4	0.34
Metal	6	0.52
Oxidized Zirconium on Polyethylene	2	0.17
N/A	392	33.99
Implant type		
Cemented	36	3.12
Cementless	300	26.02
Not Specified	217	18.82
N/A	600	52.04
Readmission within one-year post-THA	237	20.55

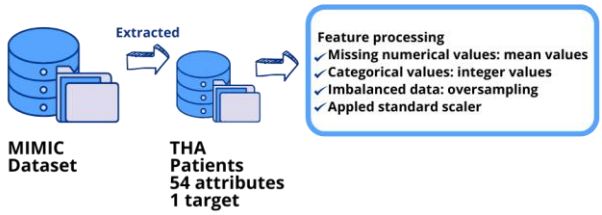


Fig. 2. Preprocessing steps.

Since the dataset is imbalanced, with 237 TRUE and 916 FALSE, a simple random oversampling method to duplicate examples in the minority class was applied using *imbalanced-learn* library [64]. The oversampling method produced equal size of data for each target class, which were 916 TRUE and 916 FALSE.

III. RESULTS

A. Algorithm Selection and Hyperparameter Optimization (HO)

The experiment was conducted in two modes for each algorithm: without and with hyperparameter optimization (HO). A limitation of the bayes-opt library is its restriction to numeric attributes, preventing the optimization of categorical and Boolean hyperparameters. Five numerical hyperparameters were optimized for each algorithm, except for Support Vector Machine and Elastic-net Penalized Logistic Regression, which had only three due to categorical constraints. Table II presents the optimized hyperparameters, their values, and a comparison of accuracy and AUROC between Bayesian Optimization (BO) and default hyperparameters.

The convergence plot in Fig. 3 shows that six of seven algorithms (except SVM) reached a steady value within six iterations, indicating BO has effectively estimated the optimization target. Even SVM's sharp increase in the 14th iteration [Fig. 3(d)] remained within the 95% Confidence Interval (CI).

The objective plot in Fig. 4 visualizes the distribution of hyperparameter values when the BO searched for the optimized ones. The diagonal subplots represent histograms of the sampled values for each parameter and aid in visualizing the distribution of parameter values used during optimization. This illustrates how BO investigated the parameter space. The red dashed line represents the optimized hyperparameter value while the blue line plots the AUC-ROC value as the objective target around the optimized hyperparameter value. If the majority of a parameter's samples are concentrated in a specific range, it may indicate that this range has values that are near ideal. For example, four straight lines in the diagonal subplots of XGB [Fig. 4(a)] indicates that all the sampled values of expGamma (an exponential transformation of Gamma), learning_rate, max_depth, and min_child_weight between their respective ranges yield a near optimal AUC-ROC value, i.e. 0.9955. Meanwhile, the off-diagonal contour subplots show pairwise interactions between two hyperparameters, where darker colors indicate better values. The contour identifies regions where the parameter combinations yield optimal results. For example, the off-diagonal subplot between expGamma and expC (the exponential transformation of C) in Fig. 4(c) indicates that the red star value ($expGamma = 0.584$, thus $Gamma = 10^{0.584} = 3.837$) in the dark region represents the optimal value of expGamma when combined with the value of expC. In Fig. 4(a), the red star in n_estimators plot (n_estimators=1818.9744) resides within the narrow darker area, indicating that the value is optimal or near optimal.

TABLE II. THE HYPERPARAMETERS OF EACH ALGORITHM AND THEIR OPTIMIZED VALUES

Algorithm	Hyperparameters	Bounds		optimized value	without HO		with HO	
		min	max		auc-roc	fit time (s)	auc-roc	elapsed time (s)
XGB	learning_rate	0.01	0.8	0.0352	0.9955	0.18	0.9958	582.11
	min_child_weight	0	5	1.5814				
	max_depth	1	50	50				
	gamma	1e-5	1.0	0.0731				
	n_estimators	5	5000	1818.9744				
SGB	learning_rate	0.01	0.8	0.4091	0.8790	0.72	0.9936	747.03
	n_estimators	10	250	151.6208				
	subsample	0.1	0.9	0.8513				
	min_sample_split	2	25	11.3614				
	max_depth	0	500	155.4370				
RF	max_depth	0	500	34.0807	0.9945	0.66	0.9936	541.20
	max_features	0.1	0.999	0.3961				
	max_leaf_nodes	0	5000	1556.2286				
	min_samples_split	2	25	6.0441				
	n_estimators	10	250	219.7180				
SVM	C	1e-6	1e+6	1230.18891	0.7655	0.25	0.9915	6060.32
	gamma	1e-4	1e+5	2.359238				
	tol	1e-9	0.1	0.1				
DT	max_depth	1	500	179.5508	0.8782	0.03	0.8783	4.75
	min_samples_split	2	25	10.5158				
	min_samples_leaf	1	5	3.8508				
	max_features	0.1	0.999	0.5504				
	max_leaf_nodes	2	1000	715.1166				
MLP	alpha	1e-6	10	0.0226	0.9142	3.12	0.9503	485.52
	learning_rate_init	1e-6	10	0.0012				
	batch_size	10	300	65.5406				
	max_iter	100	1000	801.9782				
	tol	1e-9	0.01	0.0003				
EnPLR	tol	1e-9	0.01	0.0000002	0.6592	1.67	0.6622	923.93
	C	0.001	1000	526.6542				
	max_iter	5000	10000	8923.7024				

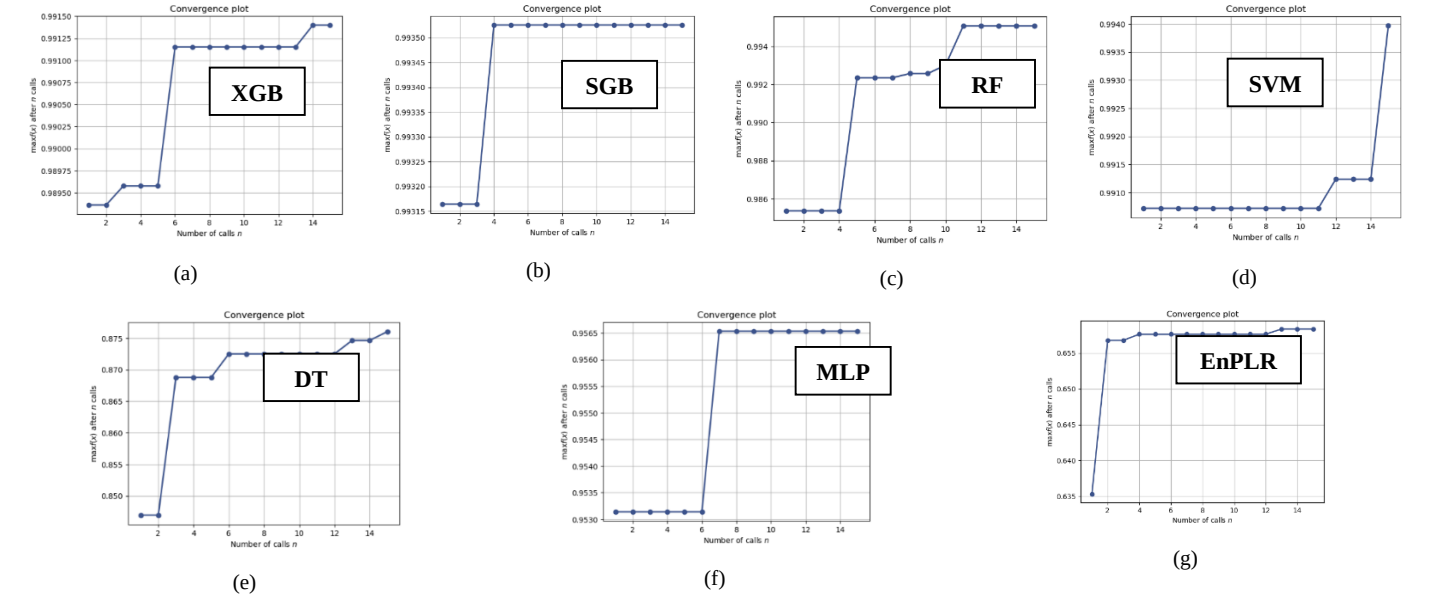
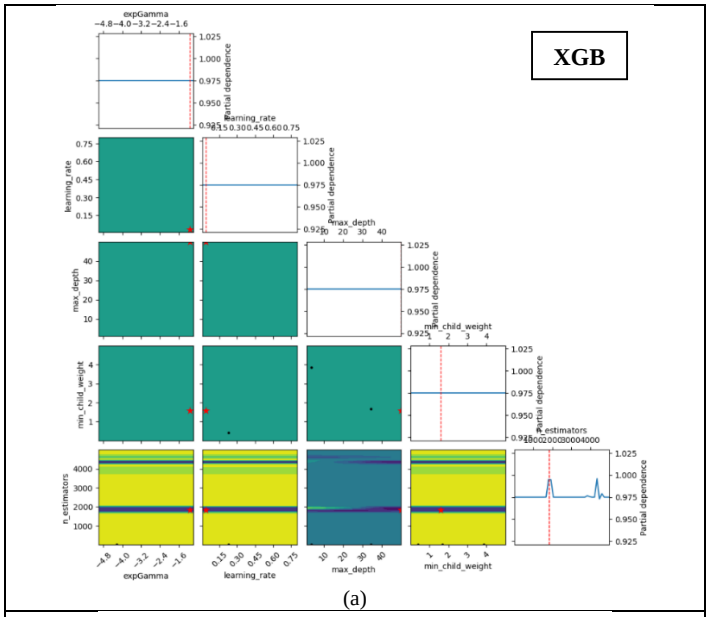
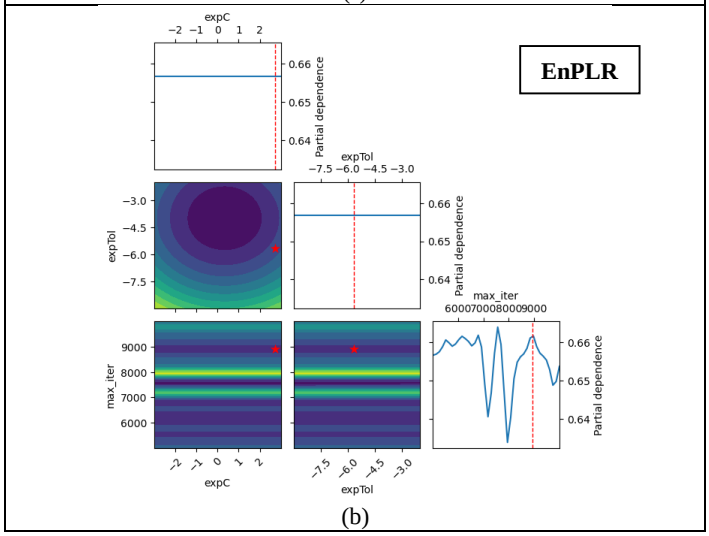


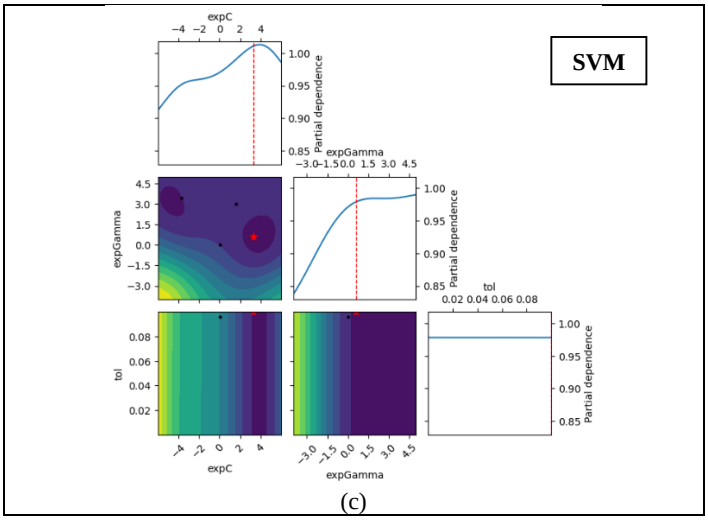
Fig. 3. The convergence plot in bayesian optimization technique for algorithm: (a) Extreme gradient boosting (XGB), (b) Stochastic gradient boosting (SGB), (c) Random forest (RF), (d) Support vector machine (SVM), (e) Decision tree (DT), (f) Multilayer perceptron (MLP), and (g) Elastic-net penalized linear regression (EnPLR).



(a)



(b)



(c)

Fig. 4. The objective plot of (a) XGB, (b) EnPLR, and (c) SVM. The red stars represent the best values of each hyperparameter.

B. *Performance Evaluation*

Table II shows the time required for each approach to complete optimization. All seven algorithms took longer with Bayesian Optimization (BO) than with default hyperparameters, with execution times increasing from 155 times longer (MLP) to 24,605 times longer (SVM). This added duration is expected in hyperparameter tuning, balancing computational cost with improved AUC-ROC performance.

Table III presents performance metrics: Accuracy, Precision, Recall, and AUC-ROC. SGB, SVM, and MLP showed improvements in all four metrics after optimization, while RF and DT performed worse. Optimization had minimal impact on XGB and EnPLR.

Additionally, MLP's default setting (max_iter = 200) failed to converge, requiring an increase to max_iter = 1000 in the optimized version. Similarly, EnPLR failed to converge in both cases, even after raising max_iter to 10,000 in the optimized version.

TABLE III. PERFORMANCE EVALUATION COMPARISON OF ALGORITHM

Alg.	Accuracy		Precision		Recall		AUROC	
	w/o HO	with HO	w/o HO	with HO	w/o HO	with HO	w/o HO	with HO
XGB	0.945	0.950	0.908	0.916	0.991	**0.991**	0.995	**0.996**
SGB↑	0.801	0.976	0.767	**0.971**	0.866	0.982	0.879	0.994
RF↓	0.978	0.967	0.965	0.947	0.991	0.989	0.994	0.994
SVM↑	0.693	**0.991**	0674	0.947	0.750	0.989	0.764	0.994
DT↓	0.878	0.820	0.814	0.780	0.985	0.894	0.878	0.878
MLP↑	0.862	0.900	0.813	0.843	0.942	0.986	0.914	0.950
EnPLR	0.630	0.632	0.623	0.624	0.668	0.672	0.660	0.662

↑: an increased value when using BO; ↓: a decreased value when using BO

IV. Discussion

Previous THA-related studies have employed machine learning techniques using institution-specific datasets, contributing to research advancements. However, many relied on Grid Search for hyperparameter tuning or default algorithm settings, limiting optimization efficiency. Only a few studies explored advanced methods such as Bayesian Optimization (BO) to enhance predictive performance, despite its potential for complex optimization problems.

This study applies BO across multiple machine learning algorithms commonly used in previous research to improve readmission prediction. Our results demonstrate that optimized hyperparameters significantly enhance AUC-ROC scores compared to default settings. While prior studies addressed different adverse event predictions, they shared a common challenge: imbalanced datasets, as adverse events are rare.

Comparing model performance, this study outperformed previous research across several algorithms. The Extreme Gradient Boosting (XGB) algorithm achieved the highest AUC-ROC score (0.996), surpassing [44], other cited works [36], [37], [39], [42], [50], and specifically [49] for predicting readmission risk. Similarly, the Multilayer Perceptron (MLP) achieved an AUC-ROC of 0.914, higher than the Neural Network in [36] (0.85). The Support Vector Machine (SVM) in this study achieved 0.994, far outperforming previous studies [39], which reported 0.74 (Length of Stay), 0.8 (Discharge), and 0.97 (Mortality). The Stochastic Gradient Boosting (SGB) algorithm improved to 0.994, higher than 0.88 in [42]. Random Forest (RF) achieved 0.994, surpassing 0.80 in [36], 0.91 in [42], and 0.83 in [47] for readmission prediction. However, Elastic-Net Penalized Logistic Regression (EnPLR) underperformed (0.662), falling below 0.732 in [37], 0.93 in [42], and 0.86 in [47]. These findings highlight the effectiveness of Bayesian Optimization in tuning machine learning models, addressing limitations in previous studies that relied on basic or exhaustive hyperparameter searches.

Additionally, the MIMIC-IV dataset used in this study provides a publicly available benchmark to validate models trained on private institutional datasets, which often restrict benchmarking and generalizability. SVM, which performed best in some studies [39], [50], showed weaker performance before optimization, while XGB, previously considered suboptimal [39], [44], demonstrated superior results. This suggests that models trained on private datasets may not always generalize well to broader patient populations.

Furthermore, Table III indicates that Recall scores are consistently higher than Accuracy and Precision, suggesting that models prioritize identifying positive cases, reducing false negatives. This is critical in medical applications, where minimizing false negatives (missed readmission cases) is more important than false positives, which can be addressed with further evaluation [65], [66], [67].

The discrepancy in comparative results between datasets may stem from differences in patient populations, clinical settings, and data preprocessing. The MIMIC-IV dataset includes a more diverse patient cohort from multiple institutions, whereas private datasets from previous studies may be more homogeneous. Additionally, the inclusion/exclusion criteria and feature selection may introduce variability in model performance.

The varying performance of models across datasets suggests that some algorithms are more sensitive to dataset structure. For instance, tree-based models (XGB, RF, SGB) performed consistently well, likely due to their robustness in handling structured tabular data. SVM, which performed best in some previous studies, showed inferior performance before optimization in our dataset, suggesting that hyperparameter tuning plays a crucial role in improving its adaptability to different data distributions. Our results show that Bayesian Optimization significantly improves performance consistency across datasets. Without proper hyperparameter tuning, models such as SVM and MLP may underperform in certain datasets due to suboptimal parameter selection. This highlights the importance of adaptive optimization techniques in ensuring machine learning models generalize well across different clinical datasets.

Despite its contributions, this study has several limitations. First, as a retrospective analysis relying on past medical records and billed ICD codes, it is susceptible to errors such as miscoding or missing values, which could impact machine learning predictions. Second, the bayes-opt library used for hyperparameter tuning is still under active development, and its latest version (1.5.1 as of July 10, 2024) does not yet support categorical and Boolean hyperparameters. As a result, key parameters such as SVM kernels, MLP activation functions, and RF criteria could not be optimized, potentially limiting model performance. Lastly, the dataset exhibited class imbalance, with significantly fewer TRUE cases than FALSE cases. While this study employed oversampling to address the issue, more advanced techniques could further enhance predictive accuracy. Future research should explore alternative resampling strategies and hyperparameter tuning methods to refine model performance.

V. Conclusion

This study successfully developed, trained, tested, and validated Bayesian Optimization for hyperparameter tuning across seven machine learning algorithms to predict readmission risk following Total Hip Arthroplasty (THA). By comparing these optimized models to previous studies that either lacked hyperparameter tuning or employed different optimization methods, the results demonstrated that Bayesian Optimization significantly enhances predictive performance. This underscores the critical role of hyperparameter tuning in maximizing model accuracy, improving decision-making, and increasing confidence in integrating machine learning into THA patient management.

Additionally, this study highlights the importance of external evaluation using a publicly available dataset, ensuring that machine learning models trained on institution-specific data can generalize effectively across different patient populations. The findings suggest that standardized evaluation is essential for ensuring model robustness and reliability in clinical applications.

While this study addressed key challenges in predictive modeling for THA readmission, limitations remain, particularly regarding class imbalance and the scope of readmission prediction. Future research may explore advanced resampling techniques to further improve model performance and investigate more specific predictive outcomes, such as infection-related readmission, revision THA, or Length of Stay.

ACKNOWLEDGMENT

This research is supported by School of Postgraduate Studies-Diponegoro University and University of Pembangunan Nasional Veteran Jawa Timur.

ETHICAL APPROVAL

This study used the MIMIC-IV database (Medical Information Mart for Intensive Care, version 2.0), a publicly available and de-identified dataset, which complies with the Health Insurance Portability and Accountability Act (HIPAA) standards. Access to the dataset was granted after completing the Collaborative Institutional Training Initiative (CITI) Program training and acceptance of the Data Use Agreement (DUA). Ethical approval for the original data collection was obtained by the Institutional Review Board (IRB) of Beth Israel Deaconess Medical Center (BIDMC), and the requirement for individual patient consent was waived. Therefore, no additional ethical approval was required for this study.

REFERENCES

[1] Australian Orthopaedic Association, "Australian Orthopaedic Association National Joint Replacement Registry," 2022. [Online]. Available: https://aoanjrr.sahmri.com/annual-reports-2022].

[2] H.-H. Bleß Miriam Kip Eds, "White Paper on Joint Replacement," Berlin, 2018. doi: https://doi.org/10.1007/978-3-662-55918-5.

[3] R. Brittain et al., "NJR statistical analysis, support and associated services," 2021. [Online]. Available: www.njrcentre.org.uk

[4] J. T. Evans, J. P. Evans, R. W. Walker, A. W. Blom, M. R. Whitehouse, and A. Sayers, "How long does a hip replacement last? A systematic review and meta-analysis of case series and national registry reports with more than 15 years of follow-up," 2019. [Online]. Available: www.thelancet.com

[5] X. Y. Mei, Y. J. Gong, O. Safir, A. Gross, and P. Kuzyk, "Long-term outcomes of total hip arthroplasty in patients younger than 55 years: A systematic review of the contemporary literature," 2019, Canadian Medical Association. doi: 10.1503/cjs.013118.

[6] I. P. Sitorus and I. H. Dilogo, "Functional and Radiological Outcome of Revision Total Hip Arthroplasty in the Indonesian National Referral Hospital," Hip Knee J, vol. 3, no. 1, pp. 2723–7826, 2022, doi: 10.46355/hipknee.v3i1.161.

[7] M. Spalević et al., "TOTAL HIP REPLACEMENT REHABILITATION: RESULTS AND DILEMMAS," Acta Medica Medianae, vol. 57, no. 1, pp. 48–53, Mar. 2018, doi: 10.5633/amm.2018.0108.

[8] S. M. Kurtz, K. L. Ong, E. Lau, and K. J. Bozic, "Impact of the Economic Downturn on Total Joint Replacement Demand in the United States: Updated Projections to 2021," JBJS, vol. 96, no. 8, 2014, [Online]. Available: https://journals.lww.com/jbjsjournal/Fulltext/2014/04160/Impact_of_the_Economic_Downturn_on_Total_Joint.2.aspx

[9] J. A. Singh, S. Yu, L. Chen, and J. D. Cleveland, "Rates of Total Joint Replacement in the United States: Future Projections to 2020–2040 Using the National Inpatient Sample," J Rheumatol, vol. 46, no. 9, p. 1134, Sep. 2019, doi: 10.3899/jrheum.170990.

[10] M. Sloan, A. Premkumar, and N. P. Sheth, "Projected Volume of Primary Total Joint Arthroplasty in the U.S., 2014 to 2030," JBJS, vol. 100, no. 17, 2018, [Online]. Available: https://journals.lww.com/jbjsjournal/fulltext/2018/09050/projected_volume_of_primary_total_joint.3.aspx

[11] M. A. Smolle et al., "Readmissions at Thirty-Days and One-Year for Implant-Associated Complications following Primary Total Hip and Knee Arthroplasty: A Population-Based Study of 34,392 Patients Across Austria," J Arthroplasty, Aug. 2024, doi: 10.1016/j.arth.2024.08.027.

[12] R. C. Clement et al., "Risk factors, causes, and the economic implications of unplanned readmissions following total hip arthroplasty," Journal of Arthroplasty, vol. 28, no. 8 SUPPL, pp. 7–10, 2013, doi: 10.1016/j.arth.2013.04.055.

[13] R. M. Greiwe, J. M. Spanyer, J. R. Nolan, R. N. Rodgers, M. A. Hill, and R. G. Harm, "Improving Orthopedic Patient Outcomes: A Model to Predict 30-Day and 90-Day Readmission Rates Following Total Joint Arthroplasty," J Arthroplasty, vol. 34, no. 11, pp. 2544–2548, Nov. 2019, doi: 10.1016/j.arth.2019.05.051.

[14] S. Zhao, J. Kendall, A. J. Johnson, A. A. G. Sampson, and R. Kagan, "Disagreement in Readmission Rates After Total Hip and Knee Arthroplasty Across Data Sets," Arthroplast Today, vol. 9, pp. 73–77, Jun. 2021, doi: 10.1016/j.artd.2021.04.002.

[15] D. E. Goltz et al., "A Novel Risk Calculator Predicts 90-Day Readmission Following Total Joint Arthroplasty," JBJS, vol. 101, no. 6, 2019, [Online]. Available: https://journals.lww.com/jbjsjournal/fulltext/2019/03200/a_novel_risk_calculator_predicts_90_day.9.aspx

[16] S. M. Kurtz, E. C. Lau, K. L. Ong, E. M. Adler, F. R. Kolisek, and M. T. Manley, "Has Health Care Reform Legislation Reduced the Economic Burden of Hospital Readmissions Following Primary Total Joint Arthroplasty?," J Arthroplasty, vol. 32, no. 11, pp. 3274–3285, 2017, doi: https://doi.org/10.1016/j.arth.2017.05.059.

[17] J. Williams, B. S. Kester, J. A. Bosco, J. D. Slover, R. Iorio, and R. Schwarzkopf, "The Association Between Hospital Length of Stay and 90-Day Readmission Risk Within a Total Joint Arthroplasty Bundled Payment Initiative," J Arthroplasty, vol. 32, no. 3, pp. 714–718, 2017, doi: https://doi.org/10.1016/j.arth.2016.09.005.

[18] M. A. Varacallo, L. Herzog, N. Toossi, and N. A. Johanson, "Ten-Year Trends and Independent Risk Factors for Unplanned Readmission Following Elective Total Joint Arthroplasty at a Large Urban Academic Hospital," J Arthroplasty, vol. 32, no. 6, pp. 1739–1746, 2017, doi: https://doi.org/10.1016/j.arth.2016.12.035.

[19] J. N. Struijs, E. F. De Vries, C. A. Baan, P. F. Van Gils, and M. B. Rosenthal, "Bundled-Payment Models Around the World: How They Work and What Their Impact Has Been," 2020. Accessed: Oct. 26, 2024. [Online]. Available: https://www.commonwealthfund.org/sites/default/files/2020-04/Struijs_bundled_payment_models_around_world_ib.pdf

[20] R. E. Mechanic, "Mandatory Medicare Bundled Payment — Is It Ready for Prime Time?," New England Journal of Medicine, vol. 373, no. 14, pp. 1291–1293, Oct. 2015, doi: 10.1056/nejmp1509155.

[21] C. Bazell, M. Alston, P. M. Pelizzari, and B. A. Sweatman, "BUndled payment US," pp. 1–5, Mar. 27, 2023. Accessed: Oct. 26, 2024. [Online]. Available: https://www.milliman.com/en/insight/what-are-bundled-payments-and-how-can-they-be-used-by-healthcare-organizations

[22] MInister of Health of the Republic of Indonesia, Regulation of the Minister of Health of the Republic of Indonesia concerning Guidelines for the Implementation of the National Health Insurance Program. Jakarta: MInister of Health of the Republic of Indonesia, 2014.

[23] M. A. Fontana, S. Lyman, G. K. Sarker, D. E. Padgett, and C. H. MacLean, "Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty?," Clin Orthop Relat Res, vol. 477, no. 6, pp. 1267–1279, 2019, doi: 10.1097/CORR.0000000000000687.

[24] M. Huber, C. Kurz, and R. Leidl, "Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning," BMC Med Inform Decis Mak, vol. 19, no. 1, p. 3, 2019, doi: 10.1186/s12911-018-0731-6.

[25] J. Sniderman, R. B. Stark, C. E. Schwartz, H. Imam, J. A. Finkelstein, and M. T. Nousiainen, "Patient Factors That Matter in Predicting Hip Arthroplasty Outcomes: A Machine-Learning Approach," Journal of Arthroplasty, vol. 36, no. 6, pp. 2024–2032, 2021, doi: 10.1016/j.arth.2020.12.038.

[26] HealthCare.gov, "Payment Bundling."

[27] NEJM Catalyst, "What Are Bundled Payments?," Catalyst Carryover, vol. 4, no. 1, Aug. 2023, doi: 10.1056/CAT.18.0247.

[28] Institute of Medicine (US) Committee on Quality of Health Care in America, Crossing the Quality Chasm: A New Health System for the 21st Century. Washington (DC; US): National Academies Press, 2001. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25057539

[29] L. B. Oldmeadow, H. McBurney, and V. J. Robertson, "Predicting risk of extended inpatient rehabilitation after hip or knee arthroplasty," J. Arthroplasty, vol. 18, no. 6, pp. 775–779, 2003, doi: 10.1016/s0883-5403(03)00151-7.

[30] M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie, "A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation," J Chronic Dis, vol. 40, no. 5, pp. 373–383, 1987, doi: https://doi.org/10.1016/0021-9681(87)90171-8.

[31] A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey, "Comorbidity measures for use with administrative data," Med. Care, vol. 36, no. 1, pp. 8–27, 1998, doi: 10.1097/00005650-199801000-00004.

[32] D. E. Goltz, S. P. Ryan, C. B. Howell, D. Attarian, M. P. Bolognesi, and T. M. Seyler, "A Weighted Index of Elixhauser Comorbidities for Predicting 90-day Readmission After Total Joint Arthroplasty," Journal of Arthroplasty, vol. 34, no. 5, pp. 857–864, May 2019, doi: 10.1016/j.arth.2019.01.044.

[33] B. Alsinglawi et al., "An explainable machine learning framework for lung cancer hospital length of stay prediction," Sci Rep, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-021-04608-7.

[34] S. D. Mohanty, D. Lekan, T. P. McCoy, M. Jenkins, and P. Manda, "Machine learning for predicting readmission risk among the frail: Explainable AI for healthcare," Patterns, vol. 3, no. 1, Jan. 2022, doi: 10.1016/j.patter.2021.100395.

[35] R. R. Isnanto, I. Rashad, and C. Edi Widodo, "Classification of Heart Disease Using Linear Discriminant Analysis Algorithm," in E3S Web of Conferences, R. R. Isnanto, H. null, and B. Warsito, Eds., EDP Sciences, 2023. doi: 10.1051/e3sconf/202344802053.

[36] C. Klemt et al., "Can machine learning models predict failure of revision total hip arthroplasty?," Arch Orthop Trauma Surg, Jun. 2022, doi: 10.1007/s00402-022-04453-x.

[37] A. A. Shah, S. K. Devana, C. Lee, R. Kianian, M. van der Schaar, and N. F. SooHoo, "Development of a Novel, Potentially Universal Machine Learning Algorithm for Prediction of Complications After Total Hip Arthroplasty," Journal of Arthroplasty, vol. 36, no. 5, pp. 1655-1662.e1, 2021, doi: 10.1016/j.arth.2020.12.040.

[38] I. Yuniar Purbasari, A. Priharyoto Bayuseno, R. R. Isnanto, T. Indah Winarni, and J. Jamari, "Artificial Intelligence and Machine Learning in Prediction of Total Hip Arthroplasty Outcome: A Bibliographic Review," in E3S Web of Conferences, R. R. Isnanto, H. null, and B. Warsito, Eds., EDP Sciences, 2023. doi: 10.1051/e3sconf/202344802054.

[39] F. H. Nham, T. Court, A. K. Zalikha, M. M. El-Othmani, and R. P. Shah, "Assessing the predictive capacity of machine learning models using patient-specific variables in determining in-hospital outcomes after THA," J Orthop, vol. 41, pp. 39–46, 2023, doi: 10.1016/j.jor.2023.05.012.

[40] C. Klemt et al., "The utility of machine learning algorithms for the prediction of patient-reported outcome measures following primary hip and knee total joint arthroplasty," Arch Orthop Trauma Surg, vol. 143, no. 4, pp. 2235–2245, 2023, doi: 10.1007/s00402-022-04526-x.

[41] O. Pakarinen, M. Karsikas, A. Reito, O. Lainiala, P. Neuvonen, and A. Eskelinen, "Prediction model for an early revision for dislocation after primary total hip arthroplasty," PLoS One, vol. 17, no. 9 September, 2022, doi: 10.1371/journal.pone.0274384.

[42] K. N. Kunze, A. V Karhade, E. M. Polce, J. H. Schwab, and B. R. Levine, "Development and internal validation of machine learning algorithms for predicting complications after primary total hip arthroplasty," Arch Orthop Trauma Surg, vol. 143, no. 4, pp. 2181–2188, 2023, doi: 10.1007/s00402-022-04452-y.

[43] K. N. Kunze, A. V. Karhade, A. J. Sadauskas, J. H. Schwab, and B. R. Levine, "Development of Machine Learning Algorithms to Predict Clinically Meaningful Improvement for the Patient-Reported Health State After Total Hip Arthroplasty," Journal of Arthroplasty, vol. 35, no. 8, pp. 2119–2123, Aug. 2020, doi: 10.1016/j.arth.2020.03.019.

[44] I. Lazic et al., "Prediction of Complications and Surgery Duration in Primary Total Hip Arthroplasty Using Machine Learning: The Necessity of Modified Algorithms and Specific Data," J Clin Med, vol. 11, no. 8, 2022, doi: 10.3390/jcm11082147.

[45] A. M. Alaa and M. van der Schaar, "AutoPrognosis: Automated Clinical Prognostic Modeling via Bayesian Optimization with Structured Kernel Learning," in Proceedings of the 35th International Conference on Machine Learning, J. Dy and A. Krause, Eds., Stockholm, Sweden, Jul. 2018, pp. 139–148.

[46] V. Digumarthi et al., "Preoperative prediction model for risk of readmission after total joint replacement surgery: a random forest approach leveraging NLP and unfairness mitigation for improved patient care and cost-effectiveness," J Orthop Surg Res, vol. 19, no. 1, Dec. 2024, doi: 10.1186/s13018-024-04774-0.

[47] J. Park et al., "Machine Learning-Based Predictive Models for 90-Day Readmission of Total Joint Arthroplasty Using Comprehensive Electronic Health Records and Patient-Reported Outcome Measures," Arthroplast Today, vol. 25, Feb. 2024, doi: 10.1016/j.artd.2023.101308.

[48] M. Korvink, C. W. Hung, P. K. Wong, J. Martin, and M. J. Halawi, "Development of a Novel Prospective Model to Predict Unplanned 90-Day Readmissions After Total Hip Arthroplasty," Journal of Arthroplasty, vol. 38, no. 1, pp. 124–128, Jan. 2023, doi: 10.1016/j.arth.2022.07.017.

[49] J. Slezak, L. Butler, and O. Akbilgic, "The role of frailty index in predicting readmission risk following total joint replacement using light gradient boosting machines," Inform Med Unlocked, vol. 25, Jan. 2021, doi: 10.1016/j.imu.2021.100657.

[50] F. C. Kuo, W. H. Hu, and Y. J. Hu, "Periprosthetic Joint Infection Prediction via Machine Learning: Comprehensible Personalized Decision Support for Diagnosis," Journal of Arthroplasty, vol. 37, no. 1, pp. 132–141, Jan. 2022, doi: 10.1016/j.arth.2021.09.005.

[51] H. Shaheen, P. Marimuthu, and D. Rashmi, "The Practical Approaches of Datasets in Machine Learning," in AI Applications in Cyber Security and Communication Networks, C. Hewage, L. Nawaf, and N. Kesswani, Eds., Singapore: Springer Nature Singapore, 2024, pp. 221–227.

[52] Y. Rimal, N. Sharma, and A. Alsadoon, "The accuracy of machine learning models relies on hyperparameter tuning: student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms," Multimed Tools Appl, vol. 83, no. 30, pp. 74349–74364, 2024, doi: 10.1007/s11042-024-18426-2.

[53] M. Feurer and F. Hutter, "Hyperparameter Optimization," in Automated Machine Learning: Methods, Systems, Challenges, F. Hutter, L. Kotthoff, and J. Vanschoren, Eds., Cham: Springer International Publishing, 2019, pp. 3–33. doi: 10.1007/978-3-030-05318-5_1.

[54] R. Kohavi and G. H. John, "Automatic Parameter Selection by Minimizing Estimated Error," in Machine Learning Proceedings 1995, A. Prieditis and S. Russell, Eds., San Francisco (CA): Morgan Kaufmann, 1995, pp. 304–312. doi: https://doi.org/10.1016/B978-1-55860-377-6.50045-1.

[55] A. Goodfellow, Ian; Bengio, Yoshua; Courville, Deep Learning. The MIT Press, 2016.

[56] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," Journal of Machine Learning Research, vol. 13, no. 10, pp. 281–305, 2012, [Online]. Available: http://jmlr.org/papers/v13/bergstra12a.html

[57] A. E. W. Johnson et al., "MIMIC-IV, a freely accessible electronic health record dataset," Sci Data, vol. 10, no. 1, p. 1, 2023, doi: 10.1038/s41597-022-01899-x.

[58] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in KDD '16. ACM, Aug. 2016. doi: 10.1145/2939672.2939785.

[59] J. H. Friedman, "Stochastic gradient boosting," Comput Stat Data Anal, vol. 38, no. 4, pp. 367–378, 2002, doi: https://doi.org/10.1016/S0167-9473(01)00065-2.

[60] L. Breiman, "Random Forests," Mach Learn, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[61] C. Cortes and V. Vapnik, "Support-vector networks," Mach Learn, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[62] A. Johnson et al., "MIMIC-IV (version 3.0)," 2024, PhysioNet. doi: https://physionet.org/content/mimiciv/3.0/.

[63] D. Chicco, "Ten quick tips for machine learning in computational biology," BioData Min, vol. 10, p. 35, 2017, doi: 10.1186/s13040-017-0155-3.

[64] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," Journal of Machine Learning Research, vol. 18, no. 17, pp. 1–5, 2017, [Online]. Available: http://jmlr.org/papers/v18/16-365.html

[65] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," Artif Intell Med, vol. 23, no. 1, pp. 89–109, 2001, doi: https://doi.org/10.1016/S0933-3657(01)00077-X.

[66] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in Proceedings of the 23rd International Conference on Machine Learning, in ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, pp. 233–240. doi: 10.1145/1143844.1143874.

[67] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," Brief Bioinform, vol. 19, no. 6, pp. 1236–1246, 2018, doi: 10.1093/bib/bbx044.

Forecasting Models for Predicting Global Supply Chain Disruptions in Trade Economics

Limei Fu
Tourism Management, Hainan Vocational University of Science and Technology, Haikou, Hainan Province, 571126, China

***Abstract*—Global supply chain disruptions have evolved into a critical challenge for trade economics, and have caused them to reach across industries and economies around the globe. The ability to foresee these disruptions is crucial for policymakers, businesses, and supply chain managers who want to develop actionable strategies for stability. The current document focuses on analyzing the potential application of forecasting models to predict global supply chain disruptions and their efficacy and limitations. A comparison of statistical, machine learning, and hybrid models is performed, and the best methods for predicting disruptions arising from geopolitical events, pandemics, natural disasters, and other external factors are identified. The study considers real-world datasets and various scenario analyses to provide actionable insights. The key findings were obtained by integrating various sources of information, including trade volume fluctuations, transportation bottlenecks, and economic indicators, into predictive frameworks. It is thus a novel contribution to the field of study done by this research to build up an advanced forecasting model that can boost the resilience level and elasticity level of global supply chains, finally playing a key role in the sustainability of trade economics.**

***Keywords*—Supply chain disruptions; forecasting models; trade economics; predictive analytics; global resilience**

I. Introduction

Rightly termed the backbone of contemporary trade economics, global supply chains allow for the smooth transfer of goods, services, and capital across borders. They are integral to global commerce as they bring together in a complex yet interrelated network of economic activities manufacturers, suppliers, distributors, and consumers. However, these elaborate systems are also very sensitive to disruptions, such as geopolitical tensions and economic sanctions, as well as the ever-present threat of natural disasters and pandemics [1]. The disruption of such services has serious consequences, often leading to breaks in the flow of goods and services, affecting global economic stability, business profitability, and consumer welfare. In this context, the eventual forecasting of global supply chain [2-3] disruptions has emerged as a very important and strategic research area the main goal of which is to help the stakeholders effectively anticipate and deal with risks.

The most recent changes regarding how countries interact with one another have acted as the accelerator for the increased susceptibility of supply chains to external shocks[4]. The COVID-19 pandemic, for example, illustrated how easily these supply chains can be disrupted with halting production, shipping delays, and having inadequate materials all becoming standard. Geopolitical events such as trade wars, as well as regional conflicts, have moreover reinforced the need for strong predictive machinery. The traditional methods of managing risks such as reactivity to a problem are today not enough. Instead, proactive strategies that rely on sophisticated forecasting models are now critical [5-6].

Forecasting models refer to techniques that use historical records, statistical methods, and computational tools to predict future events and trends. Models of this kind have been useful in such global phenomena as the supply chain problems occurring in the world through the follow-up of various sets of data and sensitivity analyses of the contributing factors, like weather, trade volumes, and even political events, to the point of realizing the critical points [7-8]. Forecasting models serve to indicate risks early on, thus providing the companies and governments the power to manage necessary assets and the business continuity resourcing to minimize any interruptions. On the other hand, the use of trustworthy forecasting methods based on such models in global supply chains is challenging because of the diverse nature of disturbances and the intricate structure of the complex networks of supply and demand that are interconnected [9-10].

Our studies on the forecasting models have unearthed several forecasting models, all of which come with unique benefits and downsides. The most straightforward models are based on historical statistical data, which is why they are widely applied. It is easy to understand the long-term and seasonal changes presented through this model, but it might be hard for them to capture the unpredicted and on-the-spot ones [11]. The use of such models like the neural network or decision tree machine learning has been on the increase, as those models have a clearer picture of reality and thus provide better forecasts by capturing the non-linear relationships and picking up the data regardless of the data collection procedure. As a kind of progressive matter among researchers [12-13], Hybrid models have also emerged, which are based on a combination of statistical algorithms and machine learning techniques and thus can become a good balance between clarity and accuracy of prediction.

Although methodologies for prediction have grown by leaps and bounds, in the realm of global supply chain forecasting capabilities, there are still noticeable gaps in precision. A major hindrance is the lack of high-quality, timely data, often due to poor infrastructure in some areas. Furthermore, the complexity and mutual reliance of global supply chains [14-15] make it difficult to model interactions and cascading effects. For example, an incident in one location or industry can affect the entire supply chain, hence a holistic and systemic approach to forecasting is needed.

To address these challenges, both researchers and practitioners have noted the diversity of data sources and the integration of different fields as crucial. Concurrently, the technological advances in data analytics, artificial intelligence, and cloud computing can create new ways for us to develop complex forecasting models [16]. For example, the availability of satellite imagery along with IoT devices enables us to learn about bottlenecks and inventories in real-time while using news sentiment analysis, and social media can tell the early indications of geopolitical risks [17-18]. Therefore, the incorporation of such technologies in forecasting models could lead to greater precision and reliability, which in turn, could be helpful for decision-makers.

The urgency of the consequences of global supply chain disruptions emphasizes the need for forecasting accuracy in the economic sector. Supply chain disruptions can lead to increasing costs, a decline of productivity, and a reduction of order at the company level. Consequently, firms may suffer the resulting consequences of inflation, unemployment, and increasing economic disparities that have broader implications [19]. Furthermore, such disruptions could bring the international trade system to its knees, reduce the confidence of consumers, and bring growth economic grind to a halt. The power of forecasting models to facilitate timely and informed decision-making, to minimize these negative effects, and to contribute to the overall sustainability of global supply chains is huge [20-21].

The advantages of forecasting models impact the immediate dangers created by making such models an essential part of sustainable and ethical trade practices. Additionally, it can be said that variations in the supply chain can strongly influence weak populations such as the laborers in developing nations who rely on a constant flow of external trade to survive [22]. The predictions and prevention of disruptions in the initial stages could be achieved through modeling forecasts which are also associated with the survival of these communities and their sustainable economic development. In addition, accurate predictions can make the process environmentally sustainable by minimizing waste, optimizing the logistics, and reducing the supply chain's carbon footprint during the operations [23-24].

A. Objectives

1) To carry out the evaluation and comparison of the efficiency, precision, extensibility, and flexibility of the mobile hybrid of statistical machine learning models in forecasting global supply chain disturbances.

2) To explore the significance of applying an assorted array of data that encompasses economic variables, political alterations, and technical considerations in the modeling processes to enhance the efficiency of their prognosis.

3) To establish a specific and reliable forecasting model that can address the limitations of the previous ones and offer sound advice for minimizing supply chain risks.

This research proposes to attain set objectives that will connect the development of theories with the practical aspects of global supply chain disruption forecasting now and in the future. The findings will not only provide valuable guidance for researchers and practitioners but also contribute to the broader discourse on building resilient and sustainable supply chains.

To sum up, the escalated regularity and intensity of global supply chain interruptions require a fundamental change in modus operandi, going from a reactive to a proactive risk management framework [25-26]. By employing forecasting models' stakeholders, the ones involved in the supply chain, can point out the unforeseen problems. Hence, the potential of these models can be exploited after deeply analyzing the technical, data-related as well as conceptual problems. The research brings forward a fruitful look into the unsolved questions of this field, ultimately leading to the mobile solution, green and pro-active supply chain economy globally.

This paper consists of the following sections: In Section 2, a comprehensive literature review is presented with an emphasis on major research contributions and also the gaps addressed by the proposed model. The details of the data preprocessing, model development, and evaluation metrics are given in Section 3. The results of the forecasting model's performance are presented both quantitatively and graphically in Section 4. The important findings, practicality, and recommendations for further improvements of the findings are articulated in Section 5. Lastly, in Section 6, the paper's conclusion is drawn, and a summary of the research work's principal points, the deficiencies of the study, and the potential avenues for future research are given.

II. Literature Review

The area of global supply chain forecasting has become heavily discussed over the past few years because the occurrences and results of disruptions in trade economics are increasing. Researchers have investigated multiple models and methodologies in their search to accurately anticipate and act on these disruptions using techniques such as statistical analysis, machine learning, and data integration [27]. The current section is a summary of important investigations in this field, pointing out the progress made, the techniques used, and the drawbacks of each one to help give a complete view of the prevailing situation in research. The results from these investigations create the basis for the identification of the gaps and the suggestion of a better forecasting system.

The study conducted by Bhadra et al [28]. targets the disruptive factors brought by the Russian-Ukrainian conflict to the world food supply chain with special emphasis on South Korea's Food, Beverage, and Tobacco (F&B) sector. The research uses Autoregressive Integrated Moving Average (ARIMA) modeling to show that the escalation of the conflict leads to a negative trend in the KOSDAQ F&B sector returns. The study highlights that South Korea has to normalize the use of healthy, safe food as well as develop its domestic agri-economy for an overall self-sufficiency.

Shafipour et al. [29] proposed a new approach that is real time, yet comprehensive, in trying scenarios of the Supply Chain Network in the medical field. The authors highlighted a need for a change in perspectives towards the medical and engineering fields, a two-way approach that includes both issues, and thus they suggested an integration of the software aspects of communication with the problem-solving ones to

involve all the stakeholders in the enhancement of communication and problem-solving in the medical field. The authors came to the conclusion that by integrating all the important factors in the actual implementation process, it is possible to cover all the aspects of the project and thus ensure its success.

Zheng et al. [30] studied the impact of the COVID-19 epidemic on the medical mask supply chain through a simulation model utilizing AnyLogistix. Their findings stress the importance of having a backup facility and correctly optimizing the inventory level so that the supply chain can recover in the face of adversity. According to the same research work, the duration of the disruption period for the supply chain's downstream facilities is the main factor that affected the performance of the supply chain, which during future disturbances offers methods to better the flexibility and the resilience of medical mask supply chains through. In a medical mask supply chain, it was found that well-managed inventory levels and the provision of backup capacity both can reduce the effect of the agri-food chain crises on the selection.

Queiroz et al. [31] explored how blockchain technology (BCT) can be integrated into operations and supply chain management (OSCM) processes and practices in Brazil. The study adopted the Unified Theory of Acceptance and Use of Technology (UTAUT) model and identified trust, facilitating conditions, social influence, and effort expectancy as the most influential factors of BCT adoption. The research shed light on the challenges and barriers that blockchain technology faces in the transformation of societies, especially in developing economies such as Brazil.

Manupati et al. [32] propose a disruption prediction model in a supply chain network that involves smart contracts via blockchain technology. The authors advocate using a genetic algorithm-based methodology that intelligently addresses both pre- and post-disruption situations, thereby ensuring a holistic decision-making environment for disruption management. The research findings provide key insights for organizations in adapting and overcoming supply chain disruption caused by various, complex multi-tier supply chains.

Paul et al. [33] strive to understand how to gauge transportation disruption risks in supply chains using the tool called Bayesian Belief Network (BBN). The study highlights the most vulnerable sources of disruption and the parameters relating to those by developing a BBN-based model. The research, illustrated with a case study from the pharmaceuticals sector in Bangladesh, illustrated the power of BBN as a tool for predicting calamities in transportation and as a facilitator for supply chains development strategies that secured the transport of sensitive materials across delivery networks.

Camur et al. [34] have conducted research on the COVID-19 pandemic and geopolitical conflicts to determine their impact on global supply chains by analyzing the unpredictability of product delivery dates in logistics services. Through various regression models, including Random Forest (RF) and Gradient Boosting Machine (GBM), the study obtained the fact that tree-based models yield the best results for predicting availability dates. Results indicated that these models could be implemented to manage supply chain interruptions and thus lessen the risks involved.

TABLE I. LITERATURE COMPARISON

Authors	Focus	Key Methodology	Findings
Bhadra et al.	Supply chain disruptions in the F&B sector due to geopolitical events	ARIMA model for stock return prediction	Negative trend in F&B stock returns observed due to the Russia-Ukraine conflict; need for domestic self-sufficiency in food production
Shafipour et al.	Predicting time-to-disruptive events in supply chains using survival analysis techniques	Statistical flowgraph models (SFGMs) for time-to-event data analysis	SFGMs offer insights into system reliability, hazard functions, and identify supply chain weaknesses for better disruption management
Zheng et al.	Impact of the pandemic on the medical mask supply chain and strategies for resilience	Simulation models with AnyLogistix, Green Field Analysis, and risk analysis	Adding backup facilities and optimizing inventory levels helps improve supply chain resilience during disruptions
Queiroz et al.	Adoption of blockchain technology (BCT) in supply chain management and related barriers	UTAUT model, PLS-SEM for empirical validation of blockchain adoption factors	Key factors such as trust, facilitating conditions, and social influence significantly impact BCT adoption in supply chains in Brazil
Manupati et al.	Recovery strategies in supply chains using blockchain technology and smart contracts	Genetic algorithm-based approach for disruption prediction and recovery strategies	Integration of pre- and post-disruption strategies offers holistic decision support for managing supply chain disruptions and minimizing performance loss
Paul et al.	Assessing transportation disruptions in supply chains and their risk factors	Bayesian Belief Network (BBN) model for disruption risk analysis	BBN captures interdependencies between disruption risk factors and helps build resilient strategies for managing transportation disruptions
Camur et al.	Impact of geopolitical and pandemic disruptions on predicting product availability dates in supply chains	Regression models (RF, GBM, Random Forest, and Neural Networks) for prediction	Tree-based models (RF, GBM) perform best in predicting product availability dates, aiding in better supply chain management during disruptions
Mittal et al.	Predicting and mitigating risks in supply chains using AI-driven machine learning and deep learning models	Machine learning (ML) and deep learning (DL) models, including CNN networks	Deep CNN regression model outperforms others in predicting supply chain risks, offering insights for better resilience and stability in operations
Proposed Model	AI-Driven Global Supply Chain Forecasting	RNN-LSTM with Attention Mechanism and Ensemble Learning	Enhances accuracy, captures long-term dependencies, and adapts dynamically to emerging risks.

On the other hand, Mittal et al. [35], backed by Artificial Intelligence, have put forward a suggestion to counteract the vulnerabilities in the supply chain that arise due to the external forces like pandemics and inflation, etc. In order to explain it, the study made the best use of machine learning and deep learning that are including linear regression and convolutional neural networks (CNN), while also a different way of augmenting the data has been proposed by use of the Fuzzy C-means method. The Deep CNN regression model reaches a higher forecasting capacity than other models regarding potential risks in the supply chain and at the same time it provides information for strategists and planners on how to improve the stability and resilience of the supply chain.

III. METHODOLOGY

The methodology that has been adopted in this research has been framed in such a way that it handles the challenges of effectively forecasting the occasions of global supply chain disruptions. The integration of advanced computational techniques, data preprocessing methods, and performance evaluation frameworks makes the study a robust one that can predict disruptions and provide action-oriented insights. The methodology proposed in this paper has a systematic approach as the central idea, covering the topics of data acquisition, preprocessing, and the development, evaluation, and iterative refinement of models.

The key to the success of the proposed methodology is the careful integration of different data sources that signify the major factors of global supply chain disruptions. In particular, external factors such as geopolitical events, natural disasters, and pandemics as well as trade and economic data such as import-export volumes, exchange rates, and commodity prices are the sources of information; logistics data including shipping schedules, container availability, and transportation bottlenecks; and historical disruption records are the focus of the coverage. The advantages of using multiple data sources provide the forecasting models with a thorough understanding of the global supply chain landscape. Being in a position to use real-time data, the model can dynamically adjust itself to new situations and rising risks.

To ensure the quality and the proper use of forecasting data, the acquired data should be cleaned beforehand. The cleaning stage consists of data gathering for fixing issues like missing values, outliers, and inconsistencies; selection of the most relevant disposition of the information geared towards predicting disruptions; and conversion techniques, like normalization and scaling, for the standardization of the input data. These steps in the preprocessing process of the data will enhance the performance and reliability of the forecasting model as they will trim the noise and hence will increase the interpretability of the data.

The methodology is centered on the refining of a superior forecasting model architectural design. This study has utilized deep learning techniques like recurrent neural networks (RNNs), and in particular, long short-term memory (LSTM) networks, for this type of forecasting, which are well adapted to time series forecasting since they have a unique ability to capture temporal relationships between time steps due to their memory capacity. Moreover, the attention mechanism has been blended into the model's structure to enhance its prediction capabilities by directing the model on the most important features and the most crucial time steps. Simultaneously, ensemble methods leveraging the strengths of multiple models were employed to enhance robustness and reduce the possibility of overfitting.

The forecasting pipeline integrates scenario analysis and impact forecasting, providing a comprehensive view of different disruptions. Scenario analysis is a technique that uses the generation of both, positive, and negative what-if scenarios to help in making up-to-date predictions. The quantification of the potential consequences of disruptions on the important metrics of the supply chain, such as lead times, costs, and service levels, is being performed by the impact forecasting. Data visualization methods are then used to display the forecasting results simply and practically to make the decision-making process easier for stakeholders.

The forecasting model's performance is submitted to the evaluation using a previously determined group of metrics including accuracy, lead time, and scenario coverage. The model's accuracy to predict the disruptions accurately is the measure of its success, while the lead time is a measure of its speed of warning. Scenario coverage is the measure of how well the model will respond to any disruptions by its realization of a wide variety of alternative scenarios which ensures its generalizability in the various contexts. All the thus evaluated metrics will assist the methodology so that the suggested model would live up to the real-world applications requirement.

The proposed methodology's significant feature is its continuous improvement emphasis. The unpredictable nature of global supply chains requires that the forecasting model be constantly updated with the latest data and external factors. Through the methodology, the process of data updates, model refinements, and mechanisms for adaptability are incorporated to ensure that forecasting remains relevant and effective over time. The model method is based on the principle of iterative enhancement, which conforms to the principles of continuous learning and improvement.

The working of the proposed forecasting model within a conceptual framework such as the one presented in “Fig 1” is illustrated. The framework begins with the collection of input data from various sources such as external factors, trade and economic data, logistics, and historical disruption records. The preprocessing of the data involves data cleaning, feature selection, and data transformation to make the data suitable for analysis. The figure highlights the architecture of the core forecasting model which consists of RNNs, LSTMs, attention mechanisms, and ensemble techniques forming the model's input and output layers. The pipeline of forecasting integrates scenario analysis, impact forecasting, and data presentation to produce actionable insights. Risk management, policy advisement, and continuous monitoring are the applications of the forecasting model that are critical for the solution of supply chain disruptions. The model's accuracy metric module measures how successful the model was, how fast it was, and how deeply the scenarios were analyzed while the continuous improvement component guarantees the ability to change the

data and the model when necessary. This integrated approach is the theoretical framework of the proposed methodology and it is a detailed solution for predicting global supply chain disruptions.

Thus, adopting this structured methodology ensures that the suggested forecasting model overcomes the limitations of present methods and gives a scalable, adjustable, and practical framework for predicting worldwide supply chain disruptions. The approach of the software development variable is capable of sustaining the extremely important data of diversity, new sources of training, and procedures of incremental improvement which constitute the improvement guarantee for the model such that it can stand up against the highly complex and linking global trade environment.

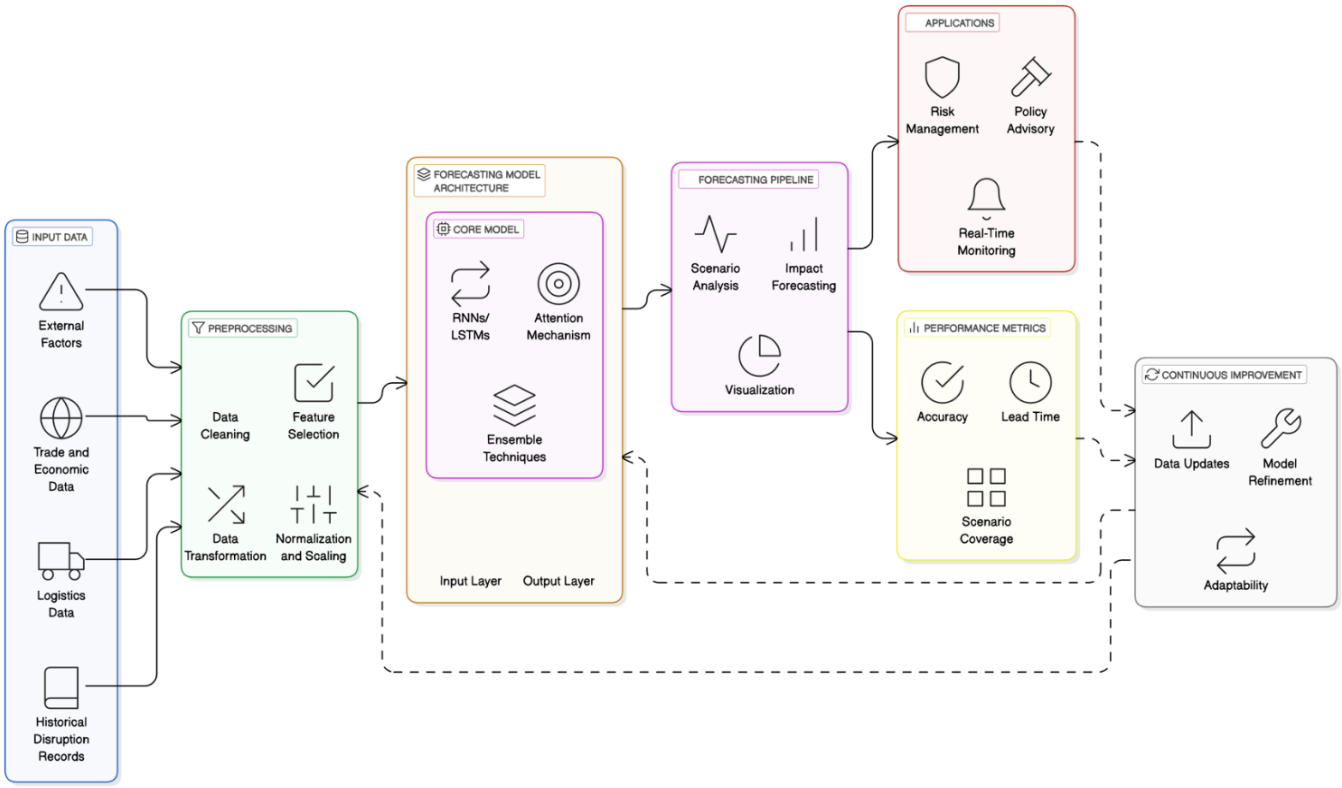


Fig. 1. Proposed model diagram.

IV. Simulation and Results

The findings of this study shed light on how good the proposed forecasting model is in predicting disruptions in the global supply chain. In the training and evaluation phase, the model was implemented using the Supply Chain Data dataset sourced from Kaggle (Dataset Reference), which encompasses the historical records of disruptions, the trends in trade volume, and logistic information. The key findings are the main focus of the discussion below in "Table II".

The appropriateness of the model's predictions was checked as both the predicted numbers and the real ones were compared for the 12 months of 2024. The results, which are shown in "Fig. 2", indicate a tight coupling between the predicted disruptions and those that happened. A case that can be cited in this case is the disruption forecasted by the model for January where it was observed that there were only two disruptions, while the number the model predicted was three. In another instance, April and July were forecasts 7 and 9 which coincided perfectly with the real figures for instance. Despite this, the former diverged from the latter slightly in some months; it is nevertheless the case that generally, the right-wrong count as well as the sum were of a considerable size thus demonstrating that the model does a remarkable job detecting violations regardless of the diversity of cases presented.

TABLE II. Supply Chain Forecasting Data

Month	Predicted Disruptions	Actual Disruptions	Lead Times (Weeks)
Jan	3	2	2.5
Feb	4	5	2.7
Mar	5	4	2.6
Apr	7	8	2.8
May	6	7	2.9
Jun	8	9	3.0
Jul	9	10	3.1
Aug	10	11	3.2
Sep	8	9	3.1
Oct	7	8	3.0
Nov	6	7	2.9
Dec	5	6	2.8

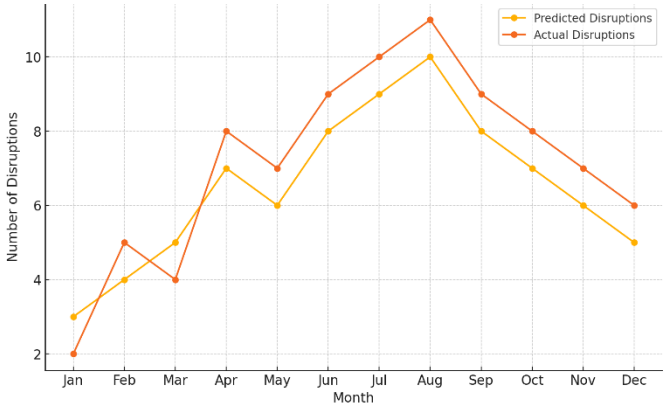


Fig. 2. Predicted vs actual supply chain dsisruptions (2024).

Also, a prime performance indicator that was evaluated was the lead time of disruption alerts which essentially shows the capability of the model to give timely warnings. At the beginning of "Fig. 3", it can be seen that the average time lead times were 2.5 weeks in January and 3.2 weeks in August. With such timings, respective stakeholders can engage in proper planning and risk mitigation measures. The fact that led times were constant in different months until the recent one suggests that the model is not only sturdy but also can change its features quite well as per the changes in the supply chain environment

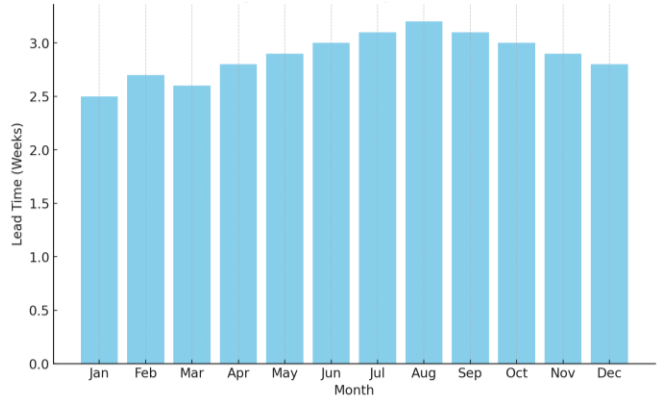


Fig. 3. Average lead times by month (2024).

The results further indicate the seasonal variability of disruption patterns. For example, it was the summer months (June to August) that saw the peak of disruptions, which could be a result of the increase in trade volume, geographical conditions, and political issues. In contrast, disruption levels were relatively low during the winter months, which points to times of low volatility in the global supply chain. These signals are crucial for policymakers and businesses to distribute resources meaningfully and control risks during the most dangerous times.

Different data sources were combined, and the model was developed ensuring each one of them did their part successfully. The researchers made use of selective outside data, for example, economic and trade data, logistics records, historical disruption trends, and sequestration were all used to identify different aspects of the model of the global supply chains ecosystem. The dynamic combination of external factors such as geopolitical events and natural disasters has contributed to the model's improved ability to precisely predict disruption. The innovative function of the model is indicative of the fact which manifold is the application of data in the construction of reliable forecasting models.

The proposed model was compared with baseline approaches including simple statistical models, which demonstrated significant improvements in accuracy and leading time. The latest technologies introduced in our model include recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and attention mechanisms enabled it to capture the subtle temporal patterns and detect the complicated data signals. Additionally, the application of ensemble methods offered some duplication to the given model through the reduction of the influence of the outliers and noise on the outcomes of the methods resulting in an improved performance of the model.

The insights provided through visualizations and the forecasting pipeline were instrumental in decision-making. Scenario analysis, along with impact forecasting, integrated into the model allows stakeholders to explore various what-if scenarios and assess the potential consequences of disruptions. The graphical representations, such as those in Figures 2 and 3, helped readers intuitively understand the results, enabling informed and timely decision-making.

While the results seem promising, certain limitations were found during the evaluation process. Minor discrepancies between predicted and actual values in some months underscore the need for continuous model refinement and updates. Moreover, the reliance on high-quality and real-time data poses obstacles, especially in regions with limited data infrastructure. The ongoing data updates and the model should be developed in such a way as to fully address the weighty concerns and test the effectiveness of the model in changing supply chain contexts.

V. Discussion

In conclusion, the results show that the proposed forecasting model has a high ability to achieve not only accurate predictions of global supply chain disruptions but also the generation of useful information that decision-makers can act on. The graphical and tabular representations of the findings showcase the model's capability to recognize trends, quantify risks, and assist in the decision-making process. The diverse data sources integrated with the leading-edge computational techniques and a commitment to continuous improvement ensure that the model is well adapted to the challenges of contemporary supply chains. It is essential that we address the issue of resilience and sustainability in the global trade economic system, and this study is part of the ongoing discussions on this subject.

VI. Conclusion

The research findings highlight the significant role of data-driven forecasting models in understanding global supply chain disruptions and their impact on trade economics. The accuracy of the proposed model was shown to be high through closely predicted and actual situations as a result the deviation for the

majority of the months is very small. Also, in supplying an average lead time of 2.5 to 3.2 weeks, the model can bring crucial warnings early enough to allow for the application of risk mitigation strategies by stakeholders. Additionally, the integration of heterogeneous data sources like trade information, logistical data, and others played a key role in obtaining these outcomes while sophisticated computational methods such as RNNs and LSTMs with attention mechanisms were the foundational elements enabling the model to understand complicated patterns of time. Furthermore, the seasonal trends unveiled in the analysis highlight the importance of the model in providing actionable insights through it, especially in high-risk seasons, thus helping to achieve the goals of risk resilience and sustainability in the global supply chains.

However, the model has certain constraints that must be resolved. While the model performs well in most months, occasional deviations in forecasts highlight the need for continuous refinement. Furthermore, the data must be of high quality, timely, and recent, which is a hard task in those regions, where the technology is not well implemented for what is required such as the data infrastructure is insufficient. Consequently, in the upcoming work the primary emphasis will be on the model's flexibility to handle cases with missing data and the model's capacity to deal with exceptional handling situations among others. The areas of research that need improvement as indicated by these gripes are also anticipated to be the ones where similar systems can be studied and modified for a lot of other comparable industry settings in the future and even the other variables can have a global effect and therefore necessitate the interference of global tools in the supply chain management.

REFERENCES

[1] K. Amir, "COVID-19 Pandemic: A Snapshot of Global Economic Repercussions and Possible Retaliations," SSRN Electronic Journal, 2020, doi: 10.2139/ssrn.3593754.

[2] R. De Souza, M. Goh, and F. Meng, "A Risk Management Framework for Supply Chain Networks," TLI Asia Pacific White Papers Series, vol. 07, no. August, 2006.

[3] X. Shi, A. Tsun, S. Cheong, and M. Zhou, "Economic and Emission Impact of Australia-China Trade Disruption: Implication for Regional Economic Integration," ERIA Discussion Paper Series, no. 387, 2021.

[4] Z. Liang, S. Adnan, and C. Leilei, "Blockchain Technology in Post-Covid Agriculture," in ACM International Conference Proceeding Series, 2022. doi: 10.1145/3512676.3512685.

[5] Baker McKenzie, "Beyond COVID-19: Supply chain resilience holds Key to recovery," Oxford Economics, 2020.

[6] B. McKenzie, "Beyond COVID-19: Supply chain resilience holds Key to recovery," Oxford Economics, 2020.

[7] I. V. Boyko, "Freight Flows in the Baltic Seaports of Russia: Factors, Trends and Perspective," Spatial Economics, vol. 17, no. 4, 2021, doi: 10.14530/SE.2021.4.168-185.

[8] G. Burstein and I. Zuckerman, "Deconstructing Risk Factors for Predicting Risk Assessment in Supply Chains Using Machine Learning," Journal of Risk and Financial Management, vol. 16, no. 2, 2023, doi: 10.3390/jrfm16020097.

[9] E. J. Boyer, "Responding to Environmental Uncertainties in Critical Supply Acquisition: An Examination of Contracting for Personal Protective Equipment (PPE) in the Aftermath of COVID-19," Journal of Public Administration Research and Theory, vol. 34, no. 2, 2024, doi: 10.1093/jopart/muad015.

[10] R. Handfield, H. Sun, and L. Rothenberg, "Assessing supply chain risk for apparel production in low cost countries using newsfeed analysis," Supply Chain Management, vol. 25, no. 6, 2020, doi: 10.1108/SCM-11-2019-0423.

[11] M. Park and N. P. Singh, "Predicting supply chain risks through big data analytics: role of risk alert tool in mitigating business disruption," Benchmarking, vol. 30, no. 5, 2023, doi: 10.1108/BIJ-03-2022-0169.

[12] D. Ivanov, "Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case," Transp Res E Logist Transp Rev, vol. 136, 2020, doi: 10.1016/j.tre.2020.101922.

[13] R. Ojha, A. Ghadge, M. K. Tiwari, and U. S. Bititci, "Bayesian network modelling for supply chain risk propagation," Int J Prod Res, vol. 56, no. 17, 2018, doi: 10.1080/00207543.2018.1467059.

[14] G. Zheng, L. Kong, and A. Brintrup, "Federated machine learning for privacy preserving, collective supply chain risk prediction," Int J Prod Res, vol. 61, no. 23, 2023, doi: 10.1080/00207543.2022.2164628.

[15] G. Baryannis, S. Dani, and G. Antoniou, "Predicting supply chain risks using machine learning: The trade-off between performance and interpretability," Future Generation Computer Systems, vol. 101, 2019, doi: 10.1016/j.future.2019.07.059.

[16] A. Brintrup et al., "Supply chain data analytics for predicting supplier disruptions: a case study in complex asset manufacturing," Int J Prod Res, vol. 58, no. 11, 2020, doi: 10.1080/00207543.2019.1685705.

[17] M. Weinke, P. Poschmann, and F. Straube, "Decision-making in Multimodal Supply Chains using Machine Learning," in Proceedings of the Hamburg International Conference of Logistics, 2021.

[18] A. S. Almasoud, "Blockchain-Based Secure Storage And Sharing Of Medical Data Using Machine Learning," in Proceedings - 2023 10th International Conference on Social Networks Analysis, Management and Security, SNAMS 2023, 2023. doi: 10.1109/SNAMS60348.2023.10375435.

[19] D. Gallego-García, S. Gallego-García, and M. García-García, "Application of the human-oriented planning model in the supply chain: From the global system to specific cases," in Advances in Science and Technology, 2023. doi: 10.4028/p-8lvU3d.

[20] N. Kalogeras et al., "State of the art in benefit-risk analysis: Economics and Marketing-Finance," 2012. doi: 10.1016/j.fct.2011.07.066.

[21] M. Gabellini, L. Civolani, A. Regattieri, and F. Calabrese, "A Data Model for Predictive Supply Chain Risk Management," in Lecture Notes in Mechanical Engineering, 2023. doi: 10.1007/978-3-031-34821-1_40.

[22] C. Atwater, R. Gopalan, R. Lancioni, and J. Hunt, "Measuring supply chain risk: Predicting motor carriers' ability to withstand disruptive environmental change using conjoint analysis," Transp Res Part C Emerg Technol, vol. 48, 2014, doi: 10.1016/j.trc.2014.09.009.

[23] T. Okada, A. Namatame, and H. Sato, "An Agent-Based Model of Smart Supply Chain Networks," 2016. doi: 10.1007/978-3-319-27000-5_30.

[24] A. Lorenc, M. Kuźnar, T. Lerher, and M. Szkoda, "Predicting the probability of cargo theft for individual cases in railway transport," Tehnicki Vjesnik, vol. 27, no. 3, 2020, doi: 10.17559/TV-20190320194915.

[25] C. Liu, T. Shu, S. Chen, S. Wang, K. K. Lai, and L. Gan, "An improved grey neural network model for predicting transportation disruptions," Expert Syst Appl, vol. 45, 2016, doi: 10.1016/j.eswa.2015.09.052.

[26] R. S. Kadadevaramth, D. Sharath, B. Ravishankar, and P. Mohan Kumar, "A review and development of research framework on technological adoption of blockchain and iot in supply chain network optimization," in 2020 International Conference on Mainstreaming Block Chain Implementation, ICOMBI 2020, 2020. doi: 10.23919/ICOMBI48604.2020.9203339.

[27] I. S. K. Acquah, C. Baah, Y. Agyabeng-Mensah, and E. Afum, "Sustainable supply chain learning and employee green creativity: Implications for disruption management and green competitiveness," in Increasing Supply Chain Performance in Digital Society, 2022. doi: 10.4018/978-1-7998-9715-6.ch012.

[28] A. D. Rusmanto, F. N. Maharani, M. Setiawan, and A. N. Arofah, "Pengaruh Stres, Keteraturan Makan, dan Makanan Minuman Iritatif Terhadap Sindrom Dispepsia Pada Mahasiswa Angkatan 2019 Fakultas

Kedokteran Universitas Muhammadiyah Malang," Jurnal Kedokteran Syiah Kuala, vol. 22, no. 4, 2022.

[29] G. Shafipour and A. Fetanat, "Survival analysis in supply chains using statistical flowgraph models: Predicting time to supply chain disruption," Commun Stat Theory Methods, vol. 45, no. 21, 2016, doi: 10.1080/03610926.2014.957856.

[30] Y. Zheng, L. Liu, V. Shi, W. Huang, and J. Liao, "A Resilience Analysis of a Medical Mask Supply Chain during the COVID-19 Pandemic: A Simulation Modeling Approach," Int J Environ Res Public Health, vol. 19, no. 13, 2022, doi: 10.3390/ijerph19138045.

[31] M. M. Queiroz, S. Fosso Wamba, M. De Bourmont, and R. Telles, "Blockchain adoption in operations and supply chain management: empirical evidence from an emerging economy," Int J Prod Res, vol. 59, no. 20, 2021, doi: 10.1080/00207543.2020.1803511.

[32] V. K. Manupati, T. Schoenherr, M. Ramkumar, S. Panigrahi, Y. Sharma, and P. Mishra, "Recovery strategies for a disrupted supply chain network: Leveraging blockchain technology in pre- and post-disruption scenarios," Int J Prod Econ, vol. 245, 2022, doi: 10.1016/j.ijpe.2021.108389.

[33] S. Paul, G. Kabir, S. M. Ali, and G. Zhang, "Examining transportation disruption risk in supply chains: A case study from Bangladeshi pharmaceutical industry," Research in Transportation Business and Management, vol. 37, 2020, doi: 10.1016/j.rtbm.2020.100485.

[34] M. C. Camur, S. K. Ravi, and S. Saleh, "Enhancing supply chain resilience: A machine learning approach for predicting product availability dates under disruption," Expert Syst Appl, vol. 247, 2024, doi: 10.1016/j.eswa.2024.123226.

[35] U. Mittal and D. Panchal, "AI-Based Evaluation System for Supply Chain Vulnerabilities and Resilience Amidst External Shocks: An Empirical Approach," Reports in Mechanical Engineering, vol. 4, no. 1, 2023, doi: 10.31181/rme040122112023m.

Developing an IoT Testing Framework for Autonomous Ground Vehicles

Murat Tashkyn, Amanzhol Temirbolat, Nurlybek Kenes, Amandyk Kartbayev
School of Information Technology and Engineering, Kazakh-British Technical University, Almaty, Kazakhstan

***Abstract*—Autonomous ground vehicles play a crucial role in the Internet of Things, offering transformative potential for applications such as urban transportation and delivery services. These vehicles can operate autonomously in uncertain environments, making reliable testing essential. This study develops and analyzes a testing framework for autonomous ground vehicles, focusing on their motion control systems and electronic modules. The research reviews testing methods for printed circuit boards (PCBs), highlighting the need for JTAG testing implementation for vehicle modules. Functional testing was conducted on key components such as cameras, LiDARs, and wireless interfaces under various conditions. Results show that JTAG testing successfully detects faults with precise localization, while functional tests confirm stable component performance. Environmental tests revealed that most components perform reliably within optimal conditions, with failures occurring at temperatures beyond ±70°C and humidity levels exceeding 90% RH. The developed testing system enhances the reliability of autonomous delivery vehicles.**

***Keywords*—Autonomous vehicle; testing system; IoT; functional testing; electronic modules; delivery automation**

I. Introduction

Intelligent autonomous vehicles are particularly relevant today. They can operate in uncertain environments and are of significant interest for a wide range of practical applications, including food delivery. This study focuses on the development and analysis of a testing system for unmanned delivery vehicles. Before such vehicles can be deployed in real-world operations, extensive testing is required to ensure their stability, reliability, and performance, as well as prior research on testing systems for unmanned delivery vehicles.

However, several challenges complicate the testing process for these vehicles.

- Testing in real-world conditions: It is essential to test delivery vehicle modules under conditions that closely mimic real environments. This includes verifying components under the influence of multiple adverse factors, such as humidity.
- Testing of PCB modules: Testing individual modules separately cannot guarantee the proper functioning of the entire delivery vehicle after assembly. Testing of all printed circuit boards (PCBs) as a unified system is necessary, since the vehicles are cyber-physical systems that perceive, process, and physically respond to information from the real world.
- Automation of testing: Manual testing introduces the risk of human error, which, even in a single component, can render the entire vehicle non-operational.
- Regulatory gaps: Unmanned delivery vehicles lack specific regulations and testing requirements, making their verification process unclear and inconsistent.

A plan of the testing system is developed, representing a set of methodologies for assessing the operational functionality and reliability of an autonomous delivery vehicle. They consist of several key subsystems, including a power supply, lighting system, motor, control system, communication, and sensing system. Thus, the delivery vehicles under investigation are complex systems comprising numerous modules. Each module requires thorough testing, as its functionality directly impacts the performance of the entire system. This would require complete disassembly and retesting of the vehicle.

The control system manages the operation of all components and coordinates their interactions. Drives are used to control movement and can be electric, hydraulic, or pneumatic, allowing the vehicle to maneuver in space, including on-the-spot rotations and 360-degree turns, which are particularly crucial in confined places. The control system also processes data from the perception system and makes decisions based on that information, without which its operation would be impossible.

At the core of the structural diagram is a computational device that governs the robot's actions through control algorithms, as shown in Fig. 1. This device processes video streams received from onboard cameras. Wheel controllers receive speed requirements for each wheel from the platform controller and manage the motor to maintain the specified speed under varying driving conditions. The peripheral controller regulates the operation of the lid motor, the locking mechanism, and the onboard lighting system.

The platform controller ensures power delivery to the platform, regulates current in each power branch, and switches to a backup power source when necessary. The power supply provides electricity to all the vehicle's systems. It can include various energy sources such as batteries, generators, or other power sources, which gives the potential for scaling the system to other types of autonomous transport or applications. Also EMC testing is a critical stage, ensuring that the system can operate with influence from external radiation sources.

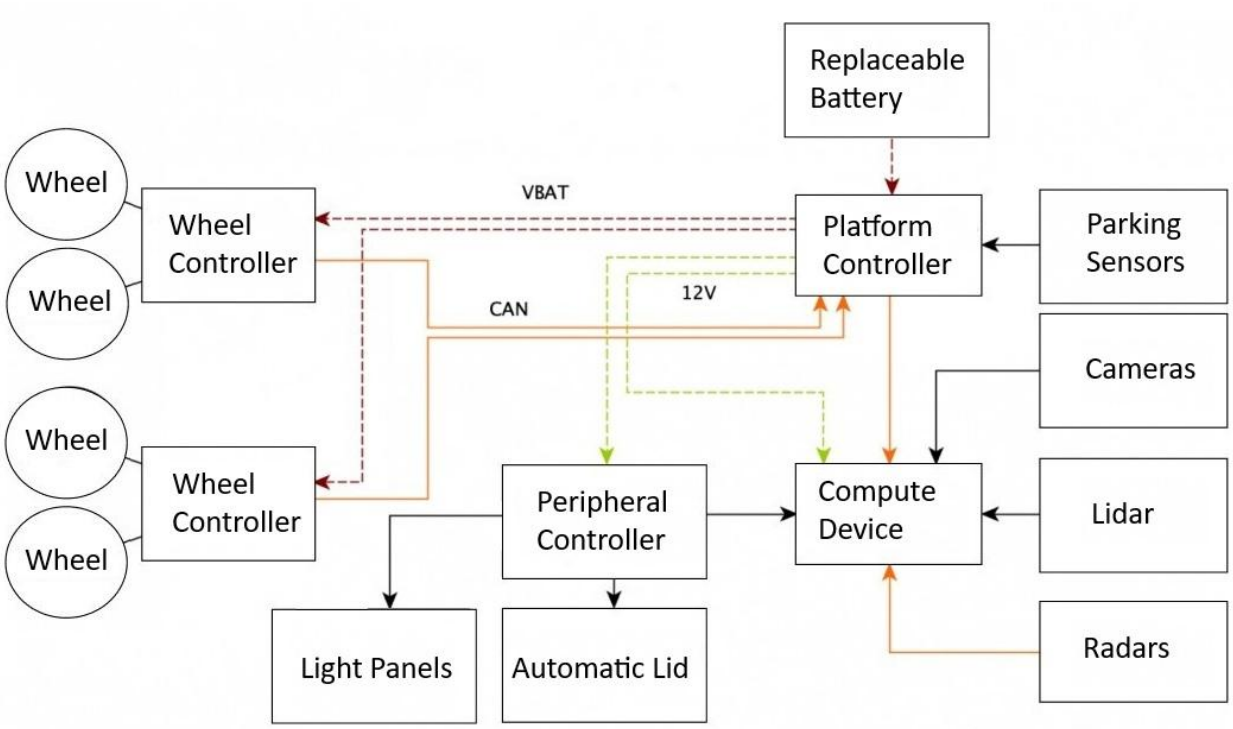


Fig. 1. The structural diagram of the vehicle under test.

Automation of the testing process can eliminate human error, enhance transparency, ensure consistent and controlled testing conditions, and allow for efficient data collection and analysis. To address these challenges, new testing methods applicable to unmanned delivery vehicles were developed. These methods ensure reliable performance in real-world conditions while meeting the demands for evaluation.

The paper is organized as follows: the related works in Section II reviews existing literature on testing systems and highlights the challenges. The methodologies in Section III details the proposed testing framework. The results in Section IV presents findings from implementing the system, focusing on performance metrics. The discussion in Section V analyzes the implications of the results, addressing limitations, and future opportunities. Finally, the conclusion summarizes the study's contributions which is given in Section VI.

II. Related Work

The research in the field of Internet of Things (IoT) and autonomous ground vehicles (AGVs) is rapidly evolving due to their significance in automation and smart transportation. IoT has been widely explored as a foundational technology for the communication and coordination of autonomous systems. Studies like Biswas and Wang emphasize the integration of IoT for data exchange between sensors, vehicles, and control systems, ensuring real-time decision-making and monitoring [1]. Similarly, Baliyan et al. have discussed the role of IoT in enhancing the efficiency and scalability of autonomous delivery vehicles [2]. The approach presented by Abdul Razak et al. [3] demonstrates how IoT-based monitoring can improve vehicle safety, which could be extended to autonomous ground vehicles for monitoring operator impairment in semi-autonomous modes due to alcohol consumption.

Testing methodologies for AGVs have been a significant focus in the literature. Son et al. presented a simulation-based testing framework to validate motion control systems in uncertain environments [4]. Their work highlighted the importance of virtual testing environments to mitigate risks during physical tests. Additionally, Brogle et al. proposed a hardware-in-the-loop (HIL) testing system for autonomous vehicles to evaluate hardware and software interactions under various operational conditions [5].

The implementation of JTAG (Joint Test Action Group) testing for PCB diagnostics has been widely adopted for automated electronic testing, as PCBs are integral to the operation of IoT-enabled systems. Techniques like boundary-scan testing (JTAG) have been explored in works such as Ling et al., which outlined the advantages of automating PCB testing processes [6]. Similarly, Yang et al. emphasized the need for adaptive testing systems to handle the growing complexity of electronic modules in autonomous systems [7].

Functional testing has been explored extensively in autonomous systems to validate their real-world applicability. Autonomous systems rely on accurate and robust sensors, including cameras, LiDARs, and wireless modules. Works by Jernigan et al. have focused on developing rigorous testing methodologies to ensure sensor reliability under varying environmental conditions [8]. Ma et al. further explored the resilience of wireless interfaces in harsh environments, which is critical for maintaining communication in delivery automation [9].

Resilience testing of AGVs and their components under extreme environmental conditions has been a critical focus area. Djoudi et al. proposed a comprehensive testing framework for unmanned delivery vehicles, emphasizing vibration and climate resistance testing [10]. Studies such as Zhang et al. have demonstrated methodologies to assess the durability of electronic modules in harsh climates, including extreme temperatures and vibrations [11]. Using microscopic traffic simulation in VISSIM and the Surrogate Safety Assessment Model (SSAM), Abuzwidah et al. [12] assessed CAV performance across 21 scenarios, highlighting substantial improvements in speed and reductions in accidents under different weather conditions. Their findings emphasize the necessity for CAVs to adapt dynamically to adverse weather for optimal safety. These analyses are crucial in identifying the limitations of functional testing, guiding the development of more effective testing protocols.

Automation in testing processes has become a cornerstone of quality assurance in AGVs. Research by Ostendorff et al. showcased improvements in boundary-scan testing techniques, enabling efficient fault detection in increasingly complex PCBs [13]. Research by Garikapati et al. demonstrated the application of automated AI testing frameworks for vehicle modules, significantly reducing manual effort and errors [14]. Similarly, Jeong et al. highlighted advancements in automated diagnostics for electronic modules, offering faster and more accurate fault detection [15].

Recent works, such as Sánchez-Martinez et al., have moved toward developing integrated testing systems combining hardware, software, and environmental validation [16]. Rahman and Thill reviewed the integration of autonomous vehicles within urban networks, focusing on the performance and the challenges of ensuring consistency [17]. Comparative studies, such as Kim and Kang, have evaluated testing methodologies for EVs, providing a framework to assess their applicability for specific use cases [18].

Advanced sensor systems such as LiDAR and cameras are pivotal in AGVs. Tang et al. examined the testing frameworks required for these sensors, focusing on accuracy, calibration, and environmental adaptability [19]. Studies like Giannaros et al. investigated the performance of wireless modules in urban and rural settings, ensuring seamless data exchange with control systems [20]. These studies emphasize the importance of testing frameworks in mitigating sensor-related failures in autonomous operations, and aligning testing protocols with specific application scenarios.

III. Methodology

The system is a comprehensive methodology for assessing the functional stability of autonomous delivery systems. Fig. 2 illustrates the testing system, which includes a series of specialized procedures aimed at identifying potential vulnerabilities and failures in the system's operation.

The testing process begins with JTAG testing of printed circuit boards, enabling the detection of possible defects in electrical circuits. Functional testing focuses on verifying the proper operation of critical components such as cameras, lidars, and wired and wireless connections.

This is followed by a series of tests for vibration and environmental resistance, designed to evaluate the system's ability to function under varying environmental conditions. Electromagnetic compatibility testing is a critical stage, ensuring that the system can operate without interference from external radiation sources. The delivery system testing framework represents a basic approach to evaluating the quality and reliability of autonomous delivery systems, ensuring their stable performance under diverse operational conditions.

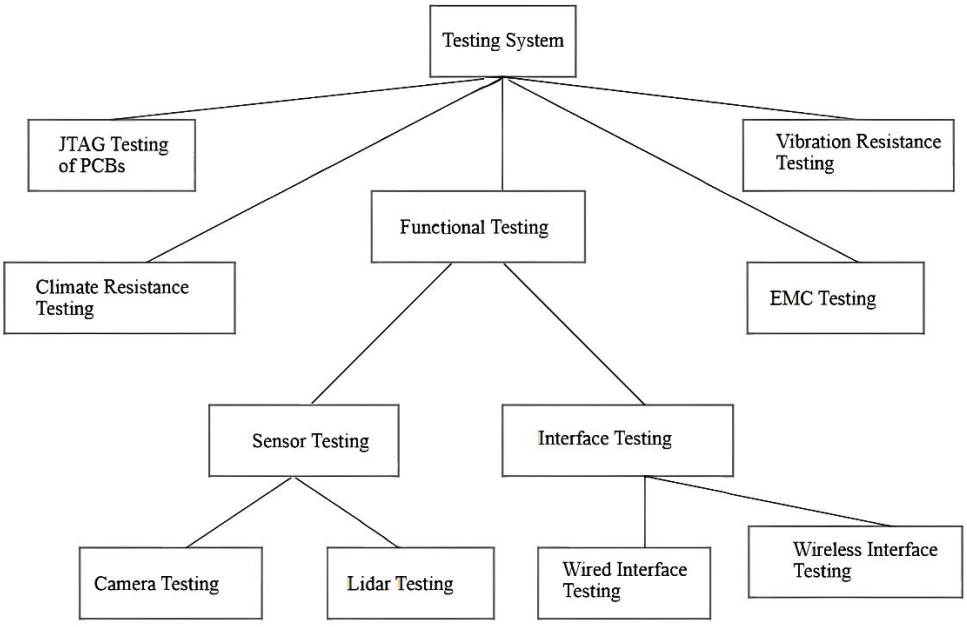


Fig. 2. The plan of the testing system.

A. PCB Testing

The testing of PCBs aims to select the most efficient method for further implementation, emphasizing the necessity of automating the testing process. Connection testing can detect missing pull-up resistors and signal "sticking" issues. This is achieved by setting specific values on the pins and comparing the read values against a predefined truth table.

Visual inspection and manual testing involve checking the quality of assembly, the presence and integrity of all components, and performing measurements using a multimeter. However, the increasing complexity of PCBs and the risk of human error reduce the efficiency of this method. An in-circuit tester (matrix testing) uses fixed sensor probes to check the integrity of soldered connections (Fig. 3 and Fig. 4). The disadvantages of this method include the high cost of the testing equipment, its large size, and the need to create a customized matrix contact field according to the PCB design.

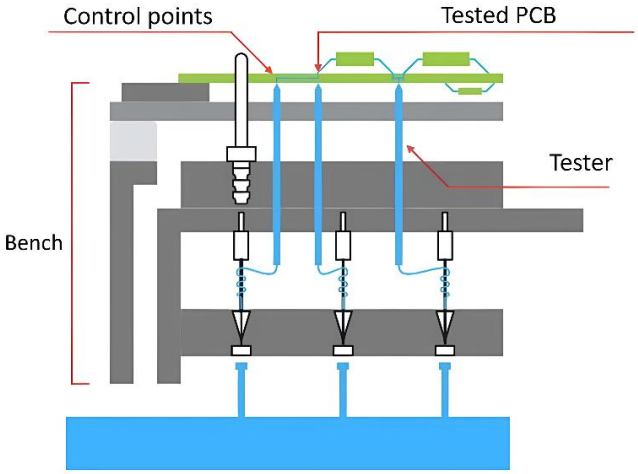


Fig. 3. Example of an in-circuit tester.

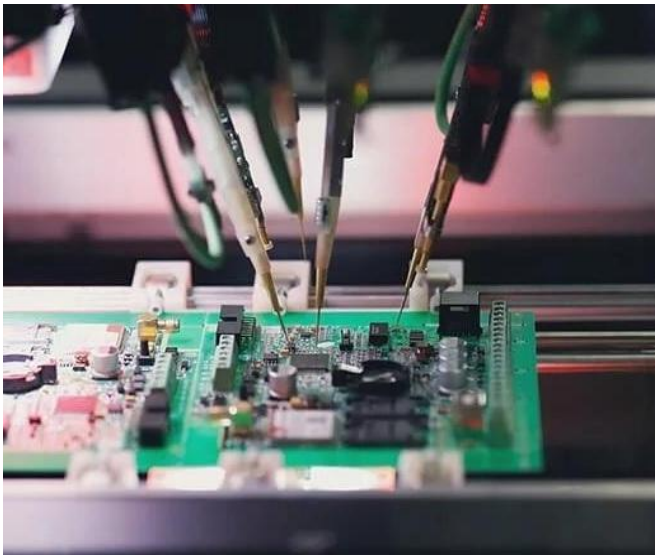


Fig. 4. Example of a flying grid tester.

JTAG testing, or connection testing, verifies whether the manufactured PCB matches the original design and identifies unintended circuit breaks or shorts [21]. For example, if the design specifies that certain chip pins must be connected somewhere on the board, the presence of the connection can be checked by applying values to one pin and reading them from others. Similarly, if the design specifies that certain pins should not be connected, JTAG testing can verify the absence of unexpected shorts by applying values to one pin and ensuring they do not influence others. The hardware of an autonomous delivery system can be tested in various ways, Table I provides more details of a comparison of PCB testing methods.

TABLE I. COMPARISON OF PCB TESTING METHODS

Method	*Error Probability*	*Performance*	*Retooling*	*Fixture Development*
Manual Testing	High	Low	Simple	Not Required
Matrix Tester	Low	High	Complex	Required
Flying grids	Low	High	Complex	Not Required
JTAG Testing	Low	High	Simple	Not Required

Thus, in the case of development and production involving numerous PCB designs manufactured in small batches, suitable testing options include purchasing a flying grids or utilizing JTAG testing. Both methods offer low error probability, high performance, and do not require the development of specialized fixtures. However, considering the cost and time required for reconfiguring the flying grids for each board, this work prioritizes JTAG testing.

B. JTAG Testing

To test a PCB using boundary scan, a BSDL (Boundary Scan Description Language) file must be downloaded from the chip manufacturer's website for each JTAG-supported chip. The supplementary text file describes the functions of the chip's pins.

The main advantage of boundary scan technology is the ability to set and read values at the pins without direct physical access. All signals between the device's core logic and its pins are intercepted by a serial scan path known as the Boundary Scan Register (BSR), which consists of a series of boundary scan cells. These cells are invisible during normal operation but can be used in test mode to set and/or read values from the device's pins or, in some cases, from the internal core logic. There are ten standard types of boundary scan cells, although manufacturers can define custom cell types to suit their hardware's functionality. The JTAG interface uses the following signal lines:

- TCK (Test Clock): To synchronize the internal operations of the state machine.
- TMS (Test Mode Select): Determines the next state of the state machine based on the rising edge of TCK.
- TDI (Test Data In): Represents data sent to the device's testing or programming logic. It is sampled on the rising edge of TCK when the state machine is in the correct state.

- TDO (Test Data Out): Represents data output from the device's testing or programming logic. It is valid on the falling edge of TCK when the state machine is in the correct state.
- TRST (Test Reset): An optional line that, if available, resets the TAP controller state machine.

JTAG is a synchronous interface, where signals are sampled on the rising edge of the clock (TCK) with the least significant bits first, and data output occurs on the falling edge. Boundary scan testing accelerates the preparation of tests for each project and eliminates the need for expensive test equipment. Furthermore, JTAG boundary scan can identify the precise location of faults, significantly simplifying diagnostics and repair.

With the increasing use of BGA (Ball Grid Array) packages, traditional PCB testing systems face limitations due to the inaccessibility of "internal" contacts. Boundary scan reduces test development costs by simplifying the management of chip pins for interaction with other board components. The standardized JTAG interface also allows individual tests to be created as library elements and reused across different projects, regardless of the JTAG-supported chips used. It is frequently used for programming chips on the board during production. When combined with boundary scan testing, this approach can save significant time and streamline the manufacturing process.

C. Functional Testing

To automate the testing process, it was necessary to choose a programming language and framework that would enable the development of a testing system with maximum simplicity and minimal programming expertise required for writing tests. Python was selected as the programming language due to its simplicity, flexibility, and widespread adoption.

In modern automated testing with Python, various testing frameworks are used. The most popular ones include:

- PyTest and PageObject;
- OpenHTF;
- Robot Framework.

The PyTest framework and the PageObject pattern allow separation of test logic from implementation, simplifying test management. However, they require significant effort during the initial development stage due to the need for low-level descriptions of testing processes and conditions for passing tests. While PyTest is excellent for integration and system testing, it may be excessive and less convenient for specific hardware board testing compared to Robot Framework. The OpenHTF library, developed by Google, was also considered. It includes a built-in graphical interface but has significant drawbacks, such as a lack of documentation, necessitating source code study, and the absence of academic work utilizing the framework.

Robot Framework offers several key advantages over other testing frameworks [22]. Its syntax is highly human-readable, making it easier to write and maintain tests, which is particularly beneficial in robotic development involving multi-disciplinary teams of hardware and software engineers. It includes built-in mechanisms for parallel test execution, detailed reporting, and logging, facilitating test analysis and statistical data collection. The framework also provides flexible mechanisms for Python integration, enabling the use of Python libraries during testing. Additionally, it allows the addition of new libraries and plugins for managing specific hardware and protocols. Thus, it is the most suitable choice for this task, offering a balance between test readability and maintainability.

To implement the delivery module testing system, readily available components were chosen, as they allow for quick hypothesis testing and reduce the development cost of the testing system. The Raspberry Pi Model 3 B+ was selected as the testing board due to its features, including an expanded 40-pin GPIO (General Purpose Input/Output) connector ideal for device testing (see Fig. 5). Additionally, the Raspberry Pi includes a CSI camera port, USB 2.0 ports, and a Micro SD slot for operating system booting and data storage. All tests conducted with the system's hardware are performed by the test bench operator.

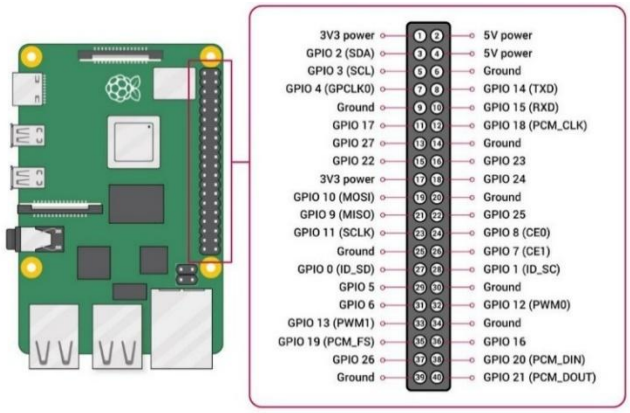


Fig. 5. External GPIO pins on rPi.

D. Testing Sensors

The Raspberry Pi 3 Model B+ Camera Module with a 5 MP resolution was chosen as the test camera because it connects via the CSI (Camera Serial Interface), similar to the cameras used in delivery vehicles, as shown in Fig. 6. This camera is classified as a MIPI (Mobile Industry Processor Interface) camera and connects directly to the VideoCore video chip through a CSI-2 (Camera Serial Interface-2) port, which helps conserve the Raspberry Pi's system resources, leaving USB ports available for other peripherals.

To test the camera's functionality, it must be confirmed that the camera can capture an image directly from the Raspberry Pi board. This involves configuring the camera in the system and installing Python libraries such as picamera for camera access and pillow for image processing. A test case was then created with the following logic: an image is captured using the camera, saved to the operating system as a file, then loaded and validated (ensuring the file is a valid image). Based on this validation, a report is generated indicating whether the test passed (Pass) or failed (Fail).

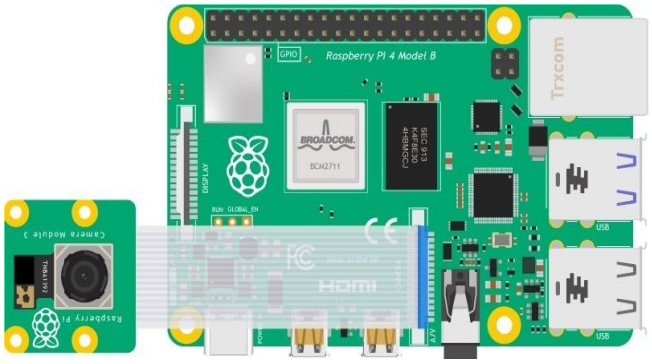


Fig. 6. The raspberry Pi 3 Model B+ camera module.

LiDARs are used for scanning and mapping the environment, aiding autonomous delivery vehicles in navigating safely. However, the diversity of LiDAR types and their operation in varying conditions complicates the development of universal testing procedures. Delivery vehicles typically use mechanical LiDARs. These feature a laser emitter (usually at a 905 nm wavelength) and a photodetector mounted on a rotating platform. This high pulse density enables the LiDAR to generate a visual 3D map of the surrounding area using a cloud of reflected points. The denser the laser pulses, the more detailed the point cloud.

Accelerometers in delivery vehicles measure acceleration along the X, Y, and Z axes. The accelerometer connects to the control electronics through the following pins:

- Power (V): Connected to the microcontroller's operating voltage.
- Ground (G): Connected to the microcontroller's ground.
- Data Signal (D): The data pin for the I²C bus, connected to the microcontroller's SDA pin.
- Clock Signal (C): The clock pin for the I²C bus, connected to the microcontroller's SCL pin.

To read values accurately from the accelerometer, the data for each axis is stored in high and low bytes. These must be combined into a single 16-bit value by performing a bit shift operation and adding the low byte to the shifted high byte.

In this section, functional testing was conducted to verify the operation of key sensors in the autonomous delivery vehicle. For instance, for testing the Wi-Fi wireless interface, a test was implemented to connect to a network using the Linux Network Manager, the most popular system for managing network connections on Linux systems. The next step involves testing the developed automated test cases.

IV. Results

A. Testing Boards

Before starting the test, ensure that the STM32F401RE board is connected to the Raspberry Pi 3. Begin by reading and verifying the board's ID. Next, test the two microcontroller pins, PA5 and PA6. Set PA5 on the STM32 to 1 (HIGH) and read the state of the corresponding pin on the Raspberry Pi. Verify that PA5 is set to 1, then similarly set and verify 0 (LOW). Repeat the same process for PA6. Finally, upload the firmware to the board and confirm that it has been successfully written, as shown in Fig. 7.

The UrJTAG utility is used for JTAG testing, allowing direct interaction with boards through the GPIO pins of the Raspberry Pi. Custom keywords were defined to structure the tests and avoid code duplication, such as opening and closing the JTAG interface and retrieving the state of a GPIO pin. The interaction with API libraries is considered but it also allows for precise fault localization, simplifying diagnostics and repair; however, its reliance on boundary scan capabilities limits its applicability to boards designed with JTAG support, potentially excluding legacy systems or simpler PCBs without such interfaces.

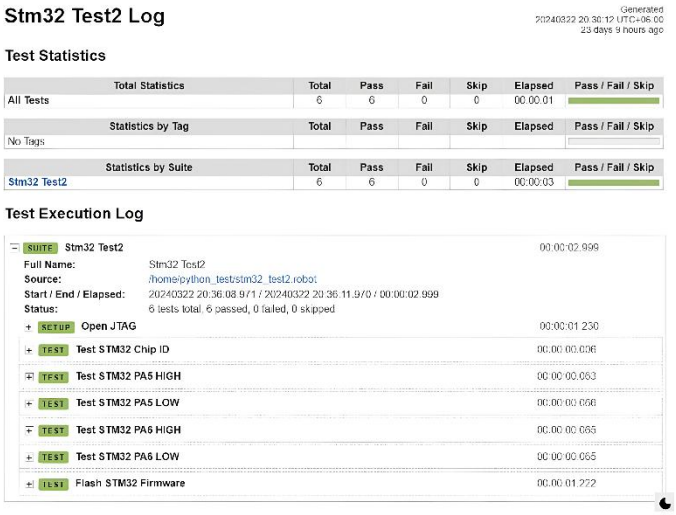


Fig. 7. JTAG testing log file.

Robot Framework tests are written in files with the .robot extension. These files use a BDD-like syntax, and the test file is named stm32_test2.robot. Test cases and keywords from the stm32_test.robot file interact with libraries by using methods belonging to those libraries.

- The jtag.py file is a library designed for working with the JTAG interface.
- The init method initializes the process for JTAG operations and sets non-blocking reading mode.
- The _set_nonblock method configures the stdout read descriptor to non-blocking mode for asynchronous interaction.
- The send method sends commands to the JTAG process, ensuring data transmission and calling flush for immediate delivery.
- The recv method reads data from the JTAG process, handles potential errors, and returns a tuple containing the execution status and response text.
- The bsdl, set_extest, and set_signal methods are used for configuring the BSDL, switching to EXTEST mode, and setting a signal on a pin, respectively.

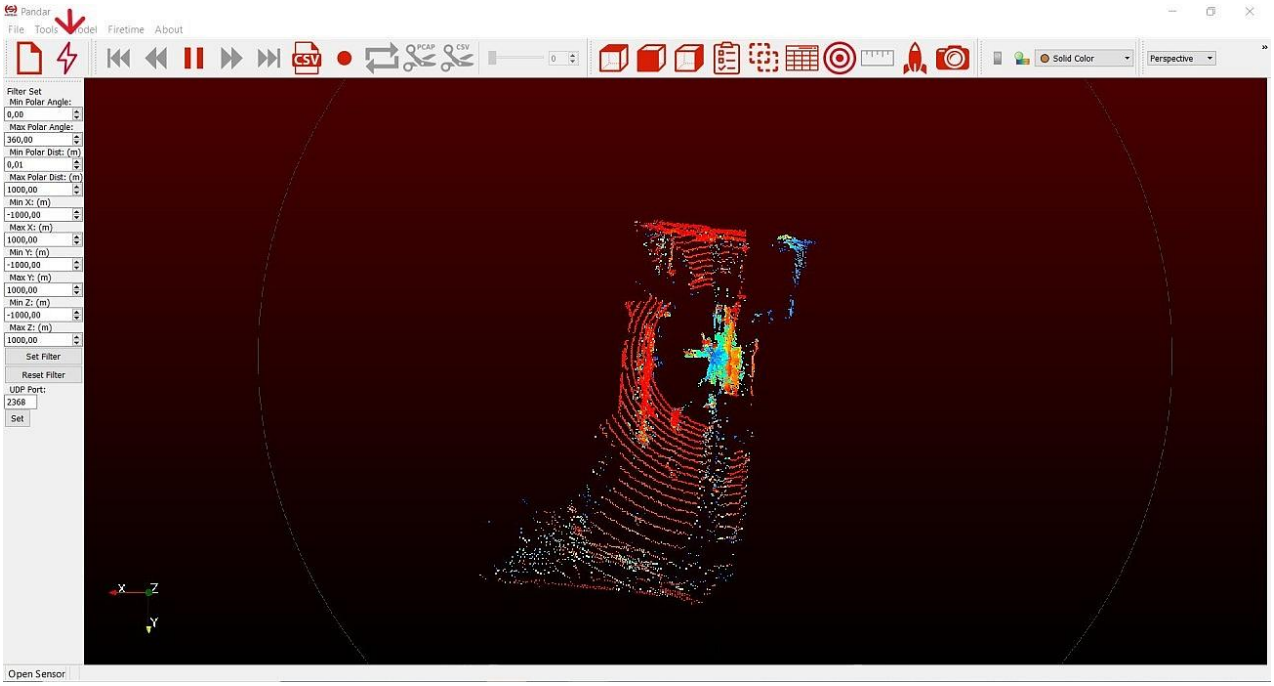


Fig. 8. LiDAR point cloud.

The detect_id method checks whether a device with a specified identifier (ID) is connected. It sends a detect command to the JTAG process and analyzes the response to determine if the specified ID is listed among the detected devices. The quit method ensures the proper termination of the JTAG process. It sends the quit command to the process and waits for it to exit. Python-based frameworks like Robot Framework has proven effective for evaluating key components such as cameras, LiDARs, and accelerometers; however, while it supports modular testing, it requires careful configuration and additional effort for integrating new device-specific libraries.

B. Testing Functional Parts

The LiDAR under test should be inspected for external mechanical damage and the condition of its lens. Prepare the necessary equipment:

- LiDAR;
- Laptop or PC with Ethernet, equipped with software for viewing the point cloud (e.g., PandarView);
- HESAI Interface Box;
- HESAI power supply;
- Ethernet patch cord.

Open PandarView and click "Receive Data from Ethernet." A point cloud should appears as in Fig. 8. This testing methodology was developed to verify the LiDAR's functionality and ensure the proper operation of its mechanical components, including the laser receiver and emitter.

We ensured that the LiDAR is not vibrating. Strong vibrations and low-frequency mechanical oscillation sounds can indicate issues with the bearings. Testing under poor visibility conditions should include assessments in fog, rain, and snowfall. A fog machine can be used to simulate these weather conditions. The greater the number of laser pulses, the denser the LiDAR point cloud. Based on various point clouds, the autonomous delivery vehicle's computational system constructs objects that form a three-dimensional representation of the surrounding environment.

Using a vibration test bench is a more effective method for assessing vibration resistance compared to driving over various road surfaces, as it allows testing under controlled, consistent conditions. Therefore, the parameters of the vibration test bench must be calculated to ensure optimal testing. Testing for mechanical factors, particularly vibration, requires preliminary calculations to select the most suitable equipment in terms of technical specifications and cost-effectiveness.

The vibration test bench is the core component and actuator of the vibration system, reproducing a specific type of vibration and transmitting it to the test object. Tests are typically conducted on a vibration stand equipped with one or more shakers, which register the sample's response to a predefined vibrational load. The market for vibration test benches is extensive and includes a wide variety of models. As a result, decisions often lean towards purchasing the most powerful vibration test bench available. For this purpose, the Tira vibration test bench, as shown in Fig. 9, with a thrust force of up to 32k, is selected as it meets the necessary requirements.

Fig. 9. Vibration test bench.

Thus, the developed system has successfully passed all tests and is suitable for testing real devices. However, for comprehensive testing of an autonomous delivery vehicle, it is insufficient to assess only its electronic modules using JTAG testing and functional tests. It is essential to conduct vibration resistance testing to evaluate the durability of electronic, electrical, and mechanical modules under vibrational stress. This is crucial as delivery vehicles may encounter potholes, gravel, and cobblestones, which impose vibrational and impact loads that could lead to the failure of individual modules or the entire vehicle [23]. The results of the developed automated tests are presented in Table II.

TABLE II. TEST CASES OF THE DEVELOPED SYSTEM

Test Name	*Steps*	*Expected Result*	*Test Passed*
All pins connected	Connect the test board to the test bench. Run. Wait for the test to complete.	Pass	Yes
Camera test (connected)	Connect the camera to Raspberry Pi. Run the test.	Pass	Yes
One pin disconnected	Disconnect PA5, run test, check result.	Fail: PA5 HIGH $0 \neq 1$	Yes
Camera test (disconnected)	Disconnect camera, run test, check result.	Fail: Photo not captured	Yes
Accelerometer test	Connect accelerometer, run test, check result.	Pass	Yes
WiFi test (disconnected)	Disable network device, run test, check result.	Fail	Yes
Vibration test	Place device on Tira, run Tira, visual check.	No cracks	Yes
EMC test	Execute EMC protocol, check connetion	Device stays connected	Yes

Environmental conditions are typically described using statistical variables such as temperature, humidity, air quality, and so on. These factors are critical for the functionality and lifespan of delivery vehicles, and manufacturers must ensure that their modules can operate within specified conditions while maintaining their stated performance characteristics.

Most climate tests are conducted in a climate chamber, as shown in Table III, which can simulate changes in temperature, humidity, dew, and frost. Additional conditions, such as dust and solar radiation exposure, are tested using dedicated chambers. Commonly measured parameters for determining the electrical safety of delivery vehicle modules include current, voltage, leakage path length between conductors, and the energy of emitted waves. It is crucial to note that electrostatic discharge (ESD) is unacceptable for electronic modules when they are not yet enclosed in a protective casing.

TABLE III. THE RESULTS FOR CLIMATE CHAMBER TESTS

Test Name	*Testing Methodology*	*Result*
Temperature Testing	Climate chamber, temperature range: -40°C to +85°C; condensation cycles and frost formation.	Most components operate within -20°C to +60°C; failures occur at ±70°C; frost impacts 30% of components
Humidity Testing	Controlled humidity chamber, 10%–95% RH	Corrosion risk increases by 40% at 85% RH; short circuits above 90% RH; risk varies by material
Dust Exposure Testing	Dust chambers with varying particle size (ISO 12103-1), Arizona test dust	Dust penetration in enclosures above 5-micron particles; 20% degradation in 3 weeks
Solar Radiation Testing	UV and infrared radiation exposure tests	UV exposure causes 15% material degradation over 1000 hours; discoloration starts at 500 hours
Electrical Safety Testing	High-voltage insulation resistance testing	Leakage current below 1mA at 1000V; insulation resistance >10MΩ meets IEC standards
Electrostatic Discharge (ESD) Testing	ESD simulator testing at 2kV–15kV discharge levels	Devices withstand up to 10kV discharge; failures start above 12kV

Once the delivery vehicle is fully assembled, it must withstand electrostatic discharge, as all electronic components are securely shielded by the casing. However, the reliance on specialized equipment introduces significant costs and operational constraints, especially for small-scale manufacturers; additionally, vibration testing under controlled conditions may not fully replicate the complexities of real-world terrain.

V. DISCUSSION

The developed testing system demonstrates significant advancements in ensuring the reliability and functional stability of autonomous delivery systems. By combining JTAG testing, functional testing, and environmental assessments, the system provides a comprehensive framework for identifying vulnerabilities and verifying the performance of key modules under diverse conditions. However, while the system achieves its primary objectives, there are opportunities for enhancement and certain limitations to address.

One notable strength of the system lies in its use of JTAG testing, which offers a highly efficient and cost-effective method for verifying PCB integrity. Unlike manual or in-circuit testing methods, JTAG testing eliminates the need for expensive test fixtures and reduces human error. It also allows for precise fault localization, simplifying diagnostics and repair. Although, its reliance on boundary scan capabilities restricts its applicability to boards designed with JTAG

support, potentially excluding legacy systems or simpler PCBs without such interfaces [24].

Functional testing, implemented through Python-based frameworks like Robot Framework, has proven effective for evaluating key components such as cameras, LiDARs, and accelerometers. The framework's human-readable syntax and Python integration streamline test development and maintenance, making it accessible to multidisciplinary teams. While Robot Framework supports modular testing, it requires careful configuration and additional effort for integrating new device-specific libraries, which may pose challenges for teams with limited resources or expertise.

Environmental testing introduces another layer of robustness by simulating real-world conditions in controlled environments [25]. The use of climate chambers for temperature, humidity, and frost testing, alongside vibration benches for mechanical stress evaluation, provides valuable insights into the durability of delivery systems. Nevertheless, the reliance on specialized equipment, such as the Tira vibration test bench, introduces significant costs and operational constraints, especially for small-scale manufacturers. Additionally, while vibration testing under controlled conditions is highly effective, it may not fully replicate the complexities of real-world terrain [26].

The inclusion of sensor-specific tests, such as those for LiDARs, camera, and accelerometers, highlights our focus on the core functionality of autonomous vehicles. Nevertheless, the diversity of sensor technologies and operational environments poses a challenge to the development of universal testing procedures [27]. For example, mechanical LiDARs with rotating components require different calibration and durability tests compared to solid-state LiDARs. Similarly, environmental factors like fog or heavy rain can disproportionately affect sensor performance, requiring further refinements to testing methodologies. The system also incorporates EMC testing to ensure devices can operate without interference [28]. This is particularly important for delivery vehicles, which rely on seamless communication between components. While the system's EMC tests have been effective, integrating real-time data collection and analysis during such tests could provide additional insights and improve overall reliability.

Despite these strengths, the system has limitations in scalability and regulatory compliance. The absence of standardized testing protocols for autonomous delivery vehicles means that manufacturers may face challenges in aligning their testing processes with emerging regulatory requirements. Furthermore, the system's reliance on high-performance test equipment, such as climate chambers and vibration benches, may not be feasible for all manufacturers, especially those operating in cost-sensitive markets.

Opportunities for improvement include the integration of machine learning algorithms to optimize test procedures and predict potential failures based on historical data [29]. Additionally, developing lightweight and portable testing solutions could reduce costs and improve accessibility for smaller manufacturers. The system could focus on expanding its adaptability and automation capabilities. Automated test execution and data analysis could significantly reduce the time and effort required for repetitive testing tasks, particularly for high-volume production scenarios. For example, integrating automated tools for capturing and analyzing LiDAR point clouds or accelerometer data could provide deeper insights into module performance under specific conditions.

Moreover, the testing framework could be extended to include real-time monitoring and diagnostics during operational testing. This critical opportunity lies in the development of modular testing architectures. Modular designs would also support scalability, allowing the testing framework to be adapted to different vehicle types or system configurations without significant reconfiguration. This would allow for dynamic adjustments in testing parameters, ensuring that modules are evaluated under a broader range of conditions, including unexpected environmental factors or system interactions [30]. Such an approach could also help detect intermittent faults that might not appear under standard test scenarios.

The incorporation of cloud-based testing and analytics could further enhance the system's capabilities [31]. A centralized platform for storing, analyzing, and sharing test results would enable manufacturers to benchmark performance across multiple production cycles or facilities. Additionally, cloud integration could facilitate collaborative development of standardized testing methodologies, allowing manufacturers to align their processes with industry best practices and emerging regulatory standards.

While the current framework emphasizes hardware testing, extending the scope to include software validation would provide a more holistic approach to system reliability. Autonomous delivery vehicles rely heavily on complex algorithms for navigation, object detection, and decision-making [32]. Testing these algorithms in simulated environments that mimic real-world scenarios could complement the hardware testing process, ensuring seamless integration and overall system robustness.

Finally, addressing regulatory alignment remains a critical area for improvement. By engaging with industry stakeholders and regulatory bodies, the system could be tailored to meet specific compliance requirements, paving the way for broader adoption in global markets [33]. Collaboration with regulatory bodies to define standardized testing frameworks could enhance the system's applicability and acceptance in the industry. Such efforts would also help establish the system as a benchmark for testing autonomous delivery vehicles, contributing to the standardization of quality assurance practices in this rapidly evolving field [34].

Overall, the developed system effectively addresses many challenges associated with testing autonomous delivery systems, further innovations in automation, modularity, and regulatory alignment will unlock new possibilities. By embracing these opportunities, the framework has the potential to become a solution for ensuring the safety, reliability, and functionality of autonomous delivery vehicles in diverse operational environments.

VI. Conclusion

The developed testing system provides a robust framework for assessing the functionality and reliability of autonomous delivery systems, combining JTAG testing, functional evaluations, and environmental assessments. It ensures comprehensive testing of critical components, including PCBs, sensors, and wireless interfaces, while maintaining cost-effectiveness through automation and streamlined processes.

However, the system has limitations. The reliance on specialized equipment like vibration benches and climate chambers can be cost-prohibitive for smaller manufacturers. Additionally, the absence of standardized protocols for autonomous delivery vehicles limits its regulatory alignment, and the diversity of sensor technologies complicates the development of universal testing methods.

Future work should focus on enhancing scalability and adaptability through modular and portable testing setups, integrating machine learning for predictive diagnostics, and expanding the framework to include software validation alongside hardware testing. Collaboration with regulatory bodies to establish standardized testing protocols and incorporating real-time data analytics will further strengthen the system's applicability and industry relevance.

References

[1] A. Biswas and H.-C. Wang, "Autonomous Vehicles Enabled by the Integration of IoT, Edge Intelligence, 5G, and Blockchain," Sensors, vol. 23, no. 4, p. 1963, 2023. doi: 10.3390/s23041963.

[2] A. Baliyan, J. S. Dhatterwal, K. S. Kaswan, and V. Jain, "Role of AI and IoT Techniques in Autonomous Transport Vehicles," in AI Enabled IoT for Electrification and Connected Transportation, Transactions on Computer Systems and Networks. Singapore: Springer, 2022, pp. 1–16. doi: 10.1007/978-981-19-2184-1_1.

[3] S. F. A. Razak, S. Yogarayan, and A. Ullah, "Preventing Impaired Driving Using IoT on Steering Wheels Approach," HighTech and Innovation Journal, vol. 5, no. 2, 2024. doi: 10.28991/HIJ-2024-05-02-012.

[4] T. Duy Son, A. Bhave and H. Van der Auweraer, "Simulation-Based Testing Framework for Autonomous Driving Development," 2019 IEEE International Conference on Mechatronics (ICM), Ilmenau, Germany, 2019, pp. 576-583, doi: 10.1109/ICMECH.2019.8722847.

[5] C. Brogle, C. Zhang, K. L. Lim and T. Bräunl, "Hardware-in-the-Loop Autonomous Driving Simulation Without Real-Time Constraints," in IEEE Transactions on Intelligent Vehicles, vol. 4, no. 3, pp. 375-384, Sept. 2019, doi: 10.1109/TIV.2019.2919457.

[6] Q. Ling and N. A. M. Isa, "Printed Circuit Board Defect Detection Methods Based on Image Processing, Machine Learning and Deep Learning: A Survey," in IEEE Access, vol. 11, pp. 15921-15944, 2023, doi: 10.1109/ACCESS.2023.3245093.

[7] J. Yang, R. Bai, H. Ji, Y. Zhang, J. Hu, and S. Feng, "Adaptive Testing Environment Generation for Connected and Automated Vehicles with Dense Reinforcement Learning," arXiv, vol. 2402.19275, Feb. 2024. doi: 10.48550/arXiv.2402.19275.

[8] M. Jernigan, S. Alsweiss, J. Cathcart and R. Razdan, "Conceptual Sensors Testing Framework for Autonomous Vehicles," 2018 IEEE Vehicular Networking Conference (VNC), Taipei, Taiwan, 2018, pp. 1-4, doi: 10.1109/VNC.2018.8628370.

[9] Z. Ma, M. Xiao, Y. Xiao, Z. Pang, H. V. Poor and B. Vucetic, "High-Reliability and Low-Latency Wireless Communication for Internet of Things: Challenges, Fundamentals, and Enabling Technologies," in IEEE Internet of Things Journal, vol. 6, no. 5, pp. 7946-7970, Oct. 2019, doi: 10.1109/JIOT.2019.2907245.

[10] A. Djoudi, L. Coquelin and R. Régnier, "A simulation-based framework for functional testing of automated driving controllers," 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), Rhodes, Greece, 2020, pp. 1-6, doi: 10.1109/ITSC45102.2020.9294454.

[11] Y. Zhang, A. Carballo, H. Yang, and K. Takeda, "Perception and sensing for autonomous vehicles under adverse weather conditions: A survey," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 196, pp. 146–177, Feb. 2023. doi: 10.1016/j.isprsjprs.2022.12.021.

[12] M. Abuzwidah, A. Elawady, L. Wang, and W. Zeiada, "Assessing the Impact of Adverse Weather on Performance and Safety of Connected and Autonomous Vehicles," Civil Engineering Journal, vol. 10, no. 9, 2024. doi: 10.28991/CEJ-2024-010-09-019.

[13] S. Ostendorff, J. Sachsse, H. Wuttke, and J. Meza Escobar, "Adaptive Test System to Improve PCB Testing in the Automotive Industry," SAE Int. J. Passeng. Cars – Electron. Electr. Syst., vol. 6, no. 1, pp. 294–300, 2013. doi: 10.4271/2013-01-1230.

[14] D. Garikapati and S. S. Shetiya, "Autonomous Vehicles: Evolution of Artificial Intelligence and the Current Industry Landscape," Big Data Cogn. Comput., vol. 8, no. 4, p. 42, 2024. doi: 10.3390/bdcc8040042.

[15] Y. Jeong, S. Son, E. Jeong, and B. Lee, "An Integrated Self-Diagnosis System for an Autonomous Vehicle Based on an IoT Gateway and Deep Learning," Appl. Sci., vol. 8, no. 7, p. 1164, 2018. doi: 10.3390/app8071164.

[16] R. Sánchez-Martinez, J. E. Sierra-García, and M. Santos, "Performance and Extreme Conditions Analysis Based on Iterative Modelling Algorithm for Multi-Trailer AGVs," Mathematics, vol. 10, no. 24, p. 4783, 2022. doi: 10.3390/math10244783.

[17] M. M. Rahman and J.-C. Thill, "Impacts of connected and autonomous vehicles on urban transportation and environment: A comprehensive review," Sustainable Cities and Society, vol. 96, p. 104649, Sep. 2023. doi: 10.1016/j.scs.2023.104649.

[18] B. Kim and E. Kang, "Toward Large-Scale Test for Certifying Autonomous Driving Software in Collaborative Virtual Environment," in IEEE Access, vol. 11, pp. 72641-72654, 2023, doi: 10.1109/ACCESS.2023.3295500.

[19] S. Tang, Z. Zhang, Y. Zhang, J. Zhou, Y. Guo, S. Liu, S. Guo, Y.-F. Li, L. Ma, Y. Xue, and Y. Liu, "A Survey on Automated Driving System Testing: Landscapes and Trends," ACM Trans. Softw. Eng. Methodol., vol. 32, no. 5, Art. 124, p. 62, Sep. 2023. doi: 10.1145/3579642.

[20] A. Giannaros, A. Karras, L. Theodorakopoulos, C. Karras, P. Kranias, N. Schizas, G. Kalogeratos, and D. Tsolis, "Autonomous Vehicles: Sophisticated Attacks, Safety Issues, Challenges, Open Topics, Blockchain, and Future Directions," J. Cybersecur. Priv., vol. 3, no. 3, pp. 493–543, 2023. doi: 10.3390/jcp3030025.

[21] F. Zhang, S. D. Paul, P. Slpsk, A. R. Trivedi and S. Bhunia, "On Database-Free Authentication of Microelectronic Components," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 29, no. 1, pp. 149-161, Jan. 2021, doi: 10.1109/TVLSI.2020.3039723.

[22] L. Jian-Ping, L. Juan-Juan and W. Dong-Long, "Application Analysis of Automated Testing Framework Based on Robot," 2012 Third International Conference on Networking and Distributed Computing, Hangzhou, China, 2012, pp. 194-197, doi: 10.1109/ICNDC.2012.53.

[23] J. Hu, T. Xu and R. Zhang, "Testing and Evaluation of Autonomous Vehicles Based on Safety of the Intended Functionality," 2021 6th International Conference on Transportation Information and Safety (ICTIS), Wuhan, China, 2021, pp. 1083-1086, doi: 10.1109/ICTIS54573.2021.9798586.

[24] E. Tramacere, S. Luciani, S. Feraco, A. Bonfitto, and N. Amati, "Processor-in-the-Loop Architecture Design and Experimental Validation for an Autonomous Racing Vehicle," Applied Sciences, vol. 11, no. 16, p. 7225, 2021. doi: 10.3390/app11167225.

[25] N. Assymkhan and A. Kartbayev, "Advanced IoT-Enabled Indoor Thermal Comfort Prediction Using SVM and Random Forest Models" International Journal of Advanced Computer Science and Applications (IJACSA), 15(8), 2024. doi:10.14569/IJACSA.2024.01508102.

[26] Z. Tahir and R. Alexander, "Coverage based testing for V&V and Safety Assurance of Self-driving Autonomous Vehicles: A Systematic Literature Review," 2020 IEEE International Conference On Artificial

Intelligence Testing (AITest), Oxford, UK, 2020, pp. 23-30, doi: 10.1109/AITEST49225.2020.00011.

[27] M. M. Long, T. T. Diep, S. H. Needs, M. J. Ross, and A. D. Edwards, "PiRamid: A compact Raspberry Pi imaging box to automate small-scale time-lapse digital analysis, suitable for laboratory and field use," HardwareX, vol. 12, Oct. 2022. doi: 10.1016/j.ohx.2022.e00377.

[28] F. Nussipova, S. Rysbekov, Z. Abdiakhmetova, and A. Kartbayev, "Optimizing loss functions for improved energy demand prediction in smart power grids," International Journal of Electrical and Computer Engineering (IJECE), vol. 14, no. 3, pp. 3415–3426, 2024. do: 10.11591/ijece.v14i3.pp3415-3426.

[29] D. Bruggner, A. Hegde, F. S. Acerbo, D. Gulati and T. D. Son, "Model in the Loop Testing and Validation of Embedded Autonomous Driving Algorithms," 2021 IEEE Intelligent Vehicles Symposium (IV), Nagoya, Japan, 2021, pp. 136-141, doi: 10.1109/IV48863.2021.957553.

[30] R. R. Arany, H. Van der Auweraer and T. D. Son, "Learning Control for Autonomous Driving on Slippery Snowy Road Conditions," 2020 IEEE Conference on Control Technology and Applications (CCTA), Montreal, Canada, 2020, pp. 312-317, doi: 10.1109/CCTA41146.2020.9206260.

[31] L. Shuguang, L. Zhonglin, W. Wenbo, Z. Yang, H. Jierui and C. Hong, "Vehicle-in-the-Loop Intelligent Connected Vehicle Simulation System Based on Vehicle-Road-Cloud Collaboration," 2022 IEEE 25th Intern. Conf. on Intelligent Transportation Systems (ITSC), Macau, China, 2022, pp. 1711-1716, doi: 10.1109/ITSC55140.2022.9922190.

[32] M. Ng, D. Jagetiya, X. Gao, H. Shi, J. Gao and J. Liu, "Real-Time Detection of Objects on Roads for Autonomous Vehicles Using Deep Learning," 2022 IEEE Eighth International Conference on Big Data Computing Service and Applications (BigDataService), Newark, CA, USA, 2022, pp. 73-80, doi: 10.1109/BigDataService55688.2022.00019.

[33] D. Schepis, S. Purchase, D. Olaru, B. Smith, and N. Ellis, "How governments influence autonomous vehicle (AV) innovation," Transportation Research Part A: Policy and Practice, vol. 178, p. 103874, Dec. 2023. doi: 10.1016/j.tra.2023.103874.

[34] T. Sever and G. Contissa, "Automated driving regulations – where are we now?," Transportation Research Interdisciplinary Perspectives, vol. 24, p. 101033, Mar. 2024. doi: 10.1016/j.trip.2024.101033.

AI-Powered Intelligent Speech Processing: Evolution, Applications and Future Directions

Ziqing Zhang
University International College, Macau University of Science and Technology, Macau 999078, China

***Abstract*—This paper provides an overview of the historical evolution of speech recognition, synthesis, and processing technologies, highlighting the transition from statistical models to deep learning-based models. Firstly, the paper reviews the early development of speech processing, tracing it from the rule-based and statistical models of the 1960s to the deep learning models, such as deep neural networks (DNN), convolutional neural networks (CNN), and recurrent neural networks (RNN), which have dramatically reduced error rates in speech recognition and synthesis. It emphasizes how these advancements have led to more natural and accurate speech outputs. Then, the paper examines three key learning paradigms used in speech recognition: supervised, self-supervised, and semi-supervised learning. Supervised learning relies on large amounts of labeled data, while self-supervised and semi-supervised learning leverage unlabeled data to improve generalization and reduce reliance on manually labeled datasets. These paradigms have significantly advanced the field of speech recognition. Furthermore, the paper explores the wide-ranging applications of AI-driven speech processing, including smart homes, intelligent transportation, healthcare, and finance. By integrating AI with technologies like the Internet of Things (IoT) and big data, speech technology is being applied in voice assistants, autonomous vehicles, and speech-controlled devices. The paper also addresses the current challenges facing intelligent speech processing, such as performance issues in noisy environments, the scarcity of data for low-resource languages, and concerns related to data privacy, algorithmic bias, and legal responsibility. Overcoming these challenges will be crucial for the continued progress of the field. Finally, the paper looks to the future, predicting further improvements in speech processing technology through advancements in hardware and algorithms. It anticipates increased focus on personalized services, real-time speech processing, and multilingual support, along with growing integration with other technologies such as augmented reality. Despite the technical and ethical challenges, AI-driven speech processing is expected to continue its transformative impact on society and industry.**

Keywords*—*Intelligent speech recognition; AI speech synthesis; speech processing; AI technology

I. INTRODUCTION

With the iterative upgrading of artificial intelligence technology, the application fields of AI technology are also expanding and developing, and at the same time, intelligent voice interaction, personalized speech generation, and other technologies are getting more and more attention in this process [1]. Among them, speech processing, as an important technology and means of personalized speech generation, involves speech signal processing, artificial intelligence, pattern recognition, phonetics, and other disciplines, and is the hot spot and difficult point in the field of speech processing research, and in the rapid development of AI technology, this field has also become a key research direction. Therefore, along with the development of the Internet and big data technology, voice processing technology (TTS), speech recognition technology (ASR), and other intelligent voice technologies are gradually maturing, the production, dissemination, and storage mechanism of the sound produces changes, and the application scenarios of AI voice are increasing, and voice products such as automobile navigation, video dubbing, intelligent speakers, cell phone assistants and so on are emerging in an endless number, and have already been realized to correspond to specific application Scenarios. Nowadays, there are two main parts of intelligent voice processing according to the function of the application scenes in social life, namely, AI synthesized voice is mainly composed of "AI voice narration" scenes without human-computer interaction, and "AI voice assistant" scenes with human-computer interaction. The two parts of the scene [2].

A. Significance of the Study

Language, as a unique tool for human communication, permeates all aspects of life, contains the rich and important emotional value, and plays an important role in the construction of civilized society, while emotion, as a characteristic unique to human beings, is usually difficult to be accurately conveyed in the process of speech processing. The development of AI technology makes it possible for the part of the speech processing process concerning human-specific expressive characteristics such as emotional value to be learned and applied by the machine, which is expected to have a profound impact on speech processing. Learn and apply, which is expected to have a profound impact on speech processing [3].

Speech, as a direct carrier of language, is the most natural way of communication among human beings and a key part of information transfer in daily life. With the rapid development of artificial intelligence research, speech recognition technology, which is carried by computers, cell phones, tablets, etc., is rapidly advancing. In the field of Human-Computer Interaction (HCI), recognizing individual phonemes or utterances of specific speakers can no longer meet the needs of learning and development of AI technology, so recognizing the hidden emotional state in speech has become a new trend in speech processing research under the rapid development of current AI technology. Therefore, the study of how to use this technology to assist speech processing to achieve intelligence in the current rapid development of AI technology is an urgent problem to be solved nowadays.

B. Conceptual Identification

1) Speech processing: Broadly speaking, people change the voice in the voice of the speaker's personality characteristics of voice processing technology collectively referred to as voice processing, which can be obtained, broad voice processing can be divided into two categories: non-specific person voice processing and specific person voice processing. Non-specific speech processing refers to the technical processing that makes the converted speech no longer sound like the original speaker's voice. In practical research and application, speech processing usually refers to the technology that changes the voice personality characteristics, such as spectrum and rhythm, of one speaker, i.e., the source speaker, to make it have the personality characteristics of another specific speaker, i.e., the target speaker, while keeping the semantic information unchanged. Generally speaking, the technical difficulty of person-specific speech processing is higher than that of non-person-specific speech processing.

The current research on linguistic AI mainly focuses on language understanding and language output, which are called "Natural Language Understanding" (NLU) and "Natural Language Generation" (NLG) respectively. "Natural Language Understanding (NLU) and Natural Language Generation (NLG). Natural Language Understanding allows computers to understand human natural language (including its intrinsic meaning) through a variety of analysis and processing, while Natural Language Generation focuses on how to allow computers to automatically generate natural language forms or systems that humans can understand. The relationship between the research contents of natural language processing and the corresponding human language ability is shown in the Table I.

TABLE I. CORRESPONDENCE TABLE BETWEEN NATURAL LANGUAGE AND HUMAN LANGUAGE ABILITY

No.	Research in Natural Language Processing	Linguistic competence of counterparts
1	Information retrieval	Language Understanding
2	Information filtering	
3	Information extraction	
4	Machine translation	Linguistic Generation
5	Automatic digest	
6	Document classification	Language Understanding
7	Text/Data mining	
8	Public opinion analysis	
9	Text editing and automatic proofreading	
10	Automatic scoring of essays	
11	Question and answer system	Linguistic Generation
12	Metaphorical computing	Language Understanding
13	Optical character recognition, OCR	
14	Speech recognition	
15	Text-to-language conversion	Linguistic Generation
16	Speaker identification/authentication/verification	Language Understanding

Using the degree of development of AI technology as a classification criterion, AI can be categorized into weak AI, strong AI, and super AI. Weak AI, also known as narrow AI. Its main features are: (1) does not have self-consciousness, its action depends on the instructions given by humans; (2) and human beings are significantly different from each other, there is still a huge gap in appearance and other aspects, it is very easy to identify; (3) cannot learn and innovate on its own, it is essentially a tool that relies on human beings to be upgraded. The above characteristics determine that weak artificial intelligence should not be qualified as a legal subject. Strong artificial intelligence, also known as general artificial intelligence, is complete artificial intelligence. Strong AI can think like a human being and is not just a tool for humans. Because it possesses the consciousness of rights, claims of rights, and the ability to realize autonomy, as well as the ability to assume responsibility relatively independently, and has a certain substantive connection with human beings, it meets the conditions for having limited legal subject qualification.

Super Artificial Intelligence, defined by Nick Bostrom as an intelligence that outperforms the human brain in almost every domain, these domains include scientific innovation, general intelligence, and social skills. This definition is open-ended. The possessor of superintelligence can be a digital computer, a computer network integration, or an artificial neural organization without regard to whether it has subjective consciousness or experience. It seems that super-artificial intelligence should be given a higher level of legal status, but this idea is still debatable due to the human fear of the risks posed by the unknown and the adherence to ethics.

Therefore, another commonly accepted classification standard at present is to categorize AI based on different application modes of AI, i.e., to classify AI into generative AI and discriminative AI. According to China's Interim Measures for the Administration of Generative Artificial Intelligence Services, which came into effect on August 15, 2023, generative AI technology refers to models and related technologies that can generate content such as text, pictures, audio, video, and so on. It has become a milestone in the history of AI because it

transcends the scope of traditional AI, is capable of generating natural language understandable to humans, performs tasks that should only be accomplished by human intellectual guidance, and is deeply involved in human daily life. Discriminative AI, also known as discriminative AI or decision-making AI, aims to train machine learning models to classify or predict based on input features. This approach is well suited for tasks such as regression and sequence labeling in areas such as image recognition, speech recognition, and natural language processing. One of the main advantages of discriminative AI is its ability to make efficient and accurate predictions, but it cannot typically generate new samples and an understanding of the generative mechanisms behind the data.

C. Purpose of the Synthesis

This paper will focus on analyzing the latest progress of intelligent speech processing technology in the wave of artificial intelligence technology and the consequent profound changes that will be brought to this research field. With the rapid development of AI technology, intelligent speech processing technology has moved from the laboratory to the road of wide application has gradually penetrated daily life, and has profoundly affected the current life and the mode of operation of various industries. The purpose of this paper is to comprehensively show the technological development of intelligent speech processing technology in terms of speech recognition accuracy, speech processing naturalness, and natural language processing intelligence by systematically combing domestic and foreign research results and application cases, and exploring how the development of AI technology will have a transformative impact on the field of intelligent speech processing, and at the same time, to look forward to the future development of intelligent speech processing. In addition, this paper will also discuss the impact of possible pitfalls in data security, privacy, and law brought about by the ongoing research in the field of intelligent speech processing, to obtain thoughts and inspiration on the impact of the development of AI technology. Overall, the research objective of this paper is to comprehensively sort out the development of intelligent speech processing technology under the development of AI technology, assess its transformative impact on the society and economy, and provide valuable reference and guidance for future research and application.

D. Structure of the Paper

This paper aims to study the development of an intelligent speech-processing technique based on AI technology and to discuss the technical and legal issues that are currently faced as well as the outlook for the future. The following is the Section structure of this thesis:

Section II mainly elaborates on the current status of intelligent speech processing, which firstly gives an overview of the relevant research progress at home and abroad, and then summarizes, combs, and summarizes the current research results at home and abroad, mainly including the overview of the achievements in the fields of speech recognition, speech processing, and speech learning paradigm.

Section III analyzes the key problems faced by the current intelligent speech processing technology based on AI technology, mainly from the main tasks and problems of speech processing technology, firstly, the tasks of the current speech processing technology are clarified, and then the completion of the speech processing tasks are analyzed in-depth by the current problems of speech processing technology, to the problems faced by the intelligent speech processing technology based on AI technology A detailed exposition is carried out.

Section IV describes the development issues in the field of intelligent speech processing based on AI technology, and this Section further explores the public data security and ethical and legal issues brought about by the technological development on top of the technological development issues in Section 3, and this Section analyzes the current intelligent speech processing field outside of the technology that may be encountered in the field of intelligent speech processing, mainly from the perspective of three aspects, namely, data privacy and security, data algorithmic discrimination and bias, and accountability and responsibility problems.

Section V provides an outlook on the future of intelligent speech processing based on the development of current technologies and application scenarios, mainly analyzing and looking forward to both technology and application levels.

Section VI concludes with a summary, which summarizes the significance of the development of intelligent speech processing technology, as well as further summarizes the technological development discussed in the content of the above Sections and the problems that may be faced, and discusses the future research direction, indicating the outlook for the future development of intelligent speech processing technology based on AI technology.

II. The Current Status of Research in the Field of Intelligent Speech Processing

A. Overview of Domestic and International Research

The work related to speech processing research can be traced back to the 1960s and 1970s, and there have been more than 50 years of research history, but it is only in the last decade or so that it has received extensive attention from both academia and industry. In recent years, the advancement of technologies such as speech signal processing and artificial intelligence deep learning, in addition, due to the explosive growth of information, the field of big data has also been developed significantly, and a considerable amount of learning and analyzing text has been obtained, therefore, the improvement of big data acquisition ability and large-scale computational performance has likewise given a strong impetus to the research and development of speech processing technology. Among them, the most important role is still the rise of speech processing methods based on artificial neural networks (Artificial neural networks, ANN), which makes the quality of speech processing further improved by the mode of deep learning. Institutions in China that have conducted speech processing research earlier include the Chinese Academy of Sciences, the University of Science and Technology of China, the National University of Defense Technology, Microsoft Research Asia, IBM China Research Institute, and so on. In recent years, many universities such as Southeast University, Nanjing University of Posts and Telecommunications, South China University of Technology, Soochow University, Harbin Institute of Technology,

Northwestern Polytechnical University, Army Engineering University, and many other companies such as Tencent, KDDI, and Baidu have also begun the research on this technology, and have successively achieved several research results. In 2016, scientists in the field of speech processing from China, Japan, and the United Kingdom organized the VCC2016. VCC2016 was organized by scientists from China, Japan, Britain, and other countries, which provided a data platform and performance scale for speech processing research. 2018 VCC2018 was also held as scheduled, and based on the world's cutting-edge artificial intelligence technology, the speech processing methods were once again pushed forward, and accordingly, the quality of speech processing was once again significantly improved. Enhancement. According to the key breakthroughs in speech recognition technology, Table II summarizes the development of speech processing and recognition technology in different time periods.

TABLE II. SPEECH PROCESSING TECHNOLOGY DEVELOPMENT

Period	Key Technology Breakthroughs	Application Areas	Main features
1950s	Syllable Recognition System	Specific Speech Conversion	Relies on hard coding with limited recognition capabilities
1960s-1970s	Rule-based and statistical modeling	Continuous speech stream processing	Recognition improved but was still limited by noise, etc.
1980s-1990s	Deep learning techniques, neural network models	Processing complex speech streams	Automatic learning capability to handle non-standard pronunciation
Early 21st century to the present	Mobile Internet and smartphone penetration	Multi-language, multi-accent recognition	Dealing with noise, accents, and speed of speech variations, a wide range of applications

B. Synthesis of Main Research Results

1) Development of speech recognition technology: For a long time, the dominant approach for speech recognition has been the Gaussian Mixture Model-Hidden Markov Model (GMM-HMM), which is based on context-dependent generative models of GMM and HMM. Neural networks were also a popular method used for speech recognition, however not as effective as GMM-HMM. Deep learning started to make an impact in the field of speech recognition after a close collaboration between academia and industry in 2010. The collaboration began with a phoneme recognition task, and the results demonstrated the capabilities of the DNN architecture and subsequent convolutional and recurrent network architectures. Their work also demonstrated the importance of moving away from the widely used MFCC features to lower-level raw speech spectrum features. Their collaboration has also yielded good results on large-vocabulary recognition tasks. The main reason for the success of DNN on large-vocabulary speech recognition tasks is that similar to the speech units in GMM-HMM, DNN uses a large-scale output layer structure. The reason for using this structure is that speech researchers expect to take advantage of context-dependent phoneme modeling techniques, which are very effective in GMM-HMM, and that this structure allows for as little change as possible to the highly efficient decoder software architecture developed for GMM-HMM. The work on large-vocabulary speech recognition by the DNN has also demonstrated that, if a large amount of labeled data can be exploited, a pre-training process similar to that for a deep confidence network is not necessary. the training process is unnecessary. Deep learning has been successful in both industry and academia in the field of speech recognition, and this success has been due to three main factors: (1) deep learning significantly reduces the speech recognition error rate compared to the best previous GMM-HMM systems; (2) due to the use of phonemes as the output of DNNs, deploying DNN-based speech recognizers requires only a small portion of the decoder to be changed; and (3) DNNs Powerful modeling capability reduces the complexity of speech recognition systems. After the success of the DNN-HMM system for speech recognition, researchers have proposed many new architectures and nonlinear units for improving speech recognition accuracy. Yu et al. proposed a tensor version of DNN by replacing one or more layers in a traditional DNN by using dual projection layers and tensor layers. The dual projection layer projects each input vector into two nonlinear subspaces. In the tensor layer, the two subspace projections interact with each other and jointly project the next layer of the entire depth architecture. The researchers also propose a method to map the tensor layer to a traditional sigmoid layer, so the tensor layer can be trained in a similar way to the sigmoid layer. The idea of time-domain convolution originated from a time-delay neural network (TDNN), which was used as a shallow network in early speech recognition. Recently when researchers used deep convolutional neural networks for phoneme recognition tasks, they found that weight sharing in the frequency domain is more compared to weight sharing in the time domain [4]. A research report also states that convolutional neural networks help in large-vocabulary continuous speech recognition tasks and that multilayer convolutional neural networks using a large number of convolutional kernels or feature maps give greater performance gains [5]. Sainath et al. explored a large number of variations of deep convolutional neural networks and found that when combined with several new methods, deep convolutional neural networks achieved the best results on several large-vocabulary speech recognition task results [6]. The most notable deep structures for speech recognition tasks are recurrent neural networks and their deep versions [7]. Although RNNs were first successful in phoneme recognition, however, due to the complexity of training RNNs, it was difficult to scale RNNs to larger speech recognition tasks [8]. Since then the learning algorithms for RNNs have been improved and better results have been achieved using RNNs on several tasks, especially using bidirectional LSTM RNNs [9]. In addition to innovations

in deep learning models for speech recognition, a large amount of work has investigated how to develop and implement better nonlinear units. The sigmoid function and the tanh function are the most commonly used nonlinear functions in DNNs, however both have limitations. For example, when the neuron nodes are close to saturation, the error function has a small value of the gradient concerning the parameters, at which point the network is slow to train. To overcome the shortcomings of the sigmoid and tanh functions, Jaitly and Hinton first used ReLU in speech recognition.

Another effective unit for speech recognition is the maxout unit, which is used to construct deep maxout networks. Deep max-out networks generate hidden layer activation values by performing a max operation or max-out operation on a fixed number of weighted inputs. This operation is the same as the maximum pooling operation in convolutional neural networks. These maximum values are the outputs of the previous layer. Thereafter, Zhang et al. generalized the max-out unit into two new types. Both of these types have been shown computationally and experimentally to work better than the ReLU units described in the previous section.

2) Research and development of speech processing technology: The goal of speech processing is to generate speech directly from text as well as additional information. To overcome the shortcomings of statistical parameter-based speech-processing algorithms, researchers have applied deep learning to the field of speech processing. Earlier speech processing techniques were mainly based on rules and rule sets, where computers converted text into speech according to predefined rules. This method requires a lot of human intervention and the results of synthesized speech are not satisfactory. Subsequently, statistical parametric speech processing appeared in the mid-1990s and became the main approach in the field of speech processing [10]. In this approach, the relationship between the text and the corresponding acoustic realizations is represented by a set of stochastic generative acoustic models, which presents three modules that are very important for speech processing: the language model, the acoustic model, and the vocoder. Among them, the task of the language model is to extract the input text into linguistic features utilizing natural language processing techniques, which have the linguistic information needed by the back-end acoustic model. The acoustic model is responsible for converting linguistic features into acoustic features, and then a separate vocoder completes the conversion of acoustic features to the original speech waveform. Hidden Markov models based on decision tree clustering, contextual correlation, and output states satisfying a Gaussian distribution are the most popular generative acoustic models. However, due to the use of shallow HMMs, this acoustic model is not adequately modeled. Some recent studies use deep learning to overcome this deficiency.

Ling et al. used Restricted Boltzmann Machine (RBM) and Deep Confidence Network as generative models to replace the traditional Gaussian model and achieved significant improvement in both subjective and objective evaluation of synthesized speech [11]. Kang et al. used the Deep Confidence Network as a generative model to represent the joint distribution of linguistic features and acoustic features and used the Deep Confidence Network to replace the decision tree and Gaussian model joint distribution of features and used deep confidence networks instead of decision trees and Gaussian models. This approach is similar to using deep confidence networks to generate digital images [12].

With the development of AI deep learning technology, speech processing technology has made a leaping breakthrough, and the iconic technology representative is the Tacotron model proposed by Google in 2017, which is an end-to-end speech processing model based on the self-attention mechanism, where the input side consists of text, which is generated by a text encoder to generate contextual text vectors with robustness, and the decoder side uses an attention-based mechanism based autoregressive decoder at the decoder side, which outputs N frames of Mel Spectrum speech features at a time. The so-called autoregressive decoding means that the N frames output in the first step become inputs in the second step, and this is repeated to generate the complete Mel Spectrogram. The Mel Spectrogram is then passed through the final high-speed convolution module of the Tacotron to generate a linear spectrogram, which is then passed through the Griffin-Lim algorithm to obtain the synthesized speech waveform. Recently, DeepMind announced its latest research breakthrough WaveNet in speech processing [13]. Subsequently, the Tacotron Generation 2 model proposed by Google Inc. in 2018 replaces the high-speed convolution module of the generation algorithm with a 3-layer long and short-term memory module and replaces the vocoder part from the GriffinLim algorithm with the deep-learning WaveNet algorithm, which is noteworthy for its synthesized quality, which is already able to reach the level of falsetto on the subjective evaluation. WaveNet utilizes real human voice clips and corresponding linguistic and phonetic features to train a convolutional neural network to be able to discriminate between linguistic and phonetic audio patterns. When using the WaveNet system, new textual information is input and the WaveNet system regenerates the entire original audio waveform to describe this new textual information.WaveNet is capable of simulating any human voice and generates speech that sounds more natural than the best speech-processing systems available today.WaveNet reduces the discrepancy between simulated-generated speech and the human voice by 50% or more [14].

The Tacotron model can generate high-quality speech processing, however, due to its autoregressive generation structure, the training speed and inference speed are not ideal [15]. Thus, in 2018, Transformer TTS proposed by the University of Electronic Science and Technology of China and Microsoft Research Asia, among others, utilized the self-attention mechanism Transformer to replace the original traditional content-based attention mechanism to accomplish non-autoregressive generation [16]. Subsequently, the FastSpeech1 and Fastspeech2 architectures proposed by Zhejiang University and Microsoft Research Asia in 2019 and 2020, respectively, succeeded in end-to-end non-autoregressive generation, which not only improved the inference speed, but also possessed a duration predictor, a pitch predictor, and an

energy predictor that could accomplish the fine-grained control of the output speech duration, pitch, energy, etc. and at the same time improves the errors of lost words and repeated words that Tacotron2 can make [17].

The VITS model is a 2021 highly expressive speech processing model that combines variational inference, normalized streaming, and adversarial training and is currently comprised of most of the speech processors used on major self-promotion platforms. VITS is the first truly end-to-end speech processing model that does not require an additional vocoder to reconstruct the waveform and directly maps characters or phonemes to waveforms. This synthesis improves the versatility of speech processing by cascading the vocoder and acoustic model of speech processing through hidden variables instead of the spectrum of the previous model [18].

3) Learning paradigms for speech recognition: In speech recognition, model training methods mainly include supervised learning, self-supervised learning, and semi-supervised learning. Traditional supervised learning employs video speech as well as corresponding labeled data and uses the loss function of label prediction for model optimization. Self-supervised learning frameworks, on the other hand, are usually divided into two phases: pre-training and fine-tuning. In the pre-training phase, unlabeled audio and video data are first utilized to train the model based on agent tasks such as mask loss prediction or audio-video cross-modal comparison learning to extract generic deep representations from the audio and video data. Then in the fine-tuning phase, supervised learning and optimization are performed for speech recognition tasks using labeled audio and video data. Semi-supervised learning, on the other hand, utilizes a portion of the labeled data to first train the speech recognition model, followed by pseudo-labeling of unlabeled audio and video speech using the trained model, and subsequently training the model jointly with all audio and video data.

The first is supervised learning. General end-to-end speech recognition directly uses text labels to train the model, and depending on the sequence-to-sequence model framework used, CTC loss, RNNT loss, or cross-entropy loss is used as the objective function for model optimization. To further constrain the model during the model training process, the researchers proposed the use of an auxiliary objective as a regularization term for training. This approach can also be regarded as a kind of multi-task learning. Sterpu et al. proposed the use of lip shape-related articulatory action units (action units) as auxiliary training targets for visual representations based on the AVAlign model to enhance audio-video modal alignment [19]. Ma et al. proposed a method based on audio representations as auxiliary training targets in a lip recognition task [20]. In this, the audio representation target is obtained by extracting the middle layer representation of the encoder using a trained speech recognition model.

The second is self-supervised learning. With the development of the self-supervised learning paradigm in recent years, speech recognition methods based on audio and video self-supervised learning have also received more and more attention. Early audio-video self-supervised learning focuses on the learning of local audio-video features, and most of the methods utilize the natural alignment properties between audio-video modalities to supervise each other's information, representative methods include AVTS, XDC, and local-global audio-video comparative learning methods, etc. LiRA proposes a cross-modal self-supervised learning method based on the pre-trained audio self-supervised learning model PHASE extracts audio representations as targets to train a visual representation extraction module [21]. Audio self-supervised learning methods have developed rapidly in the recent past, and researchers have been inspired by them and extended them to audio-video self-supervised learning, which includes AV-HuBERT based on the HuBERT model extended to audio-video speech, and AV-data2vec based on the data2vec model extended to audio-video speech. Based on the model of AV-HuBERT, Zhu et al. further introduced additional text modalities to realize the joint learning of audio, video, and text modal representations [22]. u-HuBERT further introduced additional unimodal speech in AV-HuBERT and could theoretically introduce more modalities of data [23]. The AV2vec model and the RAVEn models, on the other hand, use a teacher model based on an exponential sliding average approach to multimodal self-distillation learning for student models in training [24]. Audio-video self-supervised speech recognition has achieved better results in lip recognition as well as speech recognition tasks, and many current studies have used pre-trained self-supervised models to initialize the parameters of speech recognition models [25].

Then comes semi-supervised learning. Semi-supervised speech recognition methods are mostly inspired by semi-supervised learning methods for audio speech recognition. The Auto-AVSR model generates pseudo-labels for unlabeled speech recognition datasets using pre-trained speech recognition models and then trains the speech recognition models using a combination of labeled and pseudo-labeled data [26]. The AV-CPL method, on the other hand, proposes a semi-supervised based on continuous pseudo-label update learning method, which continuously employs the updated model for continuous optimization of pseudo-labels on unlabeled audio and video [27]. Self-supervised learning methods are limited by less labeled data, so the results are generally worse than semi-supervised learning. However, the self-supervised learning method can be combined with semi-supervised learning, i.e., using self-supervised learning at the beginning to obtain an initial speech recognition model, then using the initial model to generate text pseudo-labels for unlabeled data, and then jointly optimizing the recognition model with labeled data, which generally achieves better speech recognition results. The summary of speech recognition learning paradigms is shown in Table III.

TABLE III. SUMMARY OF SPEECH LEARNING PARADIGMS

Learning paradigm	Descriptive	Typical models or methods	Vintage	Drawbacks
Supervised learning	Using labeled speech data to train the model, the model learns the mapping relationship from speech signal to text	LSTM+CTC, DNN-HMM, RNN-HMM, etc.	1. can directly optimize the speech recognition performance 2. is effective in the case of sufficient annotation data	1. Labelling data is costly and time-consuming 2. Difficulty in covering all possible speech variations and noise environments
Self-supervised learning	Using unlabeled speech data to learn useful speech representations by designing assistive tasks	Siamese Networks, Generative Adversarial Networks (GANs), Self-Encoders	1. does not require large amounts of labeled data 2. can learn more generalized speech features 3. helps to improve the generalization ability of the model	1. complex model structures and training strategies may be required 2. The design of ancillary tasks has a large impact on performance
Semi-supervised learning	Combining labeled and unlabeled speech data for learning to improve the accuracy and generalization of models	Clustering-based methods, bias correction-based methods, etc.	1. exploits the richness of unlabeled data 2. enables improved performance with limited labeled data	1. need to balance the use of labeled and unlabeled data 2. may require complex model fusion strategies

III. KEY ISSUES IN INTELLIGENT SPEECH PROCESSING BASED ON AI TECHNOLOGY

A. Main Tasks of Speech Processing Technology

1) Speech matching: Speech matching refers to the automatic retrieval of all speech segments that have the same content as the query speech segment from a given speech database, so the extracted speech features are crucial for the speech-matching task. Since speech matching automatically retrieves all speech segments with the same content as the query speech segment from a given speech database, it can be regarded as a class of content-based speech retrieval applications, which have been widely used in music retrieval, song recommendation, and speech intelligence analysis. At the same time, speech matching is a class of unsupervised learning tasks, and the techniques applicable to speech matching can be applied to other unsupervised learning tasks in the field of machine learning, thus the study of speech matching algorithms has important academic value. According to the above scenario, it can be seen that the speech matching task is a typical class of speech processing tasks because the key to speech matching is the extraction of speech features.

2) Multimodal speech recognition: Human-computer interaction interfaces for intelligent machines, such as smartphones, home robots, and self-driving cars, are becoming more and more common in daily life. Speech recognition that is robust to noise is the key to achieving effective human-machine interaction [28]. Multimodal speech recognition is considered one of the effective solutions for robust speech recognition. In human-computer interaction systems, the machine can not only receive the operator's speech signal, but also observe the operator's behavioral information, such as body movement and changes in mouth shape, and this behavioral information can help the machine recognize the operator's speech signal, and the study of multimodal speech recognition has an important application value in human-computer interaction systems [29]. In addition, multimodal speech recognition is a supervised learning task that involves the fusion of information from multiple sources, which has important academic research value. Multimodal speech recognition is also a typical speech-processing task [30].

B. Technical Issues in Intelligent Speech Recognition and Processing

In recent years, with the continuous increase of current training data, deep learning models, and learning paradigms, the effect of speech recognition has been greatly improved. However, there are still many challenges in speech recognition, and future research needs to pay more attention to the following points. First, current speech recognition still performs poorly in more open and complex environments, such as home scenarios and multiple people talking freely [31]. Currently, most studies are still overly focused on the LRS2 and LRS3 datasets, but lip recognition as well as speech recognition on these two datasets are already saturated with word error rates, e.g., Auto-AVSR has a speech recognition word error rate of 0.9% on LRS3 [32]. Under other more challenging scenario test sets, such as MISP2021 and Google's YTDEV, speech recognition remains poor. In the MISP2021AVSR competition, the word error rate of the winning team was still as high as 24.6% [33]. Therefore, future research in speech recognition needs to focus on the performance of datasets with more complex conditions and avoid overfitting to single or multiple similar scene datasets [34]. Second, the performance of speech recognition in small languages with sparse data still needs to be improved. Similar to speech recognition, there are still many challenges in small-language speech recognition [35]. Compared to mono-speech, audio and video data in small languages are more difficult to capture, and many small languages even face the challenge of not being able to capture video data [36]. How to deal with this situation of missing visual data and very low resources to improve speech recognition in small languages is still a problem that needs to be further researched and solved. Third, speech recognition is relatively computationally intensive and incurs high inference computation costs in practical application deployments. Under many high signal-to-noise ratio conditions, the improvement of speech recognition relative to audio speech recognition is not significant and brings little benefit [37]. Therefore, how to further reduce the computational burden of speech recognition, achieve lightweight computation of the vision module, and at the same time realize effective dynamic computation for environmental noise conditions is one of the important issues to be solved for the practicalization of speech recognition [38]. Fourth, the robustness of speech recognition in the case of missing video modalities and faces being occluded needs to be further improved [39]. Recently, researchers have

begun to pay more and more attention to such problems of speech recognition in practical applications, and the proposed methods have improved the robustness of speech recognition to some extent [40]. However, these challenges have not been fully addressed, especially when there are cases of video damage, occlusion, etc. outside the training set distribution, speech recognition may still be worse than unimodal speech recognition. Ideal speech recognition maximizes the use of incremental information in the video while avoiding the effects of distracting information that appears in the video. How best to achieve this goal remains the open question.

IV. Development Issues in the Field of Intelligent Speech Processing based on AI Technology

AI technology in the development process is accompanied by lingering controversies, such as ChatGPT since its inception has been full of controversial copyright issues, as well as the existence of a large number of the use of AI voice cloning technology on the network to clone the voices of some of the singers and secondary creation of content, which belong to the "gray area" [41 These are all "gray areas" [41]. These contents not only involve copyright issues but also bring some difficult ethical problems. The current rapid development of AI technology has led to three major problems as summarized in Table IV.

TABLE IV. AI-Based Intelligent Speech Processing Problem

Type of problem	Explicit description	Frequency/number of reports
Data Privacy and Security	Voice data is being collected on a large scale, increasing the risk of privacy breaches	In recent years, an average of dozens of related privacy breaches have occurred each year, affecting millions of users.
Algorithmic bias and discrimination	Bias in training data leads to unfair recognition results	Several studies have reported that at least half of intelligent speech-processing systems suffer from varying degrees of algorithmic bias.
Responsibility and accountability	Difficulty in defining the responsible party when there is a problem with the speech recognition system	Every year there are dozens of legal disputes caused by errors in voice recognition systems, most of which involve unclear delineation of responsibility.

1) Data privacy and security issues: Generative AI models require a large amount of data for training at the initial stage, and data privacy issues are involved in whether the source of the data is traceable, authentic, and licensed by the information source. In the process of training generative AI models, the protection of user data is very important and must be done when conducting experiments. Several protection measures can be taken to address this issue. First, the data can be anonymized, the user data can be desensitized, and sensitive information, such as name, address, phone number, etc., can be deleted or replaced while ensuring accuracy, and encryption technology can be used to encrypt the user data to ensure that only authorized personnel can access it [42]. Second, user data can be stored in an isolated environment, data access rights can be set, and secure communication protocols can be used for transmission. Based on this, training data selection is performed by choosing appropriate training data in the dataset and avoiding the use of training data containing sensitive information, such as publicly available datasets or using processed data [43]. Again, the security of the model should be ensured, and after the model training is completed, the model needs to be evaluated for security to ensure that the model will not leak user data or have adverse effects. Finally, relevant laws and regulations, such as the Data Security Law of the People's Republic of China and the Personal Information Protection Law of the People's Republic of China, need to be strictly observed to ensure that user data are legally used and protected. In conclusion, a series of measures are needed to protect the security and privacy of user data when training and using generative AI.

2) Algorithmic bias and discrimination issues: If there is bias in the training data, then the generative AI may replicate that bias, leading to discriminatory decisions. American news commentator and social critic Lippmann famously proposed the "agenda-setting theory", which argues that "the news media influences 'the images we have in our minds about the world'", and this is also the case in generative AI [44]. For example, in 2016 Microsoft launched Tay, a conversational bot that was "portrayed" as racist only 16 hours after it went live and engaged in a conversation with a user, after which Microsoft urgently took the account offline [45]. The fact that it took only 16 hours from posting to taking the account offline suggests that a series of measures need to be taken to deal with this kind of problem. The first is the cleaning and filtering of data containing bias and discrimination during the data collection and processing phase [46]. This step involves removing or correcting inaccurate, outdated, or discriminatory data and, during the data collection process, ensuring that the data is diverse and inclusive. This means collecting data from different backgrounds, genders, ethnicities, and other characteristics to reduce the possibility of algorithmic bias and discrimination [47]. During algorithm design and training, ensure that the model is fair and unbiased. This includes the use of fair evaluation criteria, unbiased training data and algorithms, and a transparent decision-making process [48]. Algorithms also need to be audited periodically after the model has begun to perform computations to ensure that they meet the requirements of impartiality and fairness [49]. This includes examining the design of the algorithm, the training data, and actual runs, as well as assessing its impact on different populations. Bias detection and correction techniques can also be used to identify and correct bias in algorithms, where available [50]. This includes the use of techniques such as unsupervised learning, semi-supervised learning, or reinforcement learning to improve

the performance of the model and reduce bias. In terms of personnel requirements, there is a need to educate and train algorithm designers and users to increase their awareness of bias and discrimination issues and equip them with the skills to recognize and address these issues.

3) Responsibility and accountability issues: When AI systems lead to undesirable consequences, how to pursue responsibility is a complex issue. Similar to how to select the author of the textual content produced by generative AI, there are many controversies around this issue, some scholars believe that the responsibility should be borne by the AI system, and some scholars believe that it should be borne by the program developer, after all, the entire algorithmic mechanism is from the developer's hand [51]. It can be seen that pursuing responsibility is a complex issue that needs to be considered from several aspects. First, it is necessary to determine who the owners and users of the AI system are, i.e., the responsible subjects. This involves multiple parties such as hardware and software vendors, data providers, application developers, and users. Second, it is necessary to understand information about the operation mechanism, algorithm design, training data, and actual application scenarios of the AI system. This helps to analyze the causes of adverse consequences and attribute responsibility. Again, it is important to assess the extent of the impact of the adverse consequences of the AI system on individuals, organizations, and society for a specific time, including economic losses, privacy leakage, security issues, etc. Finally, it is important to determine the type of responsibility, including technical, legal, ethical, and other responsibilities, based on the adverse consequences and impacts of AI systems [52]. The discussion of this issue, should not stop at determining the responsible person, but should also look for solutions, based on the type of responsibility and the actual situation, including compensation for losses, improvement of system design, correction of algorithms, and strengthening of supervision. At the same time, an accountability mechanism should be established to hold AI systems accountable for adverse consequences. There is a need to establish internal accountability systems, implement transparent and traceable management programs, and strengthen regulation and legal sanctions. In addition, there is a need to strengthen cooperation and communication, mainly among government departments, enterprises, social organizations, and other institutions, to jointly solve problems that are likely to arise from AI systems.

V. Future Outlook for Intelligent Speech Processing

A. Forecast of Technology Development Trends

The future development of speech-processing technology will mainly rely on the continuous development of deep learning and neural network technology. With the continuous upgrading of hardware equipment and the continuous optimization of algorithms, the quality and naturalness of speech processing technology will continue to improve. The current speech processing technology can already realize realistic speech processing, but there are still some shortcomings, such as the rhythm and rhyme of the voice is not natural. Future speech processing technology will pay more attention to the improvement of these aspects to realize more realistic speech processing. Future speech-processing technology will pay more attention to personalized services and experiences. With the continuous development of artificial intelligence technology, future speech processing technology will be able to personalize speech processing according to the user's needs and preferences, and can synthesize any person's voice based on small or zero samples, providing speech processing services that are closer to the user's needs. Future speech processing technology will pay more attention to real-time speech processing. Real-time speech processing can provide users with a more natural and smooth voice interaction experience, providing a broader application space for the development of voice interaction technology.

B. Forecast of Application Development

With the development of artificial intelligence technology and other related technologies, the field of intelligent speech processing will increasingly focus on the personalized needs and experiences of users. Future speech processing technologies will be able to better provide personalized services and experiences, such as personalized speech processing based on user needs and interests, thereby increasing user satisfaction and loyalty.

In addition, multilingual support and cross-cultural communication will be facilitated by deep learning from a large amount of text and data. Speech-processing technology will be expected to support more languages and dialects so that it can better meet the needs of users from different countries and regions and realize cross-cultural communication [53]. Further, speech processing technology will also realize automatic translation and conversion between multiple languages, providing users with more convenient and diversified services. At the same time, the future speech processing technology will be able to better combine with other integrated media technologies, such as images, videos, text, etc., to realize richer and more vivid forms of media expression [54]. For example, in TV news, speech processing technology can be combined with video and text to realize more vivid and intuitive news broadcasting. Based on the above applications, it is foreseeable that future speech processing technology is expected to be combined with augmented reality technology to realize a more intelligent and convenient user experience. Based on the trend of deep integration of speech processing technology and artificial intelligence technology, the key applications of technology in the future may be concentrated in the following areas, as shown in Table V:

TABLE V. KEY APPLICATION AREAS FOR INTELLIGENT SPEECH PROCESSING

Application Areas	Specific application	Key technologies
smart home	Smart speaker, smart appliance control	Speech Recognition, Speech Synthesis, Internet of Things
smart city	Intelligent Transportation, Intelligent Security	Deep learning, computer vision, big data analytics
health care	Disease diagnosis, drug development	Natural language processing, machine learning, big data analytics
financial	Intelligent Risk Control, Intelligent Investment	Machine learning, natural language processing, big data analytics
teach	Intelligent Tutoring, Language Learning	Speech Recognition, Natural Language Processing, Intelligent Recommender Systems

VI. CONCLUSION

Intelligent speech processing technology has made significant progress under the impetus of AI technology, which not only realizes a qualitative leap in speech recognition and speech processing accuracy but also promotes the innovation of human-computer interaction. Through advanced technologies such as deep learning, intelligent speech processing systems can more accurately understand the emotional and semantic information in human speech and generate natural and smooth speech. With the continuous maturation of technology and the expansion of application scenarios, intelligent speech processing technology will play an important role in more fields, and at the same time, its wide application has also brought about profound impacts on social structure, occupational structure, and privacy laws. Looking ahead, intelligent speech-processing technology will continue to innovate and develop, bringing more convenience and possibilities to human society.

REFERENCES

[1] Bostrom, N. (1998). How long before superintelligence?. *International Journal of Futures Studies*.

[2] Hinton, G., & Shallice, T. (1991). Lesioning an attractor network: investigations of acquired dyslexia. *Psychological Review*, 98(1), 74-95.

[3] Yu, D., Deng, L., & Seide, F. (2013). The deep tensor neural network with applications to large vocabulary speech recognition. *IEEE Transactions on Audio Speech and Language Processing*, 21(2), 388-396.

[4] Abdel-Hamid, O., Mohamed, A. R., Jiang, H., & Penn, G. (2012). Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition. *IEEE International Conference on Acoustics*. IEEE.

[5] Sainath, T. N., Kingsbury, B., Mohamed, A. R., Dahl, G. E., Saon, G., & Soltau, H. (2013). Improvements to deep convolutional neural networks for LVCSR. *Automatic Speech Recognition & Understanding*. IEEE.

[6] Pardede, H. F., Adhi, P., & Zilvan, V. R. A. K. D. (2023). Deep convolutional neural networks-based features for indonesian large vocabulary speech recognition. *IAES International Journal of Artificial Intelligence*, 12(2), 610-617.

[7] Graves, A., Mohamed, A. R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. Acoustics, Speech, and Signal Processing, 1988. ICASSP-88. *1988 International Conference on* (Vol.38). IEEE.

[8] Wei, A., Zhang, H., & Zhao, E. (2025). DereflectFormer: Vision Transformers for Single Image Reflection Removal. *International Conference on Pattern Recognition*. Springer, Cham.

[9] Graves, A., Jaitly, N., & Mohamed, A. R. (2013). Hybrid speech recognition with deep bidirectional lstm. *IEEE*.

[10] Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., & Oura, K.(2013). Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, 101(5), 1234-1252.

[11] Ling, Z. H., Deng, L., Yu, D. (2013). Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10): 2129-2139.

[12] Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., & Kavukcuoglu, K. (2016). Wavenet: a generative model for raw audio. DOI:10.48550/arXiv.1609.03499.

[13] Miao, Y., Metze, F., Rawat, S. (2013). Deep max out networks for low-resource speech recognition. *IEEE Workshop on Automatic Speech Recognition and Understanding*: 398-403.

[14] Zhang, X., Trmal, J., Povey, D., & Khudanpur, S. (2014). Improving deep neural network acoustic models using generalized maxout networks. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 215–219. doi.org/10.1109/ICASSP.2014.6853589

[15] Humphrey, E. J., & Bello, J. P. (2012). Rethinking automatic chord recognition with convolutional neural networks. *2012 11th International Conference on Machine Learning and Applications* (ICMLA), 2, 357–362. doi.org/10.1109/ICMLA.2012.232

[16] Humphrey, E. J., Bello, J. P., & LeCun, Y. (2012). Moving beyond feature design: Deep architectures and automatic feature learning in music informatics. Proceedings of the 13th International Society for Music Information Retrieval Conference, 403–408.

[17] Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2011). Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Communications of the ACM*, 54(10), 95–103. doi.org/10.1145/2001269.2001295

[18] Sterpu, G., Saam, C., & Harte, N. (2020). How to teach DNNs to pay attention to the visual modality in speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, 1052–1064. doi.org/10.1109/TASLP.2020.2983042

[19] Ma, P. C., Petridis, S., & Pantic, M. (2022). Visual speech recognition for multiple languages in the wild. *Nature Machine Intelligence*, 4, 930–939. doi.org/10.1038/s42256-022-00523-1

[20] Ma, P. C., Wang, Y. J., Petridis, S., & Pantic, M. (2022). Training strategies for improved lip-reading. In 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 8472–8476). *IEEE*. doi.org/10.1109/ICASSP43922.2022.9747620

[21] Korbar, B., Tran, D., & Torresani, L. (2018). Cooperative learning of audio and video models from self-supervised synchronization. arXiv preprint arXiv:1807.00230.

[22] Alwassel, H., Mahajan, D., Torresani, L., & Ghanem, B. (2019). Self-supervised learning by cross-modal audio-video clustering. arXiv preprint arXiv:1911.12667.

[23] Ma, S., Zeng, Z. Y., McDuff, D. J., & Song, Y. (2021). Contrastive learning of global and local video representations. *In Advances in Neural Information Processing Systems* (Vol. 34, pp. 7025–7040).

[24] Ma, P. C., Mira, R., Petridis, S., & Pantic, M. (2021). LiRA: Learning visual speech representations from audio through self-supervision. *In Interspeech 2021* (pp. 3011–3015).

[25] Hsu, W. N., Bolte, B., Tsai, Y. H. H. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 3451–3460. doi.org/10.1109/TASLP.2021.3122291

[26] Petridis, S., & Pantic, M. (2016). Deep complementary bottleneck features for visual speech recognition. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2304–2308). *IEEE*. doi.org/10.1109/ICASSP.2016.7472081

[27] Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017). Lip reading sentences in the wild. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3444–3453). *IEEE*. doi.org/10.1109/CVPR.2017.367

[28] Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2022). Deep audiovisual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12), 8717–8727. doi.org/10.1109/TPAMI.2021.3058582

[29] Serdyuk, D., Braga, O., & Siohan, O. (2022). Transformer-based video front-ends for audio-visual speech recognition for single and multi-person video. *In Interspeech* 2022 (pp. 2833–2837).

[30] Makino, T., Liao, H., Assael, Y., et al. (2019). Recurrent neural network transducer for audio-visual speech recognition. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 905–912). *IEEE*. doi.org/10.1109/ASRU46091.2019.9003751

[31] Ma, P. C., Petridis, S., & Pantic, M. (2021). End-to-end audiovisual speech recognition with conformers. In 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7613–7617). *IEEE*. doi.org/10.1109/ICASSP39728.2021.9413580

[32] Petridis, S., Stafylakis, T., Ma, P. C., & Pantic, M. (2018). Audio-visual speech recognition with a hybrid CTC/attention architecture. In 2018 IEEE Spoken Language Technology Workshop (SLT) (pp. 513–520). *IEEE*. doi.org/10.1109/SLT.2018.8639589

[33] Assael, Y. M., Shillingford, B., Whiteson, S., & de Freitas, N. (2016). LipNet: Sentence-level lipreading. arXiv preprint arXiv:1611.01599.

[34] Wand, M., Koutník, J., & Schmidhuber, J. (2016). Lipreading with long short-term memory. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6115–6119. https://doi.org/10.1109/ICASSP.2016.7472852

[35] Stafylakis, T., & Tzimiropoulos, G. (2017). Combining residual networks with LSTMs for lipreading. *Interspeech 2017*, 3652–3656. doi.org/10.21437/Interspeech.2017-85

[36] Gao, R. H., & Grauman, K. (2021). VisualVoice: Audio-visual speech separation with cross-modal consistency. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 15490–15500. doi.org/10.1109/CVPR46437.2021.01525

[37] Ma, P. C., Martinez, B., Petridis, S., & Pantic, M. (2021). Towards practical lipreading with distilled and efficient models. 2021 *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 7608–7612. doi.org/10.1109/ICASSP39728.2021.9413581

[38] Prajwal, K. R., Afouras, T., & Zisserman, A. (2022). Sub-word level lip reading with visual attention. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 5152–5162. doi.org/10.1109/CVPR52688.2022.00508

[39] Boulanger-Lewandowski, N., Bengio, Y., & Vincent, P. (2013). Audio chord recognition with recurrent neural networks. *Proceedings of the 14th International Society for Music Information Retrieval Conference* (ISMIR 2013), 335–340.

[40] Van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep content-based music recommendation. *Advances in Neural Information Processing Systems* (NIPS 2013), 2643–2651.

[41] Lin, M., Chen, Q., & Yan, S. (2013). Network in network. arXiv preprint arXiv:1312.4400.

[42] Zeiler, M. D., & Fergus, R. (2013). Stochastic pooling for regularization of deep convolutional neural networks. arXiv preprint arXiv:1301.3557.

[43] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9), 1904–1916. doi.org/10.1109/TPAMI.2015.2389824

[44] Ouyang, W., Luo, P., Zeng, X., Qiu, S., & Wang, X. (2014). DeepID-Net: Multi-stage and deformable deep convolutional neural networks for object detection. arXiv preprint arXiv:1409.3505.

[45] Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *In European Conference on Computer Vision* (ECCV 2014) (pp. 818–833). Springer. doi.org/10.1007/978-3-319-10590-1_53

[46] Chen, H., Du, J., Hu, Y., & Dai, L. (2021). Correlating subword articulation with lip shapes for embedding aware audiovisual speech enhancement. *Neural Networks*, 143, 171–182. doi.org/10.1016/j.neunet.2021.07.021

[47] Hu, D., Li, X. L., & Lu, X. Q. (2016). Temporal multimodal learning in audiovisual speech recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3574–3582). Las Vegas, NV, USA: IEEE. doi.org/10.1109/CVPR.2016.389

[48] Ninomiya, H., Kitaoka, N., Tamura, S., Iribe, Y., & Takeda, K. (2015). Integration of deep bottleneck features for audio-visual speech recognition. *In Proceedings of Interspeech 2015* (pp. 563–567). Dresden, Germany: ISCA.

[49] Mroueh, Y., Marcheret, E., & Goel, V. (2015). Deep multimodal learning for audio-visual speech recognition. *In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 2130–2134). South Brisbane, Queensland, Australia: IEEE. doi.org/10.1109/ICASSP.2015.7178370

[50] Takashima, Y., Aihara, R., Takiguchi, T., & Ariki, Y. (2016). Audio-visual speech recognition using bimodal-trained bottleneck features for a person with severe hearing loss. *In Proceedings of Interspeech 2016* (pp. 277–281). San Francisco, CA, USA: ISCA.

[51] Yu, J. W., Zhang, S. X., Wu, J., & Xie, L. (2020). Audio-visual recognition of overlapped speech for the LRS2 dataset. *In ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6984–6988). Barcelona, Spain: IEEE. doi.org/10.1109/ICASSP40776.2020.9054551

[52] Li, J. H., Li, C. D., Wu, Y. F., & Xie, L. (2023). Robust audio-visual ASR with unified cross-modal attention. *In ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 1–5). Rhodes Island, Greece: IEEE.

[53] Wang, J. D., Qian, X. Y., & Li, H. Z. (2022). Predict-and-update network: Audio-visual speech recognition inspired by human speech perception. arXiv preprint arXiv:2209.01768.

An Enhanced Whale Optimization Algorithm Based on Fibonacci Search Principle for Service Composition in the Internet of Things

Yun CUI
Hebei Chemical & Pharmaceutical College, Shi Jiazhuang 050026, China

***Abstract*—Service composition in the Internet of Things (IoT) poses significant challenges owing to the dynamics in IoT ecosystems and the exponential increase in service candidates. This paper proposes an Enhanced Whale Optimization Algorithm (EWOA) by introducing the Fibonacci search principle for service composition optimization to overcome certain shortcomings of conventional approaches, including slow convergence and being stuck in local optima, in addition to imbalanced exploration-exploitation trade-offs. The proposed EWOA combines the application of nonlinear crossover weights with a Fibonacci search to optimize the global exploration and local exploitation searches of the basic version, thereby producing a better solution. Several simulations were performed for IoT functions. Among the experiments involving different QOS-based service compositions, the results show that the EWOA achieves superior and faster convergence capability with enhanced convergence compared to recent methods.**

***Keywords*—Service composition; Internet of Things; quality of service; whale optimization; Fibonacci search**

I. INTRODUCTION

The Internet of Things (IoT) is driving a revolution in modern technologies, enabling the world to grow further connected, and billions of devices can communicate and share data effectively [1]. This rapid growth of services has led to a vast repository of functionalities that address different application areas [2]. However, this immense size and dynamism introduce a severe challenge: efficiently composing multiple IoT services to satisfy user demands [3]. Efficient service composition guarantees the realization of desired functionalities and optimizes the quality attributes of execution time, reliability, and cost in dynamic IoT ecosystems [4].

Traditional optimization methods and swarm-intelligence-based algorithms offer great promise for solving service composition problems [5]. However, these methods typically incur severe limitations due to slow convergence rates, premature stagnation at local optima, or failure to balance exploration and exploitation efficiently [6]. These challenges impede efficient solutions, particularly for large-scale and complex IoT service composition problems; hence, advanced approaches are required to ensure their robustness and scalability [7].

This paper presents an improved QoS-based service composition using a new metaheuristic approach. The contribution of this work is to improve the solution accuracy, accelerate convergence, and overcome local optima entrapment by addressing the limitations of existing algorithms. Thus, this work aims to find a globally optimal balance between local exploitation and global exploration and ensure a more effective and efficient service composition in dynamic IoT environments.

These include the development of the Fibonacci search principle to improve the performance of the WOA on global optimization problems and applying nonlinear weights in the WOA to maintain balance within its processes. In general, this contributes to reaching the optimality of a solution by increasing exploration during the initialization stage, thus creating faster convergence, whereas adopting this modified WOA to tackle the optimal solution in IoT service composition. Intensive validation showed the excellent performance of this algorithm owing to better convergence speed and stability in results, hence providing a concrete base for more challenging tasks within IoT service optimization processes.

The remainder of this paper is organized as follows: Section II presents a comprehensive literature review, discussing previous studies and existing optimization techniques for IoT service composition. Section III details the proposed algorithm. Section IV describes the experimental setup and presents the simulation results. Section V provides a critical discussion and comparative analysis of the findings. Finally, Section VI concludes the study, summarizing key contributions and outlining potential future research directions.

II. LITERATURE REVIEW

The authors in study [8] propose an Artificial Neural Network-based Particle Swarm Optimization (ANN-PSO) hybrid algorithm for cloud-edge computing to improve QoS factors. They utilized a formal verification process through functional transitions and constraint logic to ensure correctness regarding functional and non-functional aspects. Their approach shows enhanced memory, response time, availability, and price, yielding higher fitness values than others.

In study [9], a Hidden Markov Model (HMM) integrated with Ant Colony Optimization (ACO) was suggested for IoT service composition. HMM was trained for QoS prediction, and the Viterbi approach enhanced the emissions and transitions. The ACO algorithm identified optimal service paths, achieving better response time, reliability, energy consumption, and cost than prior approaches.

The authors of study [10] developed a semantic middleware to address IoT service composition challenges, incorporating contextual service search and semantic analysis. Automated, scalable service composition enhanced scalability, validated on innovative city scenarios, with improved service discovery, selection, and composition metrics versus existing methods.

In study [11], a fuzzy-driven hybrid algorithm combining ACO and Artificial Bee Colony (ABC) methods was proposed for cloud-fog IoT service composition. The approach optimized QoS metrics, including energy consumption, availability, reliability, and cost, achieving significant performance improvements over contemporary techniques.

Moreover, in study [12], a service composition technique based on Grey Wolf Optimization (GWO) within the MapReduce methodology was presented for QoS-aware IoT applications. The model achieved energy savings, reduced response times, and enhanced availability and cost metrics, with average performance gains over baseline algorithms.

The authors in study [13] proposed an enhanced ABC with a dynamic dimensionality reduction-inspired mechanism for IoT Service Composition. Further, the dimensions of disparity adjustment among solutions enable a method that has improved convergence rates and a more balanced exploration of solution exploitation. This significantly facilitates energy consumption to enhance availability and reliability with cost metrics.

In study [14], the Discrete Adjustable Lion Optimization Algorithm (DALOA) was proposed for composing IoT services, employing sub-populations and operators like roaming, mating, and migration. The approach provided a strong balance between exploitation and exploration, achieving near-optimal QoS-aware compositions in reduced execution time. A QoS-aware service discovery, developed using WOA and GA, is proposed in study [15]. This bioinspired technique has proven efficient in selecting a way of optimally utilizing energy, data access time, and cost-effectiveness in IoT service discovery.

While some existing approaches, as summarized in Table I, represent significant advances in IoT service composition, several critical gaps remain. Most of these approaches, such as HMM-based or semantic middleware-based, are not sufficiently adaptive for wider-scale IoT environments because of computational overhead or energy inefficiency in IoT devices. The majority of them experience performance bounds in highly dynamic IoT scenarios.

TABLE I. RECENT LITERATURE IN IOT SERVICE COMPOSITION

Algorithm	Key features	Performance gains	Shortcomings
ANN-PSO [8]	Hybrid approach combining ANNs for QoS enhancement and PSO for candidate service selection. Formal verification using labeled transition systems ensures correctness.	Achieved better response time, availability, and cost efficiency. Demonstrated improved fitness function values compared to other algorithms.	High computational complexity due to the hybrid approach and formal verification methods.
HMM + ACO [9]	The hidden Markov Model predicts QoS metrics based on emissions and transitions optimized by the Viterbi algorithm. ACO is used for service path identification.	Enhanced QoS regarding energy usage, cost, reliability, and response time. Outperformed prior techniques in availability and efficiency.	Limited scalability for real-time large-scale IoT environments due to the computational overhead of HMM.
Semantic middleware [10]	Modular and context-aware semantic abstraction for discovering IoT services, semantic filtering, and lightweight automatic service composition. Validated in smart city scenarios.	Improved scalability of service discovery by 15%, selection by 20%, and composition by 40% compared to state-of-the-art methods.	It focuses primarily on scalability but lacks energy efficiency and response time optimizations.
ACO + ABC (Fuzzy-based hybrid) [11]	A hybrid algorithm combining ACO and ABC algorithms with a fuzzy logic system. Focuses on energy-aware and QoS-based service selection in cloud-fog architectures.	Reduced energy utilization by 17%, improved availability by 8%, enhanced reliability by 4%, and lowered cost by 21% on average.	Increased complexity due to hybridization and dependency on parameter tuning for optimal performance.
GWO + MapReduce [12]	Combines GWO with the MapReduce framework to enable large-scale IoT service composition optimization. Targets QoS metrics like response time, cost, and energy savings.	Achieved a 24% reduction in cost, an 11% gain in availability, a 14% drop in response time, and a 40% energy savings.	MapReduce overhead may impact performance in highly dynamic IoT scenarios.
ABC with dynamic reduction [13]	The enhanced ABC algorithm introduces a dynamic reduction mechanism. Adjusts dimension disparities among solutions dynamically for better exploration-exploitation balance.	Decreased energy consumption by 17%, increased availability by 10%, improved reliability by 8%, and lowered cost by 23% compared to alternatives.	Potential convergence issues if initial dimension disparities are not set optimally.
DALOA [14]	Discrete adaptive lion optimization algorithm with unique operators (roaming, mating, migration). Balances strong global exploration through nomad roaming and efficient local exploitation via pride searching.	Achieved the best trade-off between exploration and exploitation. Reduced execution time and provided near-optimal IoT service compositions.	Increased complexity due to multiple operators and higher execution time for larger populations.
WOA + GA [15]	Integrates WOA with genetic algorithm for efficient IoT service discovery and selection. Bio-inspired optimization enhances QoS awareness in dynamic environments.	Optimized energy utilization, reduced data access time, and improved cost-effectiveness compared to traditional methods.	Lacks adaptability to large-scale IoT environments due to limited scalability and high computational cost.

In addition, the optimal balance between exploitation and exploration limits the possibility of obtaining a globally optimal solution with high efficiency. This study attempts to fill these gaps by incorporating the Fibonacci search principle into the WOA to leverage its global optimization strengths to enhance the convergence rates, stability, and QoS outcomes in IoT service composition.

III. PROPOSED METHODOLOGY

A. Problem Formulation and Statement

An IoT service refers to functional components within the IoT environment that facilitate the interaction and exchange of information between devices [16]. These services can be defined as a triple (TDP, FDP, $QoSDP$) where TDP stands for the text description of services, providing a semantic explanation of its functionality; FDP represents the functional description of services, detailing its operations and capabilities; and QoSDP refers to the Quality of Service (QoS) characteristics associated with IoT services, describing non-functional properties like execution time, cost, reliability, and trust. The QoS attributes provide measurable criteria for assessing service performance and ensuring non-functional requirements are met. These attributes are particularly crucial when selecting services for specific tasks in IoT applications.

Abstract IoT services represent a set of service instances that perform similar or identical functions. These services are abstracted into individual tasks within a requirements workflow and ensure functional compatibility but differ in their QoS values, making them key candidates in the service composition process [17].

As shown in Fig. 1, the composition of IoT services can follow various control logic structures based on user requirements. In the loop, certain tasks are repeated iteratively for a specified number of iterations (Fig. 1(a)). In the selection fashion, tasks are chosen based on specific conditions or decision points (Fig. 1(b)). In the parallel way, multiple tasks are executed simultaneously to improve performance (Fig. 1 (c)). Lastly, in the sequential approach, tasks are executed one after the other in a predefined order (Fig. 1 (d)). Given the focus on simplicity and efficiency, this paper exclusively considers sequential structures for IoT service composition.

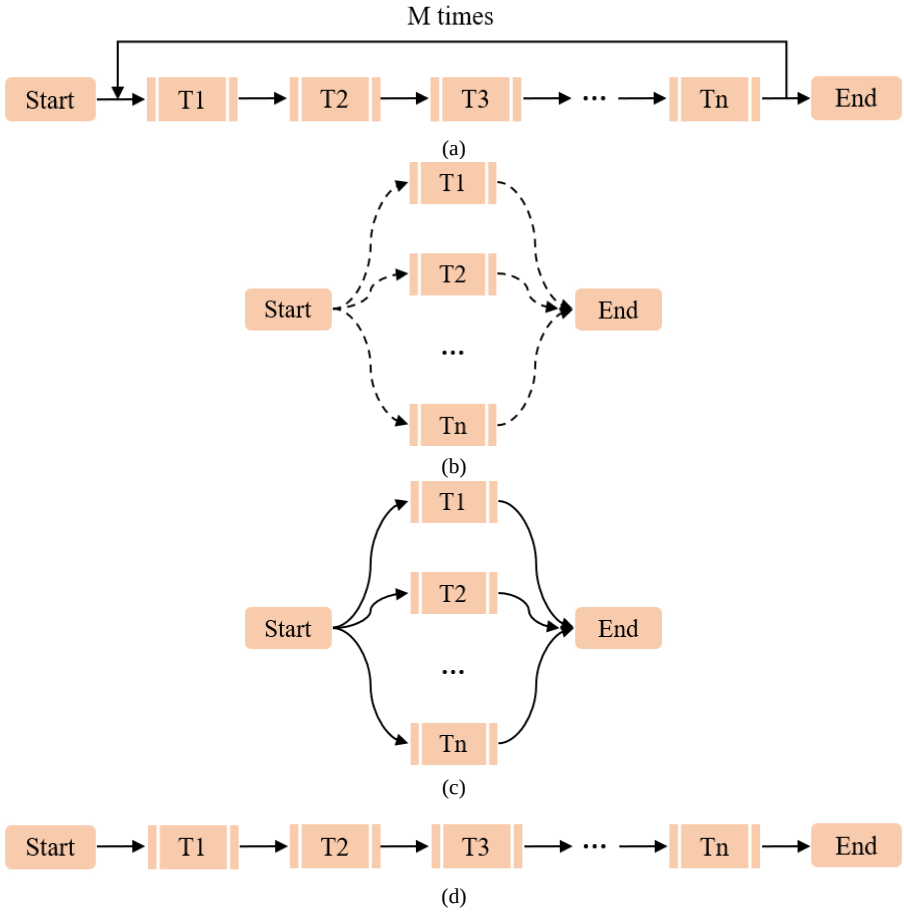


Fig. 1. IoT service composition structures: (a) Loop, (b) Selection, (c) Parallel, and (d) Sequential.

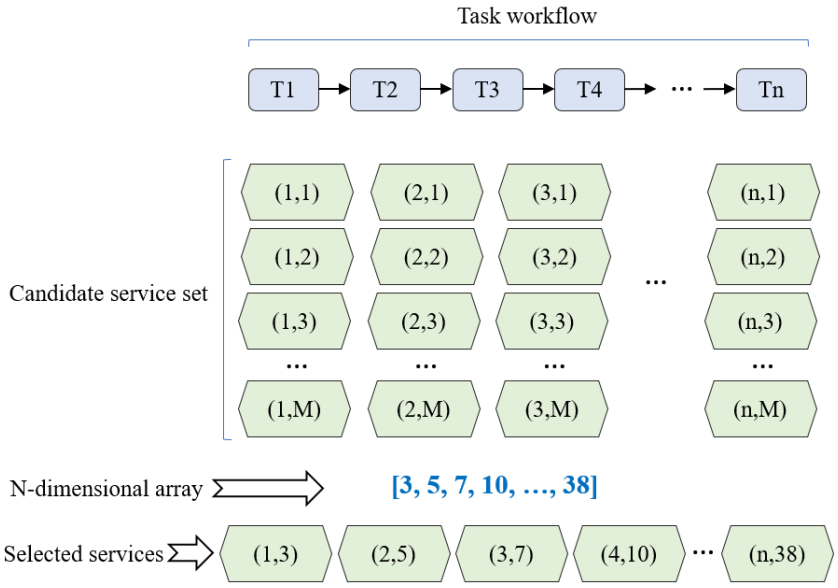


Fig. 2. Optimization process for selecting optimal IoT services from candidate sets.

The IoT service composition involves selecting specific services from a large pool of candidates to satisfy user requirements and QoS constraints. As illustrated in Fig. 2, the optimization process begins with a task workflow $T_1, T_2, \ldots, T_n$, where each task represents an abstract IoT service. For each task T_i, there exists a candidate service set $S(i, M)$, from which the most optimal service must be selected. Each task Ti has several service options $S(i, 1)$, $S(i, 2)$, …, $S(i, M)$. The selected services from all tasks are represented as an array [3, 5, 7, 10, …, 38], where each number corresponds to the selected service for a specific task. For a system with n tasks and M service candidates per task, the total number of possible combinations is M^n. This combinatorial complexity makes the service composition problem an NP-hard optimization challenge.

Several key QoS metrics are considered to evaluate the effectiveness of an IoT service composition, including reliability, credibility, service cost, and execution time. Reliability indicates the likelihood that the IoT service composition can complete tasks successfully without failure, calculated using Eq. (1) [18].

$$Q_r = \prod_{i=1}^{n} q_i^r \tag{1}$$

Credibility measures the user's trust level in the service composition based on factors such as reputation, expressed by Eq. (2) [19].

$$Q_c = \frac{1}{n}\sum_{i=1}^{n} q_i^c \tag{2}$$

Service cost refers to the monetary cost incurred by the user for utilizing the IoT service composition, calculated by Eq. (3) [20].

$$Q_{co} = \sum_{i=1}^{n} q_i^{co} \tag{3}$$

Execution time represents the total time required for the service composition to execute, including the processing time of all tasks, calculated by Eq. (4) [21].

$$Q_t = \sum_{i=1}^{n} q_i^t \tag{4}$$

The aggregated QoS value of an IoT service composition is calculated as a weighted sum of the above QoS metrics using Eq. (5).

$$Q = \sum_{k\epsilon\{r,c,co,t\}} Q_k \omega_k \tag{5}$$

Where ω_k denotes the weight assigned to each QoS metric. This weight reflects the relative importance of the corresponding attribute in the overall composition.

The objective is to select one service instance for each task in the workflow such that the aggregated QoS value Q is maximized. This optimization ensures that the composition meets user-defined QoS requirements and achieves the best possible performance in terms of execution time, cost, credibility, and reliability.

B. Enhanced Whale Optimization Algorithm

The WOA, introduced by Mirjalili and Lewis [22], is a heuristic optimization approach inspired by the hunting behavior of humpback whales. The algorithm mimics the whales' bubble-net feeding strategy as its central mechanism for solving optimization problems. The overall process of WOA is illustrated in Fig. 3. WOA consists of three main phases: encircling prey, bubble-net feeding, and searching for prey.

These phases emulate the whales' local exploitation and global exploration strategies, making WOA a versatile optimization framework.

The first phase of WOA involves encircling the prey, which represents the best solution found so far. Whales are assumed to position themselves around the prey to prepare for attack. Mathematically, this behavior is modeled using Eq. (6) and (7).

$$\vec{D} = \left|\vec{C}.\vec{X}_{best}(t) - \vec{X}(t)\right| \quad (6)$$

$$\vec{X}(t+1) = \vec{X}_{best}(t) - \vec{A}.\vec{D} \quad (7)$$

Where $\vec{A}$ *and* $\vec{C}$ are coefficient vectors, $\vec{D}$ stands for the distance between the whale and the prey, $\vec{X}(t)$ is the current position of the whale, and $\vec{X}_{best}(t)$ is the position vector of the best solution (prey) at iteration t. The vectors $\vec{A}$ and $\vec{C}$ are computed using Eq. (8) and (9).

$$\vec{A} = 2.\vec{a}.\vec{r}_1 - \vec{a} \quad (8)$$

$$\vec{C} = 2.\vec{r}_2 \quad (9)$$

Where $\vec{r}_1$ and $\vec{r}_2$ range between [0, 1], and $\vec{a}$ drops linearly from 2 to 0 with increasing iterations.

The bubble-net feeding phase simulates whales' two simultaneous strategies to capture prey: shrinking, encircling, and spiral updating. These strategies reflect both global and local search mechanisms.

Shrinking encircling reduces the distance between whales and prey over time by decreasing the range of $\left|\vec{A}\right|$. This is accomplished by progressively lowering the value of $\vec{a}$. Spiral updating mimics the spiral-shaped trajectory of whales around their prey. This phenomenon is represented mathematically using Eq. (10) and (11).

$$\vec{X}(t+1) = \vec{D}_p.e^{bl}.cos(2\pi l) + X_{best}(t) \quad (10)$$

$$\vec{D}_p = \left|\vec{X}_{best}(t) - \vec{X}(t)\right| \quad (11)$$

Where $\vec{D}_p$ refers to the distance between the whale and the prey, b defines the shape of the logarithmic spiral, and l signifies a random number in the range [-1, 1].

To combine these two strategies, WOA uses a probabilistic mechanism where a random number P determines which behavior is applied in each iteration, calculated using Eq. (12).

$$\vec{X}(t+1) = \begin{cases} \vec{D}_p.e^{bl}.cos(2\pi l) + X_{best}(t), & P < 0.5 \\ \vec{X}_{best}(t) - \vec{A}.\vec{D}, & P \geq 0.5 \end{cases} \quad (12)$$

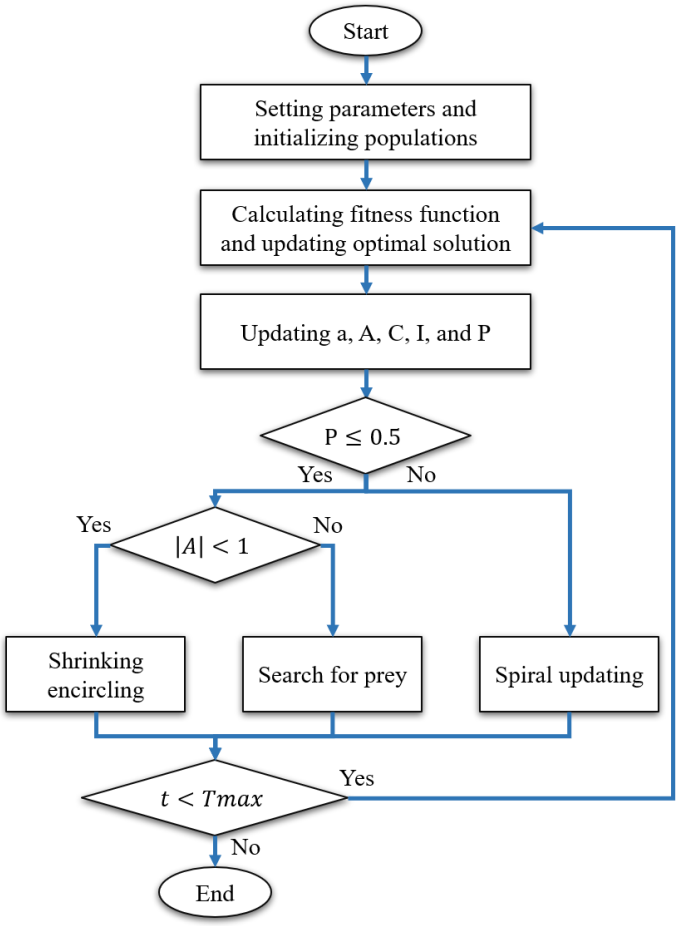


Fig. 3. WOA flowchart.

This probabilistic combination of behaviors balances global exploration and local exploitation. The final phase focuses on exploration by searching for prey. When the condition $|\vec{A}| \geq 1$ is satisfied, and whales move randomly for better solutions. This behavior is modeled using Eq. (13) and (14).

$$\vec{D} = \left|\vec{C}.\vec{X}_{rand}(t) - \vec{X}(t)\right| \quad (13)$$

$$\vec{X}(t+1) = \vec{X}_{rand}(t) - \vec{A}.\vec{D} \quad (14)$$

Where $\vec{X}_{rand}(t)$ is the position vector of a randomly selected whale. This phase prevents WOA from getting stuck on local optimum and enhances the algorithm's global search capability.

The enhanced WOA (EWOA) builds upon the original WOA by addressing its inherent shortcomings, such as slow convergence, poor accuracy, and proneness to local optima. This improvement is achieved by integrating a nonlinear cross-weight mechanism and the Fibonacci Search Method (FSM). The enhanced algorithm ensures optimal equilibrium between diversification (exploration) and intensification (exploitation), key components of any robust optimization method. Fig. 4 presents the pseudocode of the EWOA.

Input:
Initialize population $X = \{X_1, X_2, \ldots, X_n\}$
Define objective function $f(X)$
Set algorithm parameters: maximum iterations (T_{max}) and control parameters (α, A, C, I, P)
Step 1: Initialize and Evaluate
Evaluate the fitness of each solution in the population.
Identify the best solution $BEST_{sol}$.
Step 2: Iterative optimization
While iteration t is less than T_{max}
Update control parameters.
Exploration and exploitation decision:
if $P \leq 0.5$ (Exploration phase):
if $|A| < 1$:
Perform shrinking encircling mechanism to update position.
Else $|A| \geq 1$:
Conduct random search for prey to enhance exploration.
Else if $P > 0.5$ (Exploitation phase):
Perform spiral updating to refine solution accuracy.
Boundary Verification:
Check if any search agents exceed the search space limits and correct their positions if necessary.
Fitness Evaluation and Update:
Calculate the fitness of each solution.
If a better solution is found, update $BEST_{sol}$.
Increment iteration count $t = t + 1$.
End While
Output:
Return the best solution found $BEST_{sol}$

Fig. 4. Pseudocode of the EWOA.

Diversification explores the entire search space to identify global optima, while intensification involves focusing on local regions for fine-tuning solutions. The original WOA struggles with achieving an optimal balance between these two processes, often leading to stagnation at local optima. EWOA addresses this by introducing a nonlinear crossover weight to enhance solution diversity during exploration and incorporating the FSM for a more efficient local search, improving solution accuracy and convergence speed.

The EWOA incorporates crossover weights during location updates. The revised position update model is defined using Eq. (15).

$$\vec{X}(t+1) = \begin{cases} (\vec{X}_{best}(t) - \vec{A}.\vec{D}).CR1, & if\ a < 0.5 \\ \vec{D}'.e^{bl}.cos(2\pi l).CR2 + \vec{X}_{best}(t).(1 - CR2), & if\ a \geq 0.5 \end{cases} \quad (15)$$

Where $\vec{D}'$ stands for the distance between the current and best solutions, $CR1 = exp\left(tan\left(rand(1,N)\right)\right)$, $CR2 = 3.\left(0.5 - rand(1,N)\right).f$, and $f = exp\left(-\frac{Iteration}{Max\_iteration}\right)$. This mechanism ensures that solutions generated during exploration maintain sufficient diversity while enabling convergence during exploitation.

FSM is incorporated into EWOA to improve local search efficiency. FSM minimizes the search space by applying Fibonacci sequences, which guide the selection of optimal intervals. It operates as follows:

Generate Fibonacci numbers: The Fibonacci sequence $F = [F_1, F_2, \ldots, F_n]$ is expressed by Eq. (16).

$$F_n = F_{n-1} + F_{n-2},\ \ F_0 = 1,\ \ F_1 = 1 \quad (16)$$

Calculate initial points: Two points t_1 and t_2 are defined in the search range [LL, UL] using Eq. (17).

$$t_1 = LL + \frac{F_{n-2}}{F_n}.(UL - LL)$$
$$t_2 = UL - \frac{F_{n-2}}{F_n}.(UL - LL) \quad (17)$$

Where UL and LL define the upper and lower bounds of the range.

Evaluate function values: Compare the function values at t_1 and t_2:

If $(t_2) > (t_1)$, shift the range to the left.

If $(t_1) > (t_2)$, shift the range to the right.

IV. RESULTS

The effectiveness of the proposed EWOA was evaluated for optimizing the QoS-based IoT service composition optimization problem. EWOA effectiveness was evaluated against standard WOA [22], DALOA [14], and Genetic Algorithm (GA) [23] using three key evaluation criteria: effectiveness, convergence, and stability. The tests were carried out in a Windows 10 system powered by an Intel Core i7-12700F processor, 16 GB RAM, and PyCharm Community Edition 2022.3.

The experiments used randomly generated datasets based on the QoS value ranges defined in Table II. Four QoS attributes, including execution time, service cost, credibility, and reliability, were evaluated for candidate IoT service instances. The dataset scales were represented as $A \times I$, where A denotes the number of abstract service tasks, and I signifies the number of candidate services per task. The datasets included the following scales: 10×50, 10×100, 20×50, 20×100, 30×50, and 30×100. Each experiment was repeated 100 times to ensure robustness, and the results were analyzed to measure the algorithm's performance under varying scales and iterations.

TABLE II. QOS VALUE RANGES

Attributes	Reliability	Credibility	Service cost	Execution time
Ranges	(0.1,1]	(2,10]	(0,100]	(0,60]

The effectiveness of the algorithms was assessed by the average fitness values obtained after 100 global iterations. EWOA demonstrated significantly higher efficacy than WOA, DALOA, and GA, as shown in Table III and Fig. 5-7. Key observations include:

- For smaller scales (e.g., 10×50), EWOA slightly outperformed other algorithms, achieving higher-quality solutions.
- For larger scales (e.g., 30×100), EWOA's advantage became more evident, delivering higher fitness values with a broader margin.

TABLE III. FITNESS VALUES FOR VARIOUS SERVICE COMPOSITION SCALES

No. of abstract service tasks	No. of candidate services	Fitness values			
		GA	DALOA	WOA	EWOA
10	50	3.75	4.13	4.22	4.92
	100	4.02	4.21	4.36	4.85
20	50	6.49	6.75	7.06	9.53
	100	7.16	7.53	8.12	9.61
30	50	9.15	9.91	10.58	13.91
	100	10.13	11.54	12.02	14.05

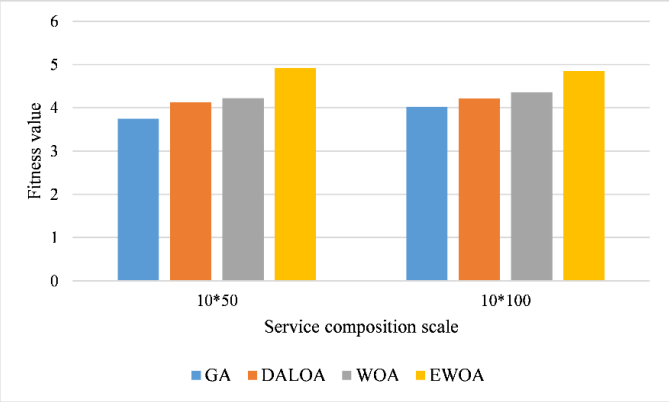


Fig. 5. Fitness value comparison for 10 service tasks.

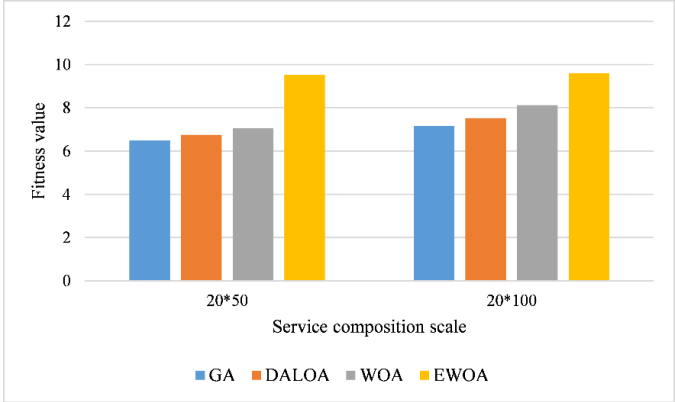


Fig. 6. Fitness value comparison for 20 service tasks.

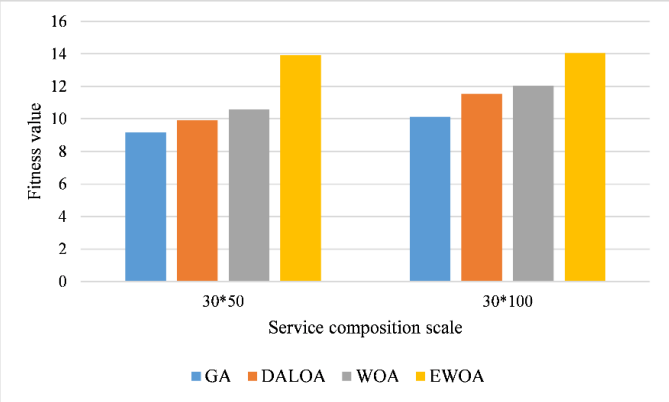


Fig. 7. Fitness value comparison for 30 service tasks.

The nonlinear crossover weights dynamically adjust exploration/exploitation, ensuring sufficient solution diversity at the beginning and accurate refinement in later iterations. The Fibonacci search strategy effectively narrowed the search space, leading to higher solution accuracy and better optimization of QoS attributes.

The convergence performance of the algorithms was analyzed based on their fitness values over iterations, as depicted in Fig. 8 – Fig. 10. EWOA consistently converged faster and to better solutions than WOA, DALOA, and GA. Notable findings include:

- At smaller scales (e.g., 10×50), EWOA achieved convergence within fewer iterations than other algorithms.
- At larger scales (e.g., 30×100), EWOA showed a significant fitness advantage and faster convergence speed.

The nonlinear crossover weights ensured efficient exploration in the early iterations, preventing premature convergence to local optima. The Fibonacci search strategy refined the best solutions in the exploitation phase, accelerating convergence toward the global optimum.

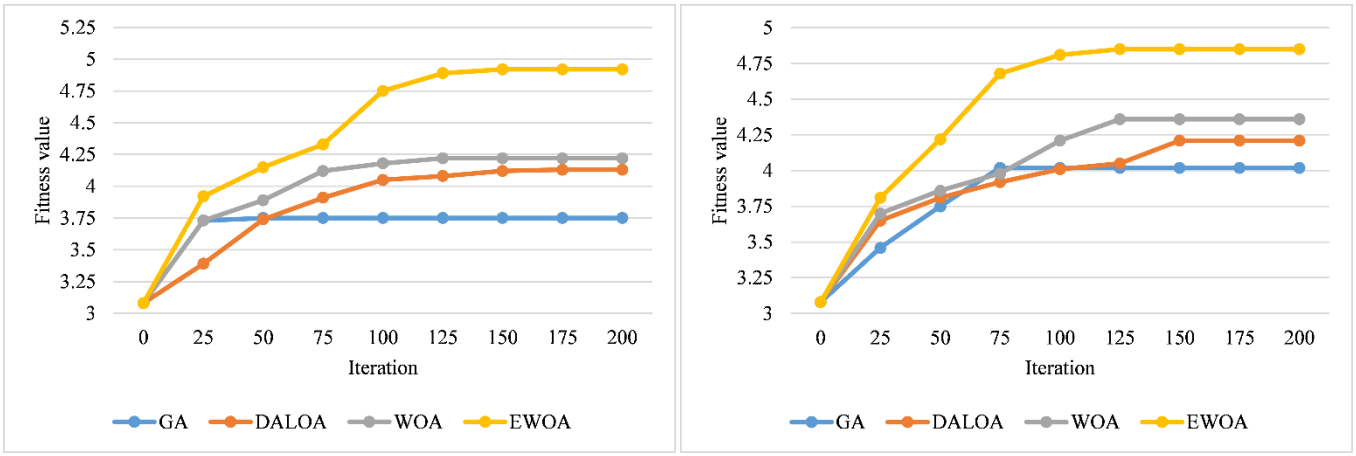


Fig. 8. Convergence performance comparison: (a) 10 service tasks and 50 candidate services, (b) 10 service tasks and 100 candidate services.

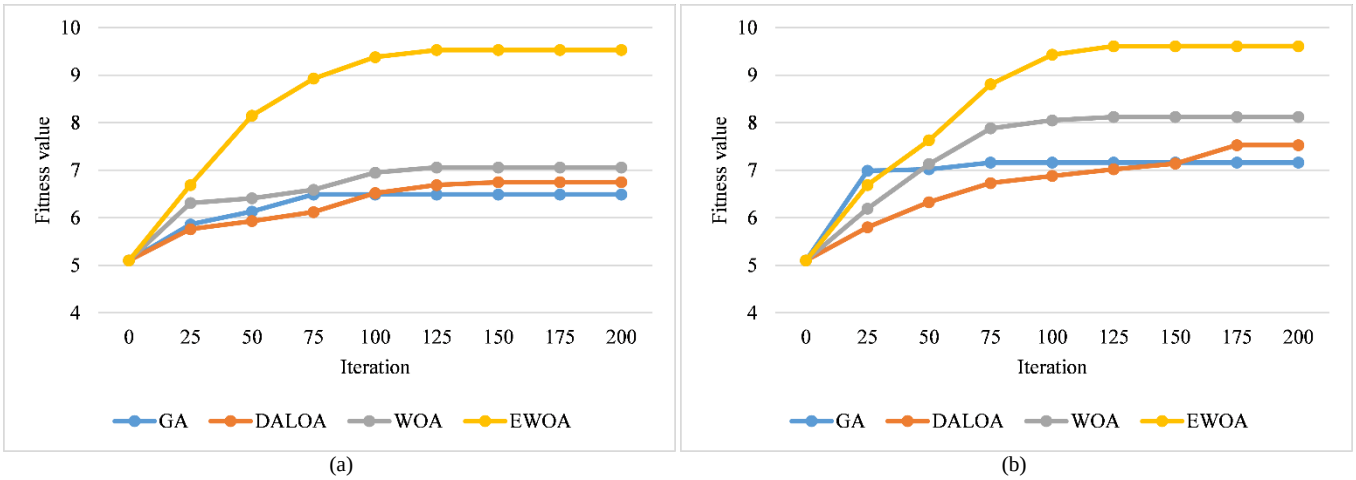


Fig. 9. Convergence performance comparison: (a) 20 service tasks and 50 candidate services, (b) 20 service tasks and 100 candidate services

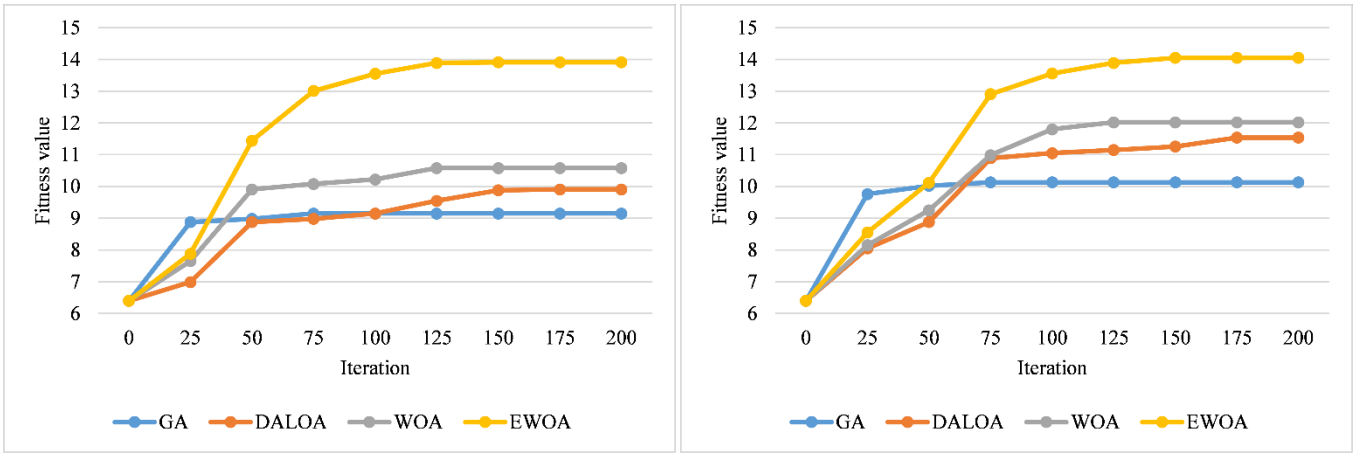


Fig. 10. Convergence performance comparison: (a) 30 Service tasks and 50 candidate services, (b) 30 Service tasks and 100 candidate services.

TABLE IV. STANDARD DEVIATION FOR VARIOUS SERVICE COMPOSITION SCALES

No. of abstract service tasks	No. of candidate services	Standard deviation			
		GA	DALOA	WOA	EWOA
10	100	0.0831	0.0526	0.0415	0.0089
20	100	0.1173	0.1085	0.1019	0.0269
30	100	0.1569	0.1503	0.1494	0.0612

The stability of the algorithms was measured in terms of standard deviation of optimal fitness values across 100 experiments, as shown in Table IV and Fig. 11. Lower standard deviation values indicate higher stability. Key findings include:

- EWOA exhibited significantly lower standard deviation values compared to WOA, DALOA, and GA, particularly at larger scales (e.g., 30×100).
- As scales increased, the standard deviation of all algorithms rose. However, EWOA maintained superior stability, consistently delivering reliable results with minimal variation.

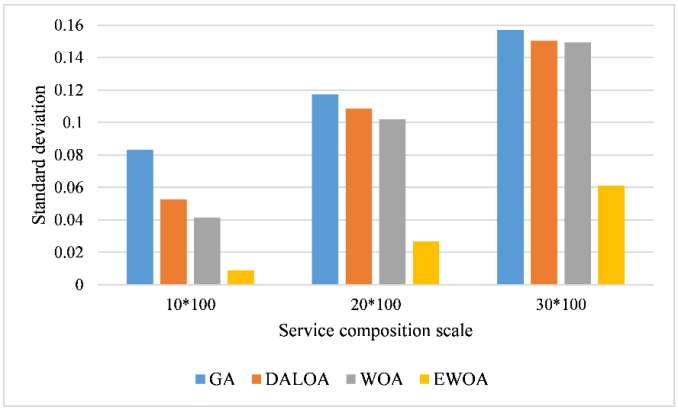


Fig. 11. Standard deviation comparison.

The nonlinear crossover weights provided adaptive adjustments, ensuring robustness against variations in initial population diversity. The Fibonacci search strategy reinforced solution refinement, minimizing the impact of random fluctuations in search trajectories.

V. DISCUSSION

The results demonstrate that the proposed EWOA significantly improves the optimization of QoS-based IoT service composition compared to standard WOA, DALOA, and GA. One of the key advantages of EWOA is its superior performance across various dataset scales, particularly for larger problem instances, where it consistently achieved higher-quality solutions. The nonlinear crossover weights effectively balanced exploration and exploitation, preventing premature convergence and ensuring sustained search diversity. Additionally, the integration of the Fibonacci search strategy enabled precise solution refinement by efficiently narrowing the search space, ultimately leading to better fitness values and improved QoS attribute optimization. These results validate the effectiveness of the proposed modifications, particularly in handling complex IoT service composition scenarios where scalability and solution accuracy are critical.

Furthermore, the convergence analysis confirms that EWOA consistently outperforms its counterparts in both speed and solution quality. The rapid convergence observed in smaller-scale datasets indicates that EWOA is highly effective even for less complex problems. However, its performance advantage becomes more pronounced as the problem scale increases, demonstrating its robust scalability. Additionally, stability analysis reveals that EWOA maintains lower standard deviation values, signifying its ability to deliver consistent and reliable results across multiple runs. This is primarily due to the adaptive adjustments of nonlinear weights, which dynamically regulate search behavior, and the Fibonacci search refinement, which enhances exploitation precision. The findings underscore the suitability of EWOA for large-scale IoT service

composition problems, offering a highly efficient, stable, and scalable optimization framework.

VI. CONCLUSION

This study proposed EWOA to resolve the QoS-based IoT service composition optimization problem. EWOA mitigated some disadvantages of conventional optimization algorithms, particularly slow convergence, the tendency toward local optima, and inability to balance exploration and exploitation. With nonlinear crossover weights combined with the Fibonacci search strategy, the EWOA optimized global exploration and local exploitation, providing outstanding performance in achieving optimization. The experimental test validated the efficiency of EWOA under different composition scenarios, from small-scale to large-scale problems. EWOA had better fitness, higher convergence speed, and more substantial stability than WOA, DALOA, and GA in all cases. Dynamic adjustment of the nonlinear crossover weights regulated solution diversity and refinement during optimization, whereas the Fibonacci search strategy improved efficiency in local search and prevented falling into suboptimal solutions.

EWOA proved to be especially helpful in large-scale composition, while its ability to handle increased problem complexity led to significant gains over existing algorithms. Moreover, its computational efficiency and stability over multiple runs indicated the appropriateness of real-world IoT scenarios. Future work will focus on integrating the EWOA with dynamic composition models to handle real-time and evolving QoS requirements. This could be further enhanced by hybridizing the EWOA with other metaheuristics for optimization problems that are highly complex and multidimensional. Given its capability, the EWOA represents a promising trend toward development in the field of optimization solutions within an IoT context.

REFERENCES

[1] A. Choudhary, "Internet of Things: a comprehensive overview, architectures, applications, simulation tools, challenges and future directions," Discover Internet of Things, vol. 4, no. 1, p. 31, 2024.

[2] B. Pourghebleh, N. Hekmati, Z. Davoudnia, and M. Sadeghi, "A roadmap towards energy - efficient data fusion methods in the Internet of Things," Concurrency and Computation: Practice and Experience, vol. 34, no. 15, p. e6959, 2022.

[3] L. Camargo, J. Pauletti, A. M. Pernas, and A. Yamin, "VISO approach: A socialization proposal for the Internet of Things objects," Future Generation Computer Systems, vol. 150, pp. 326-340, 2024.

[4] V. Hayyolalam, B. Pourghebleh, M. R. Chehrehzad, and A. A. Pourhaji Kazem, "Single - objective service composition methods in cloud manufacturing systems: Recent techniques, classification, and future trends," Concurrency and Computation: Practice and Experience, vol. 34, no. 5, p. e6698, 2022.

[5] D. K. K. Reddy, J. Nayak, H. Behera, V. Shanmuganathan, W. Viriyasitavat, and G. Dhiman, "A systematic literature review on swarm intelligence based intrusion detection system: past, present and future," Archives of Computational Methods in Engineering, vol. 31, no. 5, pp. 2717-2784, 2024.

[6] Z. Zhang, L. Tan, D. Martín, L. Qian, M. Khishe, and P. Jangir, "A dual-adaptive stochastic reinforcement chimp optimization algorithm for fire detection and multidimensional problem solving," Scientific Reports, vol. 14, no. 1, p. 31226, 2024.

[7] S. Paul, S. De, and S. Bhattacharyya, "Emerging trends in computational swarm intelligence: A comprehensive overview," Recent Trends in Swarm Intelligence Enabled Research for Engineering Applications, pp. 1-40, 2024.

[8] M. Hosseinzadeh et al., "A hybrid service selection and composition model for cloud-edge computing in the internet of things," IEEE Access, vol. 8, pp. 85939-85949, 2020, doi: https://doi.org/10.1109/ACCESS.2020.2992262.

[9] S. Sefati and N. J. Navimipour, "A qos-aware service composition mechanism in the internet of things using a hidden-markov-model-based optimization algorithm," IEEE Internet of Things Journal, vol. 8, no. 20, pp. 15620-15627, 2021.

[10] S. Berrani, A. Yachir, B. Djamaa, S. Mahmoudi, and M. Aissani, "Towards a new semantic middleware for service description, discovery, selection, and composition in the Internet of Things," Transactions on Emerging Telecommunications Technologies, vol. 33, no. 9, p. e4544, 2022.

[11] M. Hamzei, S. Khandagh, and N. Jafari Navimipour, "A quality-of-service-aware service composition method in the internet of things using a multi-objective fuzzy-based hybrid algorithm," Sensors, vol. 23, no. 16, p. 7233, 2023.

[12] A. Vakili, H. M. R. Al - Khafaji, M. Darbandi, A. Heidari, N. Jafari Navimipour, and M. Unal, "A new service composition method in the cloud - based internet of things environment using a grey wolf optimization algorithm and MapReduce framework," Concurrency and Computation: Practice and Experience, vol. 36, no. 16, p. e8091, 2024.

[13] G. Xiao, "Toward Optimal Service Composition in the Internet of Things via Cloud-Fog Integration and Improved Artificial Bee Colony Algorithm," International Journal of Advanced Computer Science & Applications, vol. 15, no. 5, 2024.

[14] S. Ait Hacène Ouhadda, S. Chibani Sadouki, A. Achroufene, and A. Tari, "A Discrete Adaptive Lion Optimization Algorithm for QoS-Driven IoT Service Composition with Global Constraints," Journal of Network and Systems Management, vol. 32, no. 2, p. 34, 2024.

[15] X. Liu and Y. Deng, "A new QoS-aware service discovery technique in the Internet of Things using whale optimization and genetic algorithms," Journal of Engineering and Applied Science, vol. 71, no. 1, p. 4, 2024.

[16] B. Pourghebleh, V. Hayyolalam, and A. Aghaei Anvigh, "Service discovery in the Internet of Things: review of current trends and research challenges," Wireless Networks, vol. 26, no. 7, pp. 5371-5391, 2020.

[17] B. Pourghebleh and N. J. Navimipour, "Data aggregation mechanisms in the Internet of things: A systematic review of the literature and recommendations for future research," Journal of Network and Computer Applications, vol. 97, pp. 23-34, 2017, doi: https://doi.org/10.1016/j.jnca.2017.08.006.

[18] D. Rupanetti and N. Kaabouch, "Combining Edge Computing-Assisted Internet of Things Security with Artificial Intelligence: Applications, Challenges, and Opportunities," Applied Sciences, vol. 14, no. 16, p. 7104, 2024.

[19] L. Wei, Y. Yang, J. Wu, C. Long, and B. Li, "Trust management for Internet of Things: A comprehensive study," IEEE Internet of Things Journal, vol. 9, no. 10, pp. 7664-7679, 2022.

[20] C. Liu, Z. Su, X. Xu, and Y. Lu, "Service-oriented industrial internet of things gateway for cloud manufacturing," Robotics and Computer-Integrated Manufacturing, vol. 73, p. 102217, 2022.

[21] H. Gao, Y. Zhang, H. Miao, R. J. D. Barroso, and X. Yang, "SDTIOA: modeling the timed privacy requirements of IoT service composition: a user interaction perspective for automatic transformation from BPEL to timed automata," Mobile Networks and Applications, pp. 1-26, 2021.

[22] S. Mirjalili and A. Lewis, "The whale optimization algorithm," Advances in engineering software, vol. 95, pp. 51-67, 2016.

[23] Q. Li, R. Dou, F. Chen, and G. Nan, "A QoS-oriented Web service composition approach based on multi-population genetic algorithm for Internet of things," International Journal of Computational Intelligence Systems, vol. 7, no. Suppl 2, pp. 26-34, 2014.

SQRCD: Building Sustainable and Customer Centric DFIS for the Industry 5.0 Era

Ruchira Rawat[1], Himanshu Rai Goyal[2], Sachin Sharma[3], Bina Kotiyal[4]
Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, India[1, 2]
Amity School of Engineering and Technology, Amity University, Punjab, India[3]
Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun, India[4]

***Abstract*—Artificial Intelligence (AI) is considered a big turning point for the financial industry. Introducing Artificial General Intelligence (AGI) enhances the capability of all the areas where AI shows its power. The development of AGI is directly proportional to the need for more advanced automation by enhancing the features quick responsive, customization/personalization and the refined decision making capabilities in different industries. The current study aims to discuss the respondent's views on the adoption of an AGI-enabled Sustainability, Quick responsiveness, Risk management, Customer-centric, and Data privacy (SQRCD) system in Digital Financial Inclusion System (DFIS). A total of 630 responses were collected from the respondents belonging to 90 different finance institutes. The result shows that SQRCD had a significant positive relationship with the attitude to adopt an AGI enabled-SQRCD system. The three cultural dimensions of Hofstede's theory power distance index, collectivism-individualism, and uncertainty acceptance are also taken as moderators. The effect of moderators is seen in the different relationships. The study develops a direct hypothesis to analyze the adoption of a new financial system which includes the mentioned factors. The result of the study is beneficial in the development of a renewed financial system where the mentioned parameters are essential for Industry 5.0.**

Keywords*—Artificial General Intelligence (AGI); Digital Financial Inclusion System (DFIS); industry 5.0; customer-centric; sustainability

I. Introduction

AI is considered as a big turning point for the financial industry. It plays an important role in process automation, consumer service and risk management [1]. Introducing AGI enhances the capability of all the area where AI is showing its power. AGI is the domain of AI which can perform tasks as human do and can learn and solve the problem in various domains. With the new upcoming technologies like big data, machine learning, cloud computing scientists are able to form many newer AI algorithms which are able to analyze the large amount of data and can predict and learn human intelligence [2]. The development of AGI is directly proportional to the need of more advanced automation by enhancing the feature quick responsive, customization/personalization and the refine decision making capabilities in different industries. Financial inclusion is the practice of making less expensive financial services and products in reach to both consumers and businesses is termed as financial inclusion. It seeks to guarantee that everyone, especially those with low incomes or disadvantages, get financial services in reduced cost. The use of digital technologies to offer financial services to under privileged people is known as DFIS. For people who are currently unbanked or financially marginalized, DFIS seeks to make financial services sustainable, reasonable and accessible. Among the many advantages of DFIS are reduced expenses, a decreased chance of financial crimes, and economic empowerment. A new stage of industrial development called Industry 5.0 highlights collaboration and cooperation between intelligent machines and people [3]. A civilization where digital technology is pervasive and permeates every part of daily life is known as a digital society. It is distinguished by the use of information and communication technology (ICT) for communication, teaching, purchasing, and selling, among other purposes. The paper is organized as follows. The literature review is covered in Section II, the proposed methodology is presented in Section III, the results and discussion are covered in Section IV, and finally, the conclusion is presented in Section V.

II. Literature Review

AI algorithm analyse a large amount of data to identify the pattern by which a fraudulent activity can be predict [4]. AI tools are playing an important role in the area of customer service to assist customer in various field. Personalization is one of the features of AGI. Advance AI algorithms are able to find the patterns and can predict the behaviour of customer and according to that can suggest service that suits to a particular customer. This is the way to fulfil everyone's need according to the choice and it will enhance the financial planning. Automation is another feature of AGI [5]. AGI can examine the large amount of data and enable the financial institution to take quick decisions about the information, service, product, investment, fraud etc. It will improve the efficiency of finance sector as well as it leads to reduced human error. Many insurance businesses have large data quantities that are difficult to handle due to some semi-automated or use of error-prone manuals. Managing human and machine-generated data is tedious because it lacks a uniform Master Data Management (MDM) system. Myriad technologies drive it challenging to provide data accuracy, consequently, steadfast systems are directed toward wisdom that can be used. Thus, it shows the need for AI and ML keys to enhance productivity. AI and ML are changing insurance businesses by enhancing functioning efficiency, increasing fraud detection, personalizing customer interactions, predictive perspicuity through the processing of big data, and enhancing sales approaches [6]. These advances offer tailored

solutions, resulting in higher customer satisfaction by providing more immediate service and lessened costs. By improving customer experience, personalization, data processing, fraud detection, and operational effectiveness. AI and ML are improving the insurance industry. As revealed by Ageas UK's image identification for cutting expenses, auto claims, delays and automation expedites claim processing. An algorithm that enhances fraud detection is US Lemonade, the US's anti-fraud algorithm. AI-based chatbots and virtual assistance enhance Customer satisfaction. AI improves goal-oriented marketing, increasing industry profitability and inducing leads [7]. According to the researcher, e-finance is a fast-developing topic that needs prevention from financial disasters to reduce financial errors. It concentrates on vital topics like the growing prominence of FinTech, bankruptcy prediction, fraud detection, decision-making, and creditworthiness evaluation. ML models like Support Vector Machines (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN) have proven to be very predictive in the field of finance. It shows how AI is being used in finance. The work presented offers a complete study of AI's revolutionary impact on digital banking. It also spotlights the possible current and future developments [8]. The author [9] compares that how AGI is different from ANI and ASI. He discusses about the methods to achieve AGI which include human brain emulation and AIXI (a theoretical model). The paper analyzes the requirement and feasibility of each method based upon technologies and finds that integrated cognitive architectures are the most appropriate option for achieving AGI. The author [10] highlight the ethical and social challenges in the transition from machine learning to AGI. The author [11] states that the prime focus of AGI is to replicate human cognitive abilities in various domain. Although the technologies like big-data and deep learning are in development process but the achievement of AGI is still uncertain. The research describes about the role of AI in finance sector. Like it automates operations, improve efficiency and reduce cost. It also offers personalized service. By analysing data in real time AI helps in detecting fraud and assess risks. By improving risk management, automating repetitive operations, streamlining financial procedures, boosting investment management, guaranteeing regulatory compliance, and facilitating improved decision-making, the incorporation of AI technology in corporate finance offers substantial benefits [12] [13]. The author [14] state about the different domain of AI like ANI-weak AI and AGI-strong AI. Further [15] describes about the difference of ANI and AGI by stating that ANI specializes in specific tasks and AGI can perform tasks in multiple domains as human-like level. AGI is more versatile as compare to ANI. The potential of AGI in Industry 4.0, Industry 5.0 and Society 5.0 is discussed by [16]. AGI has potential to revolutionize the whole sectors by improving productivity, creativity and customization. Although the number of obstacles is there for its implementation including ethics, safety, data privacy, trust etc. For appropriate development of AGI a multidisciplinary strategy which include government, academia and civil society is needed. In China, the author [17] examines the alleviating impacts of digital inclusive finance on the financing restrictions encountered by small and medium enterprises (SMEs). It implements a statistical technique in a two-fold fixed effect model. It is responsible for data collection from individuals and firms and considers both i.e. time-specific and entity-specific effects. This model helps to reduce the constraints in digital inclusive finance. It also underlines the different characteristics of this alleviation such as regional mismatches, domain, and attributes in traditional financial services. The work finalizes that digital inclusive finance improves SMEs' inner funding origins by declining financing costs and managing leverage levels. To assess the credibility of AI in finance, the author introduces a cluster of integrated metrics named as SAFE stands for (Sustainability, Accuracy, Fairness, Explainability). It employs two statistical techniques Lorenz curve and Lorenz Zonoids. The work spotlights the limitations of ML in real-time d, notably concerning extreme data events and the demand for vital metrics. The author [18] expand earlier work to develop metrics. The SAFE framework has outperformed traditional metrics as per the practical demonstration using bitcoin price data. The suggested metrics seek to help stakeholders, financial authorities as well as asset management firms by delivering reliable evaluations of AI methods in finance, eventually improving conviction in AI applications. The author [19] examines the role of AI in improving access to digital finance within the inclusive financial system. The paper focuses on the significant benefits of employing robo advisor, while also discussing the potential risks posed by AI like market manipulation and system risks. Employing automated services will restrict the high commission charges thus lowering the costs and increasing accessibility for moderate savers. The research emphasizes the importance of a robust legal and regulatory framework that prioritizes market safety, consumer protection, and market integrity while promoting financial inclusion. The research work also highlights the requirement of prudent AI that strikes a balance between the rights of individual privacy and regulatory. It concerns the possibility of RegTech where AI is the regulatory technology. Sustainability, resilience and human-centric are the foundation of Industry 5.0 discussed by the author [20]. These three points help in creating a balanced industrial environment that incorporates state-of-the-art technologies with social and environmental stewardship. The employment of AI and ML in digital finance delivers improved efficiency [21], enhanced risk management, improved customer experience, and notable cost savings. It allows organizations to concentrate on strategic missions, stimulating invention in business pinnacles. Also increases the customer satisfaction by streamlining operations. To fight against monetary cybercrime via effectual peculiarity detection, automatic feature extraction, and real-time analytics can be enhanced through deep learning. As per [22] it processes enormous datasets, facilitating organizations to pinpoint deceitful stirs promptly and accurately. The author [23] discusses improved productivity and sustainability in manufacturing by incorporating human intelligence with robots. The focus is on human-robot collaboration, environmental sustainability, and the formation of unique roles like Chief Robotics Officer. The authors [24] [25] discusses the losses for users and institutions caused due to credit card fraud that significantly impacts the financial sector. The various challenges related to data are also focused such as

limited public data, high false alarm rates, growing fraud tactics, class imbalance, etc. It scours deep learning strategies compared to traditional algorithms for showing superior performance, and fraud detection, indicating real-world potential on three financial datasets. The author in [26] analyzes the influence of AI on cybersecurity, it examines the nine-banking data in Qatar's banking sector, emphasizing AI's role in improving security, challenges in enactment, potential hazards from AI mishandling, and susceptibilities in AI instruments. It highlights the demand for a competent workforce and regulatory obedience. AGI analysis applies myriad cognitive architectures however their underlying operational likenesses suggest the potential of a unified software framework, enabling easier implementation, analysis, and comparison. Developing such a framework could accelerate AGI's progress through code reuse and standardization [27] [28] [29].

III. Proposed Methodology

The main key features of industry 5.0 are shown in Fig. 1.

1) Human centric approach: Human demands and interests are central to the production process in an industry that is human-centric. Industry 5.0 asks what technology can do for workers rather than what workers can achieve with modern technology.

2) Sustainability goal: By creating circular economy procedures, a sustainable industry assists companies in lessening their environmental effect. Reducing waste, energy use and greenhouse gas emissions, as well as preventing the depletion and deterioration of natural resources, are other sustainability changes.

3) Resilience and adaptability: A resilient industry has a high level of resilience in its industrial production. It can support vital infrastructure during emergencies and is well-prepared for outages.

4) Integration of advanced technologies: Emerging of new technologies like AI, IoT, Blockchain are one of the major key features in Industry 5.0. Lots of new technology should be able to collaborate to achieve the theme of Industry 5.0 [30] [31].

A. Theoretical Foundation

For any new technology, service, or business to be adopted, it is critical to comprehend the culture. Because culture is multifaceted, there are many different meanings for it [32]. Numerous important people made contributions to cultural theory. "The collective mental programming that sets one group of people apart from another" is how Greek philosopher Hofstede defines culture. It has been established through research that culture affects how people adopt new or emerging technology. The culture has an impact on how new technologies are adopted. In order to explore the field of information theory, Hofstede proposed a theory. The power distance index, collectivism vs. individualism, uncertainty avoidance index, masculinity vs. femininity, and long-term vs. short-term are the five national cultural variables that form the basis of the Hofstede's theory [33] [34] [35]. While examining the adoption of technology, the author provided research on the significance of culture. The author claims that people's attitudes and perceptions of different technology are influenced by their culture [36]. It follows that the acceptance of a new, creative autonomous decision-making system depends on the inclusion of cultural factors. Among the five culture-dimensions three-culture dimension is used in the investigation. Fig. 2 shows the conceptual framework based on the proposed hypothesis.

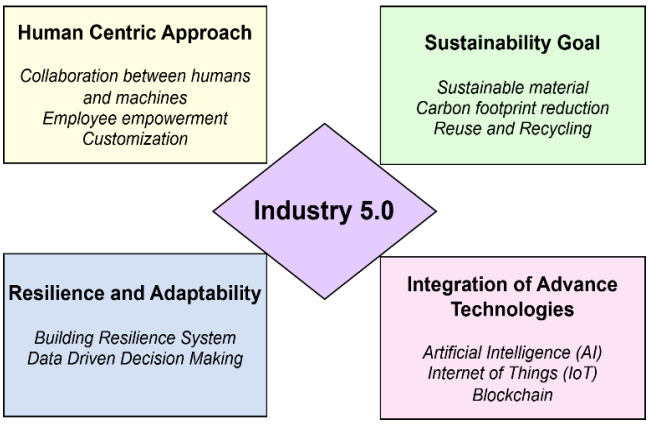


Fig. 1. Key components of industry 5.0.

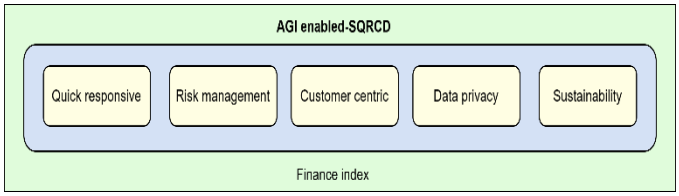


Fig. 2. Conceptual framework.

B. Conceptual Framework and Hypothesis Development

The variable quick responsive, risk management, customer centric, data privacy and sustainability influence the attitude of finance index towards the acceptance of an AGI enabled-SQRCD system. Hofstede's cultural variables power distance index, collectivism-individualism and uncertainty-avoidance moderate these relations. Table I shows the list of survey variable and Table II shows the abbreviation of the variables.

C. Hypothesis Proposed for Finance Index

Fig. 3 shows the conceptual detail framework based on the proposed hypothesis for finance prospective.

H_{qr} Quick responsive: It refers to the response time. The term quick responsive H_{qr} influence positively to the attitude of the finance index towards the acceptance of an AGI enabled- SQRCD.

H_{rm} Risk management: It refers to the capability of risk management. The term risk management H_{rm} influence positively to the attitude of the finance index towards the acceptance of an AGI enabled-SQRCD.

H_{cc} Customer centric: It refers to the process which focus about customer. The term customer centric H_{cc} influence positively to the attitude of the finance index towards the acceptance of an AGI enabled-SQRCD.

TABLE I. SURVEY VARIABLES

Variables	Items
Quick responsive	Finance index should be able to find an AGI enabled-SQRCD system where system is quick responsive while accessing service/product. Finance index should be able to find an AGI enabled-SQRCD system where system is quick responsive while dealing with customers queries.
Risk management	Finance index should be able to find an AGI enabled-SQRCD where system where risk can be managed specially while funding.
Customer centric	Finance index should be able to find an AGI enabled-SQRCD where system is design more in a customer centric way rather than a product centric way.
Data privacy	Finance index should be able to find an AGI enabled-SQRCD system where system where system is able to maintain privacy of data.
Sustainability	Finance index should be able to find an AGI enabled-SQRCD system where the system is able to take challenge to simulate sustainable development.
Power distance index	People who belongs to different income group should be able to find an AGI enabled-SQRCD system
Collectivism	People who belong to collectivist culture will more influence to adopt an AGI enabled-SQRCD system. Individual should be able to find an AGI enabled-SQRCD system
Uncertainty avoidance	People should be able to find an AGI enabled-SQRCD system where instructions are very clear and detailed what to do. People should be able to find an AGI enabled-SQRCD system where instructions and procedures are followed strictly.

TABLE II. ABBREVIATION OF VARIABLES

Variables	Abbreviations
Quick responsive	H_{qr}
Risk management	H_{rm}
Customer centric	H_{cc}
Data privacy	H_{dp}
Sustainability	H_{sn}
Power distance index	H_{pdin}
Collectivism	H_{col}
Uncertainty avoidance	H_{uav}
Attitude	H_{attd}

H_{dp} Data privacy: It refers to the capability of the process where data privacy is one of the important factors. The term data privacy H_{dp} influence positively to the attitude of finance index towards the acceptance of an AGI enabled-SQRCD.

H_{sn} Sustainability: Financial institute has to play very important role to address sustainable development. The term sustainability H_{sn} influence positively to the attitude of the finance index towards the acceptance of an AGI enabled-SQRCD.

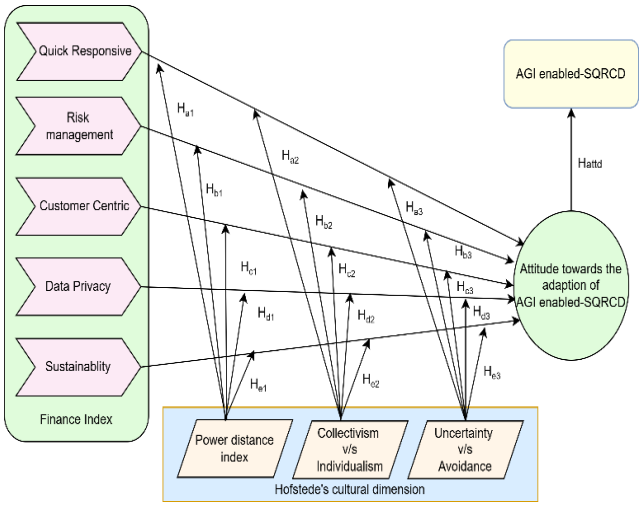


Fig. 3. Conceptual framework for finance index.

Table III shows the hypothesis created for the three constructs taken from the Hofstede's Cultural Dimension.

Fig. 4 shows the flowchart of Customer-Centric. It outlines a structured approach to enhancing customer experience and satisfaction within a financial inclusion system. It starts by identifying target customer segments, focusing on their unique needs and preferences.

TABLE III. HYPOTHESIS CREATED FOR HOFSTEDE'S CULTURAL DIMENSION

Hofstede's Cultural Dimension variables	Items	Description
H_{pdin}	H_{a1}	The relationship between the quick responsive and the attitude towards the adaption of an AGI enabled-SQRCD system is stronger in the society belongs to high/low both power distance index.
	H_{b1}	The relationship between the risk management and the attitude towards the adaption of an AGI enabled-SQRCD system is stronger in the society belongs to high/low both power distance index.
	H_{c1}	The relationship between the customer centric process and the attitude towards the adaption of an AGI enabled-SQRCD system is stronger in the society belongs to high/low both power distance index.
	H_{d1}	The relationship between the term data privacy and the attitude towards the adaption of an AGI enabled-SQRCD system is stronger in the society belongs to high/low both power distance index.
	H_{e1}	The relationship between the term sustainability and the attitude towards the adaption of an AGI enabled-SQRCD system is stronger in the society belongs to high/low both power distance index.
H_{col}	H_{a2}	The relationship between the quick responsive and the attitude towards the adaption of an AGI enabled-SQRCD system is stronger in both collectivism and individualism cultural.
	H_{b2}	The relationship between the risk management and the attitude towards the adaption of an AGI enabled-SQRCD system is stronger in both collectivism and individualism cultural.
	H_{c2}	The relationship between the customer centric process and the attitude towards the adaption of an AGI enabled-SQRCD system is stronger in both collectivism and individualism cultural.
	H_{d2}	The relationship between the term data privacy and the attitude towards the adaption of an AGI enabled-SQRCD system is stronger in both collectivism and individualism cultural.
	H_{e2}	The relationship between the term sustainability and the attitude towards the adaption of an AGI enabled-SQRCD system is stronger in both collectivism and individualism cultural.
H_{uav}	H_{a3}	The relationship between quick responsive and the attitude towards the adaption of an AGI enabled-SQRCD system is stronger in uncertainty avoidance cultural.
	H_{b3}	The relationship between risk management and the attitude towards the adaption of an AGI enabled-SQRCD system is stronger in uncertainty avoidance cultural.
	H_{c3}	The relationship between the customer centric process and the attitude towards the adaption of an AGI enabled-SQRCD system is stronger in uncertainty avoidance cultural.
	H_{d3}	The relationship between the term data privacy and the attitude towards the adaption of an AGI enabled-SQRCD system is stronger in uncertainty avoidance cultural.
	H_{e3}	The relationship between the term sustainability and the attitude towards the adaption of an AGI enabled SQRCD system is stronger in uncertainty avoidance cultural.

The next step involves collecting information through surveys and feedback mechanisms to assess customer expectations. Based on this information, personalized financial products and services are developed to ensure alignment with customer requirements. The chart emphasizes the importance of leveraging technology to provide tailored communication and support. The system incorporates regular measurement of customer satisfaction and service effectiveness, allowing for data-driven improvements using feedback process. Finally, the flow chart highlights the iterative nature of the process, ensuring that customer feedback is consistently integrated to refine offerings, enhance relationships, and foster loyalty, ultimately creating a responsive and adaptive financial service environment.

Fig. 5 shows the flowchart of sustainability-focused financial inclusion system. It outlines a structured approach to enhancing access to financial services for underserved populations while promoting environmental and social responsibility. It begins by identifying the target demographic and assessing their specific financial needs through community engagement. Next, inclusive financial products are developed, leveraging technology for accessibility and personalized education. Partnerships with local organizations are established to build trust and facilitate outreach. Financial literacy programs empower customers with knowledge, while engagement strategies foster strong relationships. The system emphasizes measuring impact through customer satisfaction metrics, allowing for continuous improvement based on feedback. Additionally, it promotes sustainable practices within financial offerings. The iterative process ensures that the system evolves to meet changing customer needs, ultimately creating a comprehensive framework that empowers individuals and communities while supporting sustainable development goals.

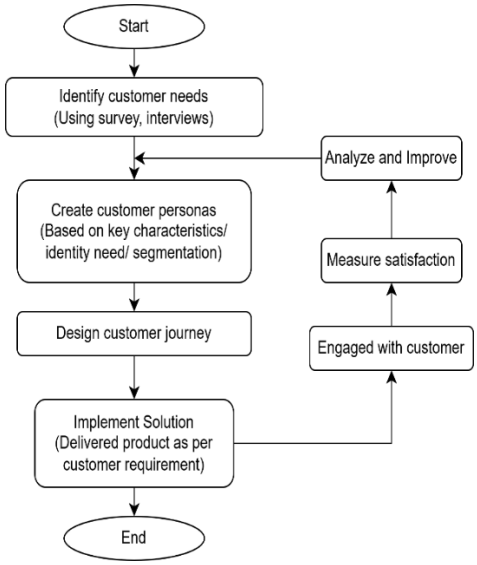


Fig. 4. Flow-chart customer centric focused intelligence financial inclusion system.

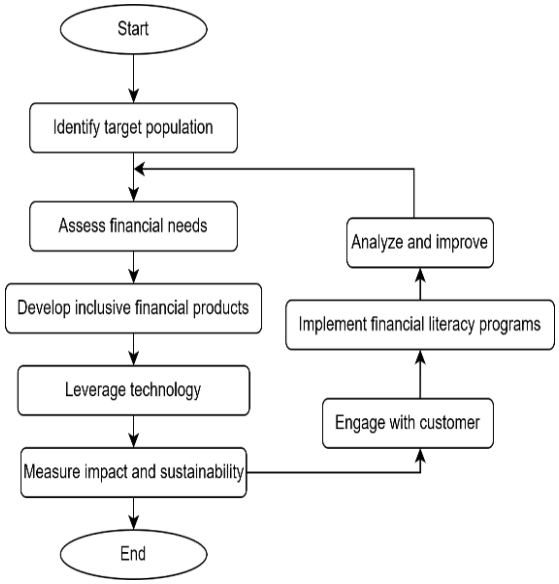


Fig. 5. Flow-chart sustainability focused intelligence financial inclusion system.

IV. Result Analysis

Using a quantitative research approach and a purposive selection technique, 630 financial service providers from various socio-economic backgrounds were selected for the sample. Table IV shows the demographic profile. Specifically, the sample design focused on people who had previously engaged with AI algorithms. Participants were invited to participate using online questionnaires. Five-point Likert scale is used for the value of survey variables, from strongly disagree to strongly agree. 1 is used as strongly disagree and 5 is used as strongly agree. This provided the data needed for a thorough analysis and understanding of the correlation between the constructs and the AGI enabled-SRQCD and also the adaption of AGI enabled-SQRCD.

TABLE IV. Demographic Profile

Type of Bank	
Public	13
Private	21
Foreign	44
Small finance	12
Total Number of Respondents (630)	
Public	91
Private	147
Foreign	308
Small finance	84

Before moving towards the thoroughly analysis the validation of each item's reliability is important. There are total five items are listed in conceptual model which are independent and one is dependent. Three items are taken from Hofstede's cultural dimension as moderators. Descriptive analysis is done to find a true picture of data involved in the study. Descriptive statistics of variable used in survey for study is shown in Table V. The graphs of mean, median, mode and std. deviation are shown in Fig. 6.

TABLE V. Descriptive Satistics

	Mean	**Median**	**Mode**	**Std. Deviation**	**Min**	**Max**
H_{qr}	4.253174603	4.25	4.5	0.54084908	1.25	5
H_{rm}	3.974206349	3.97	4.25	0.506348179	1.5	4.75
H_{cc}	3.929365079	4.25	4.5	0.730313157	1.25	5
H_{dp}	4.180952381	4.5	4.5	0.561052311	1.25	5
H_{sn}	3.986714286	4	4	0.524437011	1.33	5
H_{pdin}	4.104666667	4	4	0.582995499	1.67	5
H_{col}	4.098111111	4	4	0.510755001	1.33	5
H_{uav}	4.043253968	4.25	4.25	0.613032585	1.5	5
H_{attd}	4.253174603	4.25	4.5	0.54084908	1.25	5

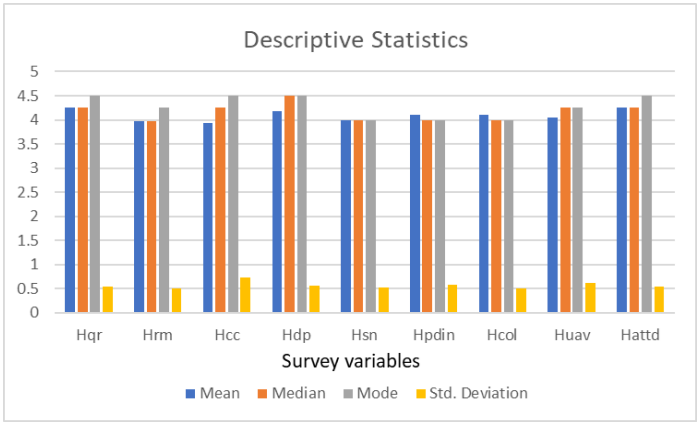


Fig. 6. Descriptive statistics.

Factor loading of each item are shown in Table VI. The value of factor loading exceed 0.7 is maintaining the guideline.

TABLE VI. FACTOR LOADING

Variable	Items	Factor Loadings
H_{qr}	H_{qr1}	0.866
	H_{qr2}	0.823
	H_{qr3}	0.777
	H_{qr4}	0.701
H_{rm}	H_{rm1}	0.711
	H_{rm2}	0.753
	H_{rm3}	0.84
	H_{rm4}	0.781
H_{cc}	H_{cc1}	0.735
	H_{cc2}	0.849
	H_{cc3}	0.842
	H_{cc4}	0.766
H_{dp}	H_{dp1}	0.76
	H_{dp2}	0.804
	H_{dp3}	0.846
	H_{dp4}	0.738
H_{sn}	H_{sn1}	0.781
	H_{sn2}	0.805
	H_{sn3}	0.852
	H_{sn4}	0.71
H_{pdin}	H_{pdin1}	0.719
	H_{pdin2}	0.82
	H_{pdin3}	0.838
	H_{pdin4}	0.711
H_{col}	H_{col1}	0.754
	H_{col2}	0.809
	H_{col3}	0.855
	H_{col4}	0.732
H_{uav}	H_{uav1}	0.822
	H_{uav2}	0.843
	H_{uav3}	0.88
	H_{uav4}	0.725
H_{attd}	H_{attd1}	0.738
	H_{attd2}	0.803
	H_{attd3}	0.854
	H_{attd4}	0.728

Cronbach's Alpha is calculated for each variable used in study and shown in Fig. 7. If the value of Cronbach's Alpha is found greater than 0.7 state that variable is reliable and can be used for further analysis.

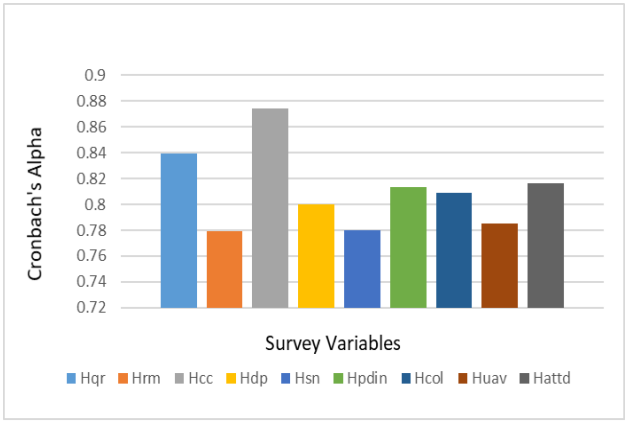


Fig. 7. Cronbach's alpha.

As shown in Table VII the value of all the variable exceed the cut off value indicate that the variables are reliable.

TABLE VII. CRONBACH'S ALPHA

Variable	Cronbach's Alpha
H_{qr}	0.839
H_{rm}	0.779
H_{cc}	0.874
H_{dp}	0.8
H_{sn}	0.78
H_{pdin}	0.813
H_{col}	0.809
H_{uav}	0.785
H_{attd}	0.816

Composite reliability of all used variables is shown in Fig. 8.

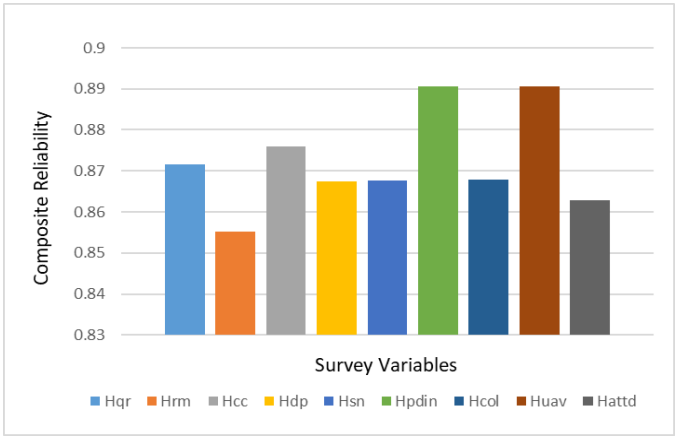


Fig. 8. Composite reliability.

Table VIII shows that the composite reliability also exceeds 0.7 states that items and variables are reliable. The Cronbach's alpha value and composite reliability exceed 0.7 confirm the guideline made by the authors [37].

TABLE VIII. COMPOSITE RELIABILITY

Variable	CR
H_{qr}	0.871597798
H_{rm}	0.855163124
H_{cc}	0.875925089
H_{dp}	0.867347575
H_{sn}	0.867622012
H_{pdin}	0.89058948
H_{col}	0.867908711
H_{uav}	0.89058948
H_{attd}	0.862774646

Average Variance Extracted (AVE) of the study variables are shown in Fig. 9.

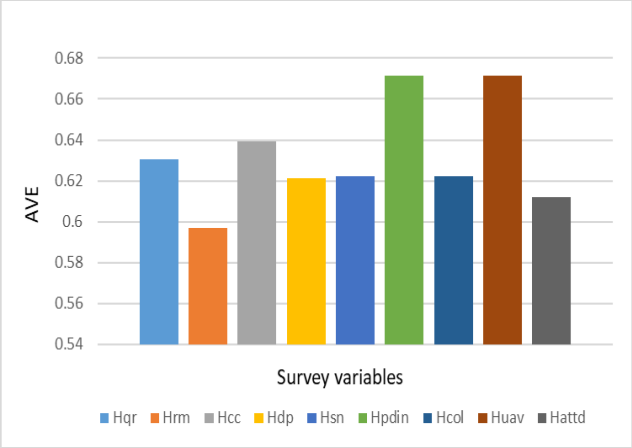


Fig. 9. Average variance extracted (AVE).

TABLE IX. AVERAGE VARIANCE EXTRACTED

Variable	AVE
H_{qr}	0.63060375
H_{rm}	0.59702275
H_{cc}	0.6391865
H_{dp}	0.621094
H_{sn}	0.6219975
H_{pdin}	0.6715895
H_{col}	0.6224615
H_{uav}	0.6715895
H_{attd}	0.61218825

Discriminant table is shown in Table X. The inter construct correlation values are less than the square root value of the Average Variance Extracted values of the respective variables shown in Table IX, thus it validates discriminant validity.

TABLE X. DISCRIMINANT TABLE

	H_{qr}	H_{rm}	H_{cc}	H_{dp}	H_{sn}	H_{pdin}	H_{col}	H_{uav}	H_{attd}
H_{qr}	.794								
H_{rm}	.71	.772							
H_{cc}	.60	.63	.799						
H_{dp}	.71	.68	.59	.788					
H_{sn}	.35	.34	.38	.72	.789				
H_{pdin}	.70	.69	.51	.58	.44	.774			
H_{col}	.55	.58	.62	.34	.56	.34	.788		
H_{uav}	.41	.50	.71	.55	.34	.56	.26	.810	
H_{attd}	.51	.56	.52	.63	.45	.59	.60	.35	.782

Initially, direct relationships tested between H_{qr}, H_{rm}, H_{cc}, H_{dp}, H_{sn} and H_{attd}. H_{qr} ($\beta = 0.43$, $p < 0.001$), H_{rm} ($\beta = 0.37$, $p < 0.001$), H_{cc} ($\beta = 0.65$, $p < 0.001$), H_{dp} ($\beta = 0.36$, $p < 0.001$), H_{sn} ($\beta = 0.68$, $p < 0.001$), H_{attd} ($\beta = 0.78$, $p < 0.001$). It states that the hypothesis H_{qr}, H_{rm}, H_{cc}, H_{dp}, H_{sn} and H_{attd} receives the support. The statistical analysis done for the examination provide the support for all six direct relations. Secondly moderator's H_{pdin}, H_{col} and H_{uav} influence checked on the relationship between H_{qr}, H_{rm}, H_{cc}, H_{dp}, H_{sn} and H_{attd}. The effects of H_{qr} X H_{pdin}, H_{rm} X H_{pdin}, H_{cc} X H_{pdin}, H_{dp} X H_{pdin}, H_{sn} X H_{pdin}, H_{qr} X H_{col}, H_{rm} X H_{col}, H_{cc} X H_{col}, H_{dp} X H_{col}, H_{sn} X H_{col}, H_{qr} X H_{uav}, H_{rm} X H_{uav}, H_{cc} X H_{uav}, H_{dp} X H_{uav}, H_{sn} X H_{uav}. H_{a1}, H_{b1}, H_{c1}, H_{d1} and H_{e1} examined with H_{pdin} as moderator. H_{a2}, H_{b2}, H_{c2}, H_{d2} and H_{e2} examined as H_{col} as a moderator. H_{a3}, H_{b3}, H_{c3}, H_{d3} and H_{e3} examined as H_{uav} as a moderator. The effect of H_{pdi} and H_{col} is Not Significant on above mention relation. The effect of H_{uav} is all were significant and supports the hypothesis.

V. FUTURE PERSPECTIVE

The future perspective of SQRCD (Sustainability, Quality, Resilience, Customer-centricity, and Digitalization) within the setting of building feasible and customer-centric DFIS for the Industry 5.0 era speaks to a major move towards making frameworks that coordinated progressed innovation, human-centric solutions, and sustainability goals. Sustainability isn't fair and natural concern but a foundational guideline for future DFIS. Within the Industry 5.0 era, which emphasizes the integration of people and machines, DFIS must consolidate eco-friendly practices, vitality effectiveness, and economical trade models. Quality affirmation will utilize AI-driven arrangements that automate testing, approval, and compliance to diminish human errors. As cyber dangers advance, securing DFIS with progressed encryption, multi-factor confirmation, and AI-driven extortion avoidance instruments is significant. Combining human imagination with machine accuracy will empower DFIS to offer versatile financial items and administrations custom-made to particular needs.

VI. Conclusion

The study proposed the direct hypotheses to analyze the adoption of an AGI enabled-SQRCD system among respondent belongs to financial service provider institute. Artificial intelligence has gained a very important place in the field of finance sector. It has been seen that that survey variables (i.e. quick responsive, risk management, customer centric, sustainability and data privacy) has positive association with the attitude of adoption of an AGI enabled-SQRCD system. Studies do not always same as actual behaviour so future work can be done to see the actual behaviour to see the adoption of AGI enabled-SQRCD system.

References

[1] Tewari, Niharika. "Artificial Intelligence in Finance and Industry: Opportunities and Challenges." Decision Strategies and Artificial Intelligence Navigating the Business Landscape. https://doi. org/10.59646/edbookc5/009 (2023).

[2] Ahmadi, Sina. "A comprehensive study on integration of big data and AI in financial industry and its effect on present and future opportunities." International Journal of Current Science Research and Review 7, no. 01 (2024): 66-74.

[3] Irfan, Mohammad, Mohammed Elmogy, and Shaker El-Sappagh, eds. The impact of AI innovation on financial sectors in the era of industry 5.0. IGI Global, 2023.

[4] Kotiyal, Bina, Heman Pathak, and Nipur Singh. "Debunking multi-lingual social media posts using deep learning." International Journal of Information Technology 15, no. 5 (2023): 2569-2581.

[5] Li, Yingbo, and Yucong Duan. "The Wisdom of Artificial General Intelligence: Experiments with GPT-4 for DIKWP." arXiv preprint (2023).

[6] Indriasari, Elisa, Ford Lumban Gaol, and Tokuro Matsuo. "Digital banking transformation: Application of artificial intelligence and big data analytics for leveraging customer experience in the Indonesia banking sector." In 2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 863-868. IEEE, 2019.

[7] Prajapati, Mr Nitin. "Influence of AI and machine learning in insurance sector." Bournemouth University Department of Computing Science, MSc. Data Science and AI 1 (2022).

[8] Najem, Rihab, Meryem Fakhouri Amr, Ayoub Bahnasse, and Mohamed Talea. "Artificial intelligence for digital finance, axes and techniques." Procedia Computer Science 203 (2022): 633-638.

[9] Rathi, Soumil. "Approaches to Artificial General Intelligence: An Analysis." arXiv preprint arXiv:2202.03153 (2022).

[10] Obaid, Omar Ibrahim. "From machine learning to artificial general intelligence: A roadmap and implications." Mesopotamian Journal of Big Data 2023 (2023): 81-91.

[11] Al-Baity, Heyam H. "The artificial intelligence revolution in digital finance in Saudi Arabia: a comprehensive review and proposed framework." Sustainability 15, no. 18 (2023): 13725.

[12] Rane, Nitin Liladhar, Saurabh P. Choudhary, and Jayesh Rane. "Artificial Intelligence-driven corporate finance: enhancing efficiency and decision-making through machine learning, natural language processing, and robotic process automation in corporate governance and sustainability." Studies in Economics and Business Relations 5, no. 2 (2024): 1-22.

[13] Prasanth, Anupama, John Vadakkan Densy, Priyanka Surendran, and Thomas Bindhya. "Role of artificial intelligence and business decision making." International Journal of Advanced Computer Science and Applications 14, no. 6 (2023).

[14] Latif, Ehsan, Gengchen Mai, Matthew Nyaaba, Xuansheng Wu, Ninghao Liu, Guoyu Lu, Sheng Li, Tianming Liu, and Xiaoming Zhai. "AGI: Artificial general intelligence for education." arXiv preprint arXiv:2304.12479 (2023).

[15] Banitaan, Shadi, Ghaith Al-refai, Sattam Almatarneh, and Hebah Alquran. "A review on artificial intelligence in the context of industry 4.0." International Journal of Advanced Computer Science and Applications 14, no. 2 (2023).

[16] Rane, J., S. K. Mallick, O. Kaya, and N. L. Rane. "Artificial general intelligence in industry 4.0, 5.0, and society 5.0: Applications, opportunities, challenges, and future direction." Future Research Opportunities for Artificial Intelligence in Industry 4.0 and 5 (2024): 2.

[17] Bu, Y., Du, X., Wang, Y., Liu, S., Tang, M., & Li, H. (2024). Digital inclusive finance: A lever for SME financing?. International Review of Financial Analysis, 93, 103115.

[18] Giudici, P., & Raffinetti, E. (2023). SAFE Artificial Intelligence in finance. Finance Research Letters, 56, 104088.

[19] Lee, J. (2020). Access to finance for artificial intelligence regulation in the financial services industry. European Business Organization Law Review, 21(4), 731-757.

[20] Rame, Rame, Purwanto Purwanto, and Sudarno Sudarno. "Industry 5.0 and sustainability: An overview of emerging trends and challenges for a green future." Innovation and Green Development 3, no. 4 (2024): 100173.

[21] Zheng, Xiao-lin, Meng-ying Zhu, Qi-bing Li, Chao-chao Chen, and Yan-chao Tan. "FinBrain: when finance meets AI 2.0." Frontiers of Information Technology & Electronic Engineering 20, no. 7 (2019): 914-924.

[22] Nicholls, Jack, Aditya Kuppa, and Nhien-An Le-Khac. "Financial cybercrime: A comprehensive survey of deep learning approaches to tackle the evolving financial crime landscape." Ieee Access 9 (2021): 163965-163986.

[23] Nahavandi, Saeid. "Industry 5.0—A human-centric solution." Sustainability 11, no. 16 (2019): 4371.

[24] Nguyen, Thanh Thi, Hammad Tahir, Mohamed Abdelrazek, and Ali Babar. "Deep learning methods for credit card fraud detection." arXiv preprint arXiv:2012.03754 (2020).

[25] AL-Dosari, Khalifa, Noora Fetais, and Murat Kucukvar. "Artificial intelligence and cyber defense system for banking industry: A qualitative study of AI applications and challenges." Cybernetics and systems 55, no. 2 (2024): 302-330.

[26] Alenzi, Hala Z., and Nojood O. Aljehane. "Fraud detection in credit cards using logistic regression." International Journal of Advanced Computer Science and Applications 11, no. 12 (2020).

[27] Snaider, Javier, Ryan McCall, and Stan Franklin. "The LIDA framework as a general tool for AGI." In Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings 4, pp. 133-142. Springer Berlin Heidelberg, 2011.

[28] Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee et al. "Sparks of artificial general intelligence: Early experiments with gpt-4." arXiv preprint arXiv:2303.12712 (2023).

[29] Dou, Fei, Jin Ye, Geng Yuan, Qin Lu, Wei Niu, Haijian Sun, Le Guan et al. "Towards artificial general intelligence (agi) in the internet of things (iot): Opportunities and challenges." arXiv preprint arXiv:2309.07438 (2023).

[30] Mamoona, Humayun. "Industrial Revolution 5.0 and the Role of Cutting Edge Technologies." International Journal of Advanced Computer Science and Applications (IJACSA) 12 (2021).

[31] Thatikonda, Ramya, Jainath Ponnala, Dileep Kumar Yendluri, M. Kempanna, Reshmi Tatikonda, and A. Bhuvanesh. "The Impact of Blockchain and AI in the Finance Industry." In 2023 International Conference on Computational Intelligence, Networks and Security (ICCINS), pp. 1-6. IEEE, 2023.

[32] Dash, Bibhu, Pawankumar Sharma, and Swati Swayamsiddha. "Organizational digital transformations and the importance of assessing theoretical frameworks such as TAM, TTF, and UTAUT: A review." International Journal of Advanced Computer Science and Applications 14, no. 2 (2023).

[33] Hofstede, Geert. "Culture and organizations." International studies of management & organization 10, no. 4 (1980): 15-41.

[34] Bokhari, Syed Asad Abbas, and Seunghwan Myeong. "Ai applications in smart city employing technology adoption model: Hofstede's cultural perspective." In 2023 2nd International Conference for Innovation in Technology (INOCON), pp. 1-6. IEEE, 2023.

[35] Jan, Jeffy, Khaled A. Alshare, and Peggy L. Lane. "Hofstede's cultural dimensions in technology acceptance models: a meta-analysis." Universal Access in the Information Society 23, no. 2 (2024): 717-741.

[36] Rawat, Ruchira, Sachin Sharma, and Himanshu Rai Goyal. "Hofstede's Cultural Dimension Driven Artificial Narrow Intelligence iDFIS for Industry 5.0 Empowered Digital Society." Recent Patents on Engineering (2024).

[37] Fornell, Claes, and David F. Larcker. "Evaluating structural equation models with unobservable variables and measurement error." Journal of marketing research 18, no. 1 (1981): 39-50.

Efficient Personalized Federated Learning Method with Adaptive Differential Privacy and Similarity Model Aggregation

Shiqi Mao[1], Fangfang Shan[2]*, Shuaifeng Li[3], Yanlong Lu[4], Xiaojia Wu[5]
School of Computer Science, Zhongyuan University of Technology, Zhengzhou 450007, Henan, China[1, 2, 3, 4, 5]
Henan Key Laboratory of Cyberspace Situation Awareness, Zhengzhou 450001, Henan, China[2]

***Abstract*—In recent years, personalized federated learning (PFL) has garnered significant attention due to its potential for safeguarding data privacy while addressing data heterogeneity across clients. However, existing PFL approaches remain vulnerable to privacy breaches, particularly under adversarial inference and client-side data reconstruction attacks. To address these concerns, we propose DP-FedSim, a novel PFL framework incorporating adaptive differential privacy mechanisms. First, to mitigate the limitations posed by fixed-layer personalization strategies, we evaluate parameter significance using the Fisher information matrix. By selectively retaining parameters with higher Fisher values, DP-FedSim reduces the noise impact, enabling more efficient dynamic personalization. Second, we introduce a layered adaptive gradient clipping method. By leveraging the mean and standard deviation of the gradients within each layer, this method allows DP-FedSim to automatically adjust clipping thresholds in response to real-time privacy demands and model states, enhancing the adaptability to various model structures. This ensures a more accurate balance between privacy preservation and model performance. Furthermore, we present a model similarity-based aggregation method utilizing cosine similarity. This technique dynamically adjusts each client's contribution to the global model update, prioritizing clients with models more similar to the global model. This improves the global model's performance and generalization by allowing DP-FedSim to better handle a variety of data distributions and client model attributes. Experimental results on multiple SVHN cifar-10 datasets show that DP-FedSim outperforms the state-of-the-art PFL algorithm by an average of 5% when data heterogeneity is at its strongest. The efficiency of the suggested modules is validated by ablation tests, and the visualization results shed light on the reasoning behind important hyperparameter settings.**

***Keywords*—Federated learning; differential privacy; gradient clipping; model aggregation**

I. INTRODUCTION

A distributed machine learning approach called federated learning (FL) [1] allows several separate devices to work together to train a single model without explicitly sharing local data, protecting privacy and avoiding data leaks. Only model updates are sent to a central server within the FL framework; each participant trains a model separately using their own local data. This decentralized method lowers network communication cost while also protecting data privacy. The performance of traditional federated learning models can be severely harmed by participant data that frequently exhibits non-independent and identically distributed (non-IID) features in real-world applications [2][3][4][5], which can significantly degrade the performance of conventional federated learning models [6][7][8][9]. To address this challenge, Personalized Federated Learning (PFL) algorithms [10][11][12][13] have been developed, incorporating personalization techniques to better accommodate the unique data distributions of individual participants. Despite the notable advances in PFL, privacy concerns remain unresolved. Specifically, model updates exchanged between clients and the central server remain vulnerable to inference and reconstruction attacks that can compromise the privacy of client data [14][15][16][17]. Therefore, integrating robust privacy-preserving techniques, particularly user-level differential privacy [18][19][20][21], is essential to ensure the security and reliability of PFL systems while protecting the privacy of sensitive data.

Despite significant progress, PFL still faces numerous challenges related to the implementation of differential privacy mechanisms. The current PFL techniques [10][22][23] frequently make strong assumptions about parameter partitioning, where clients share a set percentage of the model parameters while the rest are customized. This static approach lacks flexibility in parameter division and fails to fully account for the diversity in client data, which can hinder the performance of personalized models.

Within the context of differential privacy, gradient clipping is a key mechanism used to limit the size of gradients, thereby preventing sensitive information leakage and mitigating the issue of gradient explosion. Gradient clipping also reduces sensitivity, allowing for more efficient privacy budget usage without substantially degrading model performance [25]. However, traditional methods typically employ a fixed clipping threshold c, which can result in either over-clipping or under-clipping under different conditions, potentially impairing the performance of the model.

This paper is supported by the National Natural Science Foundation of China (No. 62302540), with author F.F.S. For more information, please visit their website at https://www.nsfc.gov.cn/ . Additionally, it is also funded by the Open Foundation of Henan Key Laboratory of Cyberspace Situation Awareness (No. HNTS2022020), where F.F.S is an author. Further details can be found at http://xt.hnkjt.gov.cn/data/pingtai/ . The research is also supported by the Natural Science Foundation ofHenan Province Youth Science Fund Project (No. 232300420422), and for more information, you can visit https://kjt.henan.gov.cn/2022/09-02/2599082.html.

In most federated learning (FL) frameworks, after a client's local training round is completed, the client sends the model updates, typically in the form of gradients, to a central server responsible for aggregation. The standard aggregation method often relies on a simple averaging of the received gradients. However, in practice, not all clients participate equally in the model training process [26]. Some clients may contribute less due to factors such as limited data, weak computational resources, or unstable network connections. This imbalance can result in the global model becoming overly dependent on clients with larger data volumes or stronger computational capabilities [27], which can hinder the model's ability to capture broader data trends, thus compromising the generalization and convergence of the global model.

However, existing studies still have obvious shortcomings in addressing the above problems. On the one hand, although some studies try to protect privacy through differential privacy techniques, they fail to adequately address the problems of inflexible parameter partitioning and overly fixed gradient trimming strategies in personalized federated learning. On the other hand, the improvement of the aggregation method also fails to effectively take into account the heterogeneity of client data and the difference in model quality, resulting in limited global model performance. Therefore, how to achieve more flexible parameter personalization, more accurate gradient tailoring, and more effective aggregation strategies under the premise of privacy protection has become a key problem to be solved in the field of personalized federated learning.

To address these challenges, we draw inspiration from the work "FedFisher: Leveraging Fisher Information for One-Shot Federated Learning" [28], which employs the Fisher information matrix to facilitate dynamic personalization of model parameters. The square of the first-order derivative of the log-likelihood function is calculated in the core mechanism to assess the contribution of parameters to the curvature of the loss function. In essence, this process captures the information content carried by each parameter. Leveraging this principle, we assess the significance of each client's model parameters through the Fisher information matrix before training begins. By retaining parameters that carry the most information, we mitigate the adverse effects of noise addition in differential privacy settings and avoid potential optimization issues stemming from inappropriate global model parameters.

Building on this, we propose an adaptive gradient pruning strategy, which introduces a hierarchical adaptive gradient clipping method. This approach automatically adjusts the clipping threshold according to the current privacy preservation requirements and the real-time state of the model. In contrast to traditional differential privacy methods that use fixed clipping bounds, this adaptive approach offers greater flexibility and can better accommodate diverse network structures and training environments. It also provides a more accurate response to privacy leakage risks. Furthermore, by allowing the use of larger learning rates, adaptive gradient clipping accelerates convergence and enhances model training performance.

Additionally, we introduce a model similarity-based aggregation method, which utilizes cosine similarity to dynamically adjust each client's contribution to the global model based on the similarity of its parameters to those of the global model. This technique is designed to better align with the data distributions and model quality of different clients, rather than simply assigning equal weight to all updates. By prioritizing updates from clients with models more similar to the global model, this method improves the overall performance and generalization of the global model, as these more similar models are likely to better capture the broader patterns and trends in the data. The following are this paper's main contributions:

- We introduce DP-FedSim, a personalized federated learning framework with adaptive differential privacy. DP-FedSim effectively integrates personalized learning with adaptive gradient tailoring, making it ideal for situations involving a high degree of client data variety and diversity.
- To address the limitations of fixed-layer approaches in traditional personalized federated learning, we leverage the Fisher information matrix to enable a dynamic approach to customization. The Fisher information matrix quantifies the importance of parameters, and under the same additive noise, parameters with higher Fisher values are more sensitive to noise, resulting in greater performance degradation. In order to lessen the effect of noise and improve model performance, we maintain parameters with higher Fisher values.
- We propose a novel adaptive gradient clipping method based on the mean and standard deviation of gradients within each layer. This method accelerates the training process and improves model performance while ensuring privacy protection. In addition, we introduce a model similarity-based aggregation strategy that effectively combines model updates from diverse clients. This method addresses the challenge of heterogeneous client data, where updates may differ significantly, by dynamically adjusting the contributions of client models based on their similarity to the global model.

This paper's remaining sections are organized as follows: We evaluate relevant work in Section II. Section III outlines the foundational concepts relevant to this study. Section IV describes the proposed methods in detail, including the implementation of the three key approaches. Section V presents a performance comparison of DP-FedSim with state-of-the-art methods, and Section VI concludes the paper.

II. Related Work

A. Personalized Federal Learning

A machine learning paradigm called Personalized Federated Learning (PFL) allows a central server to plan model training for dispersed clients without having direct access to their data. The primary objective of PFL is to address data heterogeneity by learning customized models for each client. Mainstream PFL approaches include LG-FedAvg [23], FedBABU [30], PPSGD [29], and FedBN [31]. Both FedBABU [30] and LG-FedAvg [23] adopt fixed local

parameters to exploit local data performance for personalization. However, their static partitioning approaches limit the flexibility required for handling diverse data distributions. The FedPer algorithm [10] introduces personalization layers, which are appended to the base model. During training, the base model's parameters are globally aggregated, while the parameters of the personalization layers remain local and are not aggregated. Similarly, FedBN [31] incorporates one or more batch normalization (BN) layers, which remain fixed during training and do not participate in global aggregation. By keeping certain layers locally, these methods improve customization while better accommodating local data peculiarities, however they might not be as flexible in terms of model adaption.

Additionally, algorithms such as GPFL [32], pFedMe [24], and FedAMP [33] aim to learn both global and personalized feature representations. These methods strike a balance between global model consistency and local personalization, allowing models to capture common features across clients while preserving unique, client-specific information. Despite their effectiveness, these approaches are still constrained by the inflexibility of their personalization mechanisms and may see a decrease in performance as a result of using noisy global parameters directly under the differential privacy (DP) technique.

B. Differential Privacy

Differential privacy (DP) is a crucial technique for protecting data privacy, designed to minimize the risk of identifying individual records when statistical information is shared. In the context of federated learning, user-level differential privacy has been widely adopted [18][19][20][34]. This technique quantifies privacy preservation through two key parameters, (ε, δ), where smaller values of ε and δgenerally imply stronger privacy guarantees but add extra noise, which might impair model performance, to the federated learning process. In federated learning, user-level DP is usually accomplished in two steps: first, local updates are clipped and noise is added before being sent to the server. Clipping reduces the effect of local updates, further improving privacy, and the noise is adjusted based on the sensitivity of the function being assessed. Although these steps greatly improve privacy, the required noise addition and gradient clipping may cause performance issues and delayed convergence.

Recent studies have explored ways to mitigate the negative impact of noise and clipping on performance. Sparsification and uniform regularization approaches, for example, are used by LUS and BLUR [19] to mitigate the impacts of noise and accelerate model convergence. The DP-FedSAM [20] algorithm enhances robustness to noise by employing the Sharpness-Aware Minimization (SAM) optimizer, which identifies more reliable places of convergence. Furthermore, PPSGD [29] uses customisation to enhance performance without compromising privacy. In spite of the progress in these methods, research on gradient pruning and clipping within personalized federated learning remains limited. In this paper, we aim to optimize personalized federated learning algorithms through adaptive gradient tailoring, contributing to the ongoing development of personalized federated learning approaches.

C. Federated Learning Aggregation Algorithm

Model aggregation is a core component of federated learning, where the local models from clients are aggregated in each communication round to generate an updated global model. There are two main types of aggregation: parameter-based aggregation and output-based aggregation, with the distinction based on the aggregation target. One of the first and most popular federated learning algorithms is FedAvg [1], which aggregates models by average parameters from all clients, weighted by the size of each client's dataset. The FedProx algorithm [35] modifies the aggregation process by introducing a proximal term in the objective function to control the impact of local models and ensure convergence. Meanwhile, FedNova [36] improves upon FedAvg by normalizing and scaling local updates based on each client's local iteration count, which helps achieve fairer aggregation. A data agnostic distributional fusion model, which depicts the client's heterogeneous data distribution as a global collection comprised of multiple virtual fusion components with varying parameters and weights, is used by the FedFusion algorithm [37] to characterize the global data distribution.

Although these methods contribute to the development of model aggregation in federated learning, challenges remain. For instance, the FedNova algorithm introduces additional computational complexity due to the need to track and adjust local iteration counts, which increases the computational burden on both the clients and the server, ultimately affecting model convergence speed.

III. Preliminary

A. Personalized Federal Learning

In personalized federated learning (PFL), the primary objective is to train a model that can adapt to the unique data distribution of each client while preserving a degree of global consistency across all clients. This is typically achieved by decomposing the model parameters into two distinct components: one set of globally shared parameters and another set of client-specific personalized parameters, as illustrated in Fig. 1.

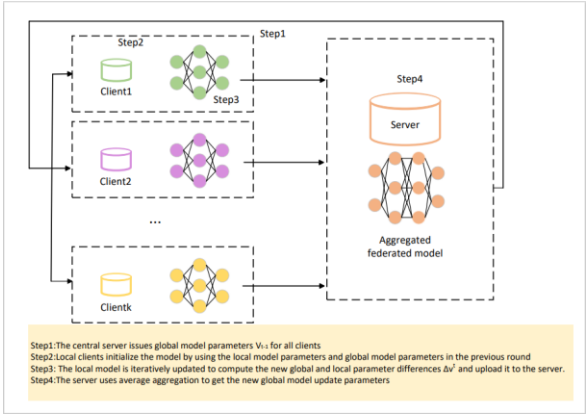


Fig. 1. The key processes of traditional personalized federated learning are depicted. First, a central server publishes global model parameters for use by clients. The model is then repeatedly updated by the client using local data to calculate parameter variances, which are then uploaded to the server. Finally, the server creates global model updates by integrating the variances using an average aggregation approach.

Suppose we have k clients, each possessing a unique dataset, denoted as $D_k = \{(x_i, y_i)\}_{i=1}^{N_k}$, where N_k is the size of the dataset for client k, and i indexes the data samples. In common personalization methods, the model parameter vector w_i is typically decomposed into two components: a local part and a global part, represented as w = (u, v). The objective of personalized federated learning is to update the model parameters according to a specific optimization process, as outlined in Eq. (1).

$$\min_{\nu, u_{1:k}} \left\{ f(\nu, u_{1:k}) := \frac{1}{m} \sum_{i}^{k} f_i(\nu, u_i) \right\} \tag{1}$$

Let $u_{1:k}$ denote $(u_1, \ldots, u_k)$, and let $f_i(v, u_i)$ represent the average loss of parameters v and u_i over the entire dataset D_i for client i, where $i = 1, \ldots, N_k$. The personalized differential privacy mechanism involves two iterative steps:

Local Iteration: During the local iteration, each client i receives the global model parameters v^{t-1} from the server while retaining the local parameters $u_{1\,i}^{t-}$ from the previous round. The model is initialized as $w_{1\,i}^{t-} = (v^{t-1}, u_{1\,i}^{t-})$. The client then performs local updates, iteratively optimizing the parameters to obtain the updated model $w_i^t = (v_i^t, u_i^t)$. The global update $\triangle v^t$ is computed as the difference between v_i^t and v^{t-1}.

Global Update: In the global update phase, all clients transmit their local updates $\triangle v^t$ to the server. The server then aggregates these updates by averaging them across all clients, updating the global parameter as follows:

$$v^t \leftarrow v^{t-1} + \frac{1}{k} \sum_{i=1}^{m^t} \Delta v_i^t \tag{2}$$

The new global parameter v^t is then sent back to all clients to initiate the next round of updates.

B. User-Level Differential Privacy

In personalized federated learning, differential privacy offers a robust protection mechanism that effectively addresses the issue of privacy leakage. As a comprehensive privacy protection framework, differential privacy aims to facilitate the analysis of overall dataset properties while safeguarding individual information. This is achieved at the cost of a certain degree of data accuracy, thereby ensuring stringent privacy protection for user data. The ultimate goal is to prevent adversaries from determining whether a specific individual is represented in the dataset.The concept of differential privacy [38] is defined mathematically to delineate this probability gap, as articulated in Definition 1:

Definition 1 (ε, δ). Differential Privacy A randomization mechanism M satisfies (ε, δ)-Differential Privacy (ε > 0, δ > 0) if and only if, for any adjacent input datasets S and S', and for any possible set of output values R, the following holds:

$$Pr[M(S) \in R], e^{\varepsilon} \cdot Pr[M(S') \in R] + \delta \tag{3}$$

Here, δ denotes the probability of a failure in privacy protection. A randomized algorithm M satisfies (ε, δ)-DP if, for any pair of neighboring datasets D and D' differing by a single record, and for any output subset S in the range of M.

Definition 2 L2 Sensitivity Given a function M and two neighboring datasets D and D', the L2 sensitivity is defined as follows:

$$\Delta f = \max \| M(D) - M(D') \|_2 \tag{4}$$

User-level differential privacy is a specific classification within the broader framework of differential privacy. Noise must be added to the model updates that are locally calculated by each user in order to apply user-level differential privacy to customized federated learning. This approach ensures compliance with user-level differential privacy requirements. Specifically, users must incorporate noise into the gradient or model parameter updates derived from their local datasets after training.

Based on this theoretical foundation, user-level differential privacy is effectively achieved through gradient cropping and noise addition. Gradient cropping is primarily employed to regulate the model's sensitivity to individual data points. By mitigating the influence of outlier samples during a given training round, it helps protect data privacy.

However, much of the existing research focuses on fixed gradient cropping methods. If the cropping threshold is set too high, most gradients may fail to exceed this threshold, rendering the cropping process ineffective. Conversely, setting the threshold too low may excessively constrain gradient updates, hindering the model's ability to glean valuable information from the data. This can diminish the training efficiency and, ultimately, the predictive performance of the model.

Therefore, this paper investigates adaptive gradient cropping within the context of personalized federated learning. A detailed exploration of this topic is presented in Section IV.

IV. Methodology

In this study, we offer a differential privacy federated learning system that combines model similarity-based aggregation, adaptive gradient cropping, and customization based on Fisher information matrices. The proposed approach consists of three key components: Fisher personalized federated learning, adaptive gradient cropping for differential privacy, and aggregation based on model similarity.

When the client gets the global model from the server, the procedure starts. Each local client then computes the Fisher information vector F_i using the Fisher information matrix. Subsequently, the client generates computational binary masks M_{1i} and M_{2i} based on the F_i vector and a set of parameters λ. These parameters are crucial in determining which model parameters should be deemed informative and retained throughout the personalization process.

Once the binary masks M_{1i} and M_{2i} are established, they are utilized to update the local model parameters $w_{1\,i}^{t-}$. The updated model parameters w_i^t are then obtained through further training using these masks.

Next, the cropping thresholds T_s are computed by leveraging the mean and standard deviation of the gradients corresponding to each layer. The gradient parameter $\| g_i \|$ is subsequently calculated, followed by the computation of the cropping factors c_s based on the defined cropping thresholds T_s. These cropping factors are employed to control the scaling of the gradient, yielding the adjusted gradient $g_i^{'}$. The complete gradient is computed by summing the cropped gradients across all layers l.

Finally, model similarity-based aggregation is achieved by calculating the cosine similarity between the global model parameters and the parameters of the i-th client model. Specifically, we first compute the similarity S_i for each client model, followed by summing and weighting all computed similarities. This cumulative similarity is then utilized to update the global model accordingly.

The detailed algorithmic procedure is illustrated in Algorithm 1 and Fig. 2.

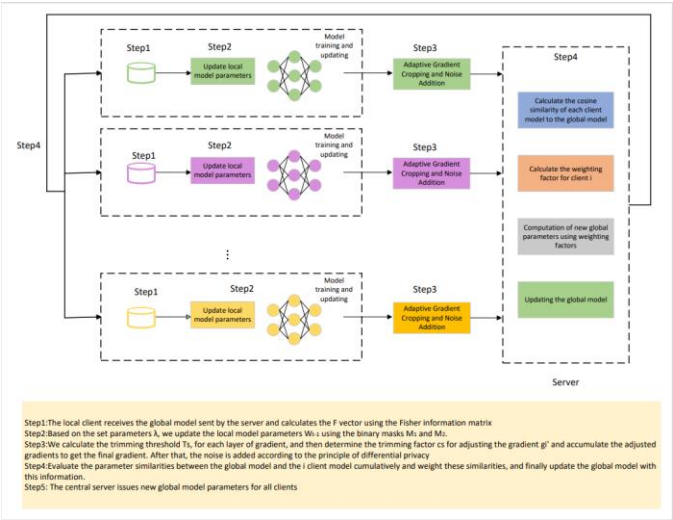


Fig. 2. An overview of DP-FedSim. The client first receives the global model and computes the F vector, then updates the local model parameters. Next, the gradient is adjusted by set thresholds and factors, and noise is added to maintain differential privacy. Finally, the server synthesizes the parameter similarities across clients and updates the global model.

A. Based on Fisher Personalized Federal Learning

Motivation: The motivation for personalization in federated learning is underscored by the use of the Fisher information matrix. The Fisher information matrix serves as an effective tool for quantifying the importance of model parameters; specifically, a larger Fisher value indicates a greater significance of the parameter in the model's predictive performance. This observation leads to the conclusion that parameters with elevated Fisher values contribute to a more substantial degradation of model performance when subjected to the same additive noise.

In conventional personalization approaches, after receiving the global model from the server, clients typically designate specific fixed layers within the network as personalization layers. However, this method is inherently limited, as it fails to account for the differential impact of noise on various parameters. To address this limitation, we propose the introduction of Fisher values as a metric for assessing the importance of model parameter information across each layer.

Fisher Personalization: To alleviate the current inflexibility issues associated with personalized federated learning, we implement a dynamic personalization strategy based on the Fisher information matrix. This approach enhances the model's adaptability to the non-independent and identically distributed (non-IID) data characteristic of individual clients. The procedure is as follows: each client i receives the distributed global model w^{t-1} from the server and subsequently computes the Fisher value F_i based on its local private dataset D_i. The retention of the previous round's local parameters is denoted as $w_{1\,i}^{t-} = (v^{t-1}, u_{1\,i}^{t-})$. In $w_{1\,i}^{t-}$, the diagonal approximation of the true Fisher value for each parameter indexed by j is calculated as:

$$F(w_{ij}) = \left(\frac{\partial \log L(w_i, D_i)}{\partial w_{ij}} \right)^2 \tag{5}$$

This formula illustrates how sensitive the model's predictions are to changes in parameter w_j and serves as a foundational element in enhancing the personalization process.

where $L(w_i, D_i)$ denotes the log-likelihood function of w_i. Then by normalizing each parameter j in layer s layer by layer, the value of fisher's F_s is achieved as

$$\hat{F}_{s,j} = \frac{F_{s,j}}{\sum_j F_{s,j}} \tag{6}$$

After we generate the layer-by-layer Fisher values f_s by the above operation, we generate two binary masks M_1 and M_2 for dynamic selection of parameters. In the event that a parameter's Fisher value is larger than or equal to λ, it is set to 1, otherwise the value is set to 0. For each parameter j in layer s, the mask is defined as follows:

$$M_1[j] = \text{sgn}(\hat{F}_{ij} - \lambda) \quad \text{and} \quad M_2[j] = 1 - M_1[j] \tag{7}$$

To choose the right parameters for customization, we next carry out an elemental multiplication between these masks and parameters. In other words that is, the parameters that had a greater Fisher value locally in the previous round are kept through the masking operation. The remaining parameters are replaced with global parameters.

$$w_i^t = M_1 \odot w_i^{t-1} + M_2 \odot w^{t-1} \tag{8}$$

where M_1 and $w_{1\,i}^{t-}$, M_2 and w^{t-1} serve as the Hadamard product, i.e., the point-by-point product between the elements. The global parameter supplied by the receiving server is w^{t-1}, whereas the local personalization parameter kept from the previous round is $w_{1\,i}^{t-}$. Updating with this method effectively retains the more informative parameters as personalized parameters, and the less informative parameters are updated by the corresponding informative global parameters.

B. Adaptive Gradient Tailoring Differential Privacy

Motivation: In the realm of differential privacy, traditional fixed gradient cropping methods exhibit inherent limitations, particularly when applied to diverse datasets and model architectures. These methods often lack the flexibility to adapt to varying circumstances, leading to issues of over-cropping or under-cropping. Such discrepancies can adversely affect both the training efficacy and the final performance of the model.

The primary motivation behind the adaptive gradient cropping (AGC) technique is its capacity to dynamically adjust cropping thresholds in response to the statistical characteristics of the weights at each layer. This dynamic adjustment mechanism enhances the stability of model training and offers the flexibility to accommodate different network architectures and training configurations, thereby allowing for more precise control over the risk of privacy leakage. A further significant advantage of the AGC technique is its facilitation of larger learning rates, which is critical for accelerating model convergence. This capability can substantially enhance the efficiency and performance of model training, resulting in a more expedient training process and improved model outcomes while maintaining robust privacy protection.

Algorithm 1: Heading

Initialize Global epochs E_g local epochs E_l participants number in t he t^{th} epoch m^t, private data of the t^{th} client $D_i = (X_i, Y_i)$, global model parameters w and client local parameters w_i, hyper-parameters λ, learning rate η,
Local Update:
For $i = 1,2,\ldots,m^t$ Server
Receive w^{t-1} from Serve

M_{1i} and $M_{2i} \leftarrow (\hat{F}_i, \lambda)$

For $e = 1, 2, ..., E_l$ do

$w_i^{t-1} = (v^{t-1}, u_i^{t-1}) \leftarrow M_{1i} \square\ w_i^{t-1} + M_{2i} \square\ w^{t-1}$

End

$$\Delta w_i^t = \sum_{l=1}^{L} \left(\min\left(1, \frac{T_s}{\square g_s \square + \grave{o}}\right) \cdot g_s \right) + \mathrm{N}\,(0, \sigma^2_{\text{noise}})$$

$$w_i^t \leftarrow w_i^{t-1} - \eta \cdot \nabla_w L(w_i^{t-1}; D)$$

End
Server Execute:
For t = 1,2,…, E_l do

For each client model weight w_i do

$$S_i \leftarrow \frac{< w_i, w_g >}{\| w_i \| \ \| w_g \|}$$

$$S_{sum} \leftarrow S_{sum} + S_i$$

End

For $i = 1, 2, ..., m^t$ do

$$w_i \leftarrow \frac{S_i}{S_{sum}}$$

$$\Delta w_i^t \leftarrow \Delta w_i^t + w_i \Delta w_i$$

End

$$w^t \leftarrow w^{t-1} + \Delta w_i^t$$

For $i = 1, 2, ..., m^t$ do

Send w^t to client i^{th}

End

End

In the context of Federated Learning (FL), the heterogeneity of data across different clients often results in substantial discrepancies in the model updates submitted by each client. Such variations may lead to excessively large or small gradient updates for some clients, consequently impacting the training stability of the global model and increasing the risk of privacy breaches. This paper presents the AGC approach, which dynamically modifies the cropping threshold according on the statistical characteristics of the gradient at each layer in order to address these issues. This method effectively addresses the complications arising from heterogeneous data, enhancing both training stability and privacy assurance.

1) Calculation of gradient trimming threshold: The fundamental principle of adaptive gradient trimming lies in performing gradient trimming on each layer's parameters of the global model while dynamically adjusting the trimming threshold to accommodate gradient variations across different data distributions. Specifically, for each layer parameter θ_i of the model, we first compute the mean μ_s and standard deviation σ_s of the corresponding gradient g_s in that layer. These statistical measures provide insight into the concentration and distribution characteristics of the gradients within that layer. The cropping threshold T_s is then calculated as follows:

$$T_s = b \times \left(1 + \frac{\sigma_s}{|\ \mu_s| + \grave{o}}\right) \tag{9}$$

where b represents a predetermined base cropping threshold, which governs the overall intensity of cropping, and ϵ is a small positive constant introduced to prevent division by zero errors. In this formulation, a larger standard deviation of the gradient for a given layer suggests a more dispersed gradient distribution, indicating the potential presence of outliers or noise. In such cases, the cropping threshold will be correspondingly elevated to mitigate excessive cropping. Conversely, when the standard deviation is small, the cropping threshold will be relatively low, thereby enforcing a stricter control over the size of the gradient.

2) Gradient trimming process: Following the determination of the cropping threshold for each layer, we proceed to trim the gradient g_s for each respective layer. Specifically, we first calculate the norm $\| g_i \|$ of the gradient, after which the cropping factor c_s is computed based on the previously established cropping threshold T_s.

$$c_s = \min\left(1, \frac{T_s}{\square g_s \square + \partial}\right) \tag{10}$$

The cropping factor cs governs the scaling of the gradient, ensuring that the cropped gradient does not exceed the predefined threshold. Ultimately, the cropped gradient $g_i^{'}$ for this layer can be expressed as follows:

$$g_s^{'} = g_s \times c_s \tag{11}$$

The complete client-side gradient is expressed as the aggregation of the cropped gradients from all layers. Assuming there are L layers, each with a corresponding cropped gradient $g_i^{'}$, the complete gradient for the client can be formulated as follows:

$$G_{\mathrm{i}} = \sum_{l=1}^{L} g_s^{'} \tag{12}$$

where G' represents the complete client-side gradient, and $g_{\mathrm{l}}^{'}$ denotes the cropped gradient for layer l.

This method preserves much of the information found in normal gradients while successfully reducing the negative impacts of anomalous gradients on the global model. As a result, it improves model training stability and reduces information loss.

Gradient update and noise addition after cropping: To further enhance privacy protection, noise is introduced to the cropped gradient $g_i^{'}$. This addition adheres to the principles of differential privacy, with the standard deviation σnoise calculated based on the base cropping threshold b and the noise multiplier noise_multiplier:

$$\sigma_{\mathrm{noise}} = \sqrt{\frac{\mathrm{b}^2 \times noise_{\mathrm{m}} ultiplier^2}{N}} \tag{13}$$

where N denotes the number of clients participating in federated learning. The noise addition process involves incorporating Gaussian noise into the gradient update at each layer, represented by the following formula:

$$g_i^{\mathrm{update}} = G_i + \mathrm{N}\ (0, \sigma_{\mathrm{noise}}^2) \tag{14}$$

This process ensures that the model's privacy is further reinforced while maintaining the integrity of the gradient through cropping. By appropriately calibrating the noise intensity, we can maximize the privacy of user data without compromising the accuracy of the model.

C. Aggregation Based on Model Similarity

Motivation: In the realm of federated learning, the selection of an appropriate aggregation strategy is pivotal to the ultimate performance of the model. The classical Federated Averaging (FedAvg) algorithm simply averages the model weights of all participating clients, with the averaging weighted by the volume of data each client holds. However, in personalized federated learning scenarios, the model updates from clients may exhibit substantial variability due to the inherent heterogeneity of their respective datasets. Consequently, straightforward weighted averaging may result in diminished global model performance or inadequate personalization.

To alleviate this problem, we propose a model similarity-based aggregation method, the core of which is to dynamically adjust the client's contribution weight in federated learning by measuring the consistency of update directions between the client's local model and the global model. The method adopts cosine similarity as the similarity metric: firstly, the parameter update vectors of the client model are cosine similar to the global update direction. The advantage of cosine similarity is that it focuses on the vector direction rather than the magnitude, which can effectively capture the synergy of the model updates, e.g., clients with high similarity in the update direction are consistent with the global trend, which can be given a higher aggregation weight to inhibit the bias caused by non independent identically distributed data leads to biased updates. Compared with Euclidean distance or Pearson correlation coefficient, cosine similarity is more robust to amplitude changes in high-dimensional sparse model parameter space, and the computational efficiency is more suitable for distributed scenarios. Through this mechanism, local model updates compatible with the global objective can be filtered out to improve the convergence speed, while retaining the client's personalized features, ultimately achieving a balanced optimization of global model performance and personalization.

3) Calculation of model similarity: In model similarity-based aggregation methods, it is essential to first quantify the similarity between each client model and the global model. In this paper, we employ cosine similarity as a measure of the degree of similarity between the client model and the global model. Cosine similarity is a widely utilized metric that computes and normalizes the inner product of two vectors, thereby deriving the angular similarity between them. In this approach, we treat the parameter vectors of the global model and each client model as high-dimensional vectors. We then compute the cosine similarity between these vectors layer by layer, ultimately taking the average similarity across all layers.

Specifically, let w_g denote the parameters of the global model and w_i represent the parameters of the i-th client model. The similarity between these two can be defined as

$$S(\mathbf{w}_i, \mathbf{w}_g) = \frac{< \mathbf{w}_i, \mathbf{w}_g >}{\square \mathbf{w}_i \square\square \mathbf{w}_g \square} \tag{15}$$

where $<w_i, w_g>$ denotes the inner product of the parameter vectors w_i and w_g, while $\|w_i\|$ and $\|w_g\|$ represent their Euclidean norms, respectively. This similarity metric assesses the directional consistency of the model updates from clients relative to the global model. In our implementation, we establish a similarity metric mechanism by calculating the cosine similarity between the parameters of the client model and the global model layer by layer.

4) Aggregation methods for similarity weighting: In the traditional Federated Averaging (FedAvg) approach, the contribution of each client to the global model update is typically determined by the proportion of its data volume. Nevertheless, the degree of similarity between the client models and the global model over several training rounds is

not taken into consideration by this strategy. To address this limitation, we propose an adjustment to the aggregation process by incorporating the cosine similarity between the client models and the global model as a weighting factor.

This process ensures that the model's privacy is further reinforced while maintaining the integrity of the gradient through cropping. By appropriately calibrating the noise intensity, we can maximize the privacy of user data without compromising the accuracy of the model.

During each training round, we first calculate the cosine similarity between each client model and the global model. To ensure the resulting weights are reasonable, we normalize these similarity values so that the sum of the weights of all clients equals 1. This normalization process effectively adjusts the contributions of the clients, allowing updates from clients that exhibit higher similarity to the global model to carry greater weight in the aggregation process, while minimizing the impact of clients with lower similarity on the global model updates.

Let S_i denote the similarity of the i-th client model. The corresponding weighting factor w_i is then computed using the following equation:

$$w_i = \frac{S_i}{\sum_{j=1}^{N} S_j} \tag{16}$$

where N represents the total number of clients participating in the training, and S_j is the similarity of the j-th client model. This weighting method facilitates a dynamic weighted aggregation strategy based on similarity, enhancing the efficiency and effectiveness of the model training process.

5) Polymerization update: Building upon the weights calculated from model similarity, this paper employs a weighted aggregation strategy to update the global model parameters. Each client model's updated value is weighted and superimposed according to its corresponding weights, resulting in the final updated value of the global model. The aggregation process is outlined as follows:

First, for each client model's update result, the update is multiplied by its corresponding weighting factor. Subsequently, the weighted update values of all clients are accumulated layer by layer. Specifically, let the update value of the i-th client be denoted as Δw_i. The update value of the global model, Δw_g, is then computed using the following formula:

$$\Delta w_g = \sum_{i=1}^{N} w_i \cdot \Delta w_i \tag{17}$$

Where w_i is the weight of the i-th client, and Δw_i represents its corresponding model update value. The aggregated update value Δw_g is subsequently applied to the global model to complete each round of model updates.

By utilizing this similarity-weighted aggregation strategy, the global model not only synthesizes data features from diverse clients but also dynamically adjusts the influence of each client based on model similarity. In the presence of diverse data distributions, this method improves the model's generalization performance.

V. Experiments

A. Experimental Setup

1) Dataset and models: We assessed DP-FedSim's performance against cutting-edge algorithms in a federated learning environment on a variety of image recognition tasks. Fashion-MNIST [39], SVHN [40], and CIFAR-10 [41] were among the datasets used in this assessment.A test set of 10,000 samples and a training set of 60,000 samples make up the Fashion-MNIST dataset. A 28 × 28 grayscale picture linked to a label from a total of 10 classes represents each sample. The 60,000 color, 32 × 32 pixel pictures that make up the CIFAR-10 dataset are divided into 10 different classes, with 6,000 images in each class. The digit classification-focused SVHN dataset, which consists of 26,032 test samples and 73,257 training samples, is taken from Street View photos. Every sample is a 32 x 32 color picture that shows the numbers 0 through 9.

For the model architectures, FEMNIST employs a simple convolutional neural network (CNN) comprising 2 convolutional layers and 2 fully connected layers. CIFAR-10, on the other hand, has a more intricate architecture that consists of three convolutional layers and three fully linked layers. The SVHN dataset is processed using a straightforward model featuring 2 convolutional layers, 1 pooling layer, and 2 fully connected layers.

2) Benchmarks: We evaluate DP-FedSim's performance against a number of cutting-edge federated learning techniques, including FedAvg [1], DP-FedAvg[42], DP-FedSAM [20], and DP-FedSAM-top [20]. The FedAvg algorithm serves as a baseline federated learning method that operates without noise. In contrast, DP-FedAvg guarantees client-level differential privacy (DP) by applying a Gaussian mechanism directly to the local updates. The DP-FedSAM algorithm addresses the adverse effects of differential privacy through the utilization of gradient perturbations, specifically incorporating a Sharpness Aware Minimization (SAM) optimizer to produce locally flat models that exhibit improved stability and robustness against weight perturbations. Additionally, DP-FedSAMtopk is a variant of DP-FedSAM that employs a top-k update thinning technique, further minimizing the magnitude of random noise by updating only the most significant portions of the model updates, thereby enhancing model performance while preserving privacy.

3) Implementation details: For our experiments involving the Fashion-MNIST, SVHN, and CIFAR-10 datasets, we model data heterogeneity across client datasets by partitioning local data from the original dataset using a Dirichlet sampling process. The sampling parameter α controls the degree of imbalance in data distribution among clients; larger values of α correspond to weaker data heterogeneity, while smaller values imply stronger heterogeneity. Our primary evaluation metric is global accuracy. In comparisons with other

algorithms, we assess accuracy under varying degrees of non-independent and identically distributed (non-IID) data partitioning, proving that our customized federated learning approach is successful in handling non-IID data.

For the Fashion-MNIST, SVHN, and CIFAR-10 datasets, we set the learning rate to 1×10^{-3} and λ to 0.4. The parameters for differential privacy are set to $\varepsilon = 2$ and $\delta = 1\times10^{-5}$. We establish the number of global rounds at 100, local update rounds at 4, and batch size at 64, with the number of clients set to 100 and a sampling rate of 0.1.

The following sections of this paper are organized into three main parts: first, we conduct a comparative analysis of our algorithm against existing differential privacy federated learning methods. Second, we perform two ablation studies focusing on adaptive gradient cropping and model similarity-based aggregation to validate their effectiveness. Finally, we present a hyperparameter analysis to further elucidate the model's performance.

B. Performance Evaluation

1) Comparative analysis: In Table I, we evaluate the global accuracy of four baseline algorithms across three datasets: Fashion-MNIST, SVHN, and CIFAR-10. To assess the impact of data heterogeneity on the performance of these algorithms, we compare all baselines while varying α within the range of {0.05, 0.1, 0.2}. The results summarized in Table 1 indicate that our proposed algorithm demonstrates superior accuracy and generalization ability under standard noise conditions. This finding underscores the enhanced performance of personalized federated learning with differential privacy.

TABLE I. OPTIMAL TEST ACCURACY OF DP-FEDSIM AND CENTRALIZED BASELINES AT DIFFERENT NON-IID SETTINGS

Data	α	FedAvg	DP-FedAvg	DP-FedSAM	DP-FedSAM-top-k	DP-FedSim
Fminst	0.05	85.12	54.48	58.37	58.64	69.36
	0.1	93.25	62.10	65.95	66.51	71.68
	0.2	93.41	64.99	71.60	72.13	72.93
SVHN	0.05	85.47	77.12	50.21	53.01	84.27
	0.1	86.27	85.28	65.89	66.78	85.83
	0.2	88.64	86.16	72.69	74.53	86.79
Cifar10	0.05	54.62	51.69	45.28	45.78	60.63
	0.1	58.86	55.76	56.78	57.02	63.01
	0.2	64.98	61.44	58.41	59.77	64.07

For instance, in the non-independent identically distributed (non-IID) setting with α= 0.2, the accuracy achieved by our algorithm on the FEMNIST dataset is 78.93%, 86.79% on the SVHN dataset, and 64.07% on CIFAR-10. It is evident that the optimal accuracies of DP-FedSim consistently surpass those of the other baseline algorithms in most cases, highlighting the effectiveness of our approach in improving computational accuracy.

Furthermore, Table I illustrates the robustness and generalization capabilities of our algorithms under varying levels of non-IID distribution, specifically with α set at 0.05, 0.1, and 0.2. The heterogeneous distribution settings among local clients complicate the training and convergence of the global model. Notably, among the four baseline algorithms, the adverse effects of heterogeneous distribution become more pronounced as α decreases.

On the SVHN dataset, our proposed algorithm (Algorithm 1) exhibits superior convergence and generalization compared to DP-FedAvg as the non-IID level diminishes. When $\alpha = 0.05$, the accuracy of Algorithm 1 exceeds that of DP-FedAvg by 7.15%, indicating a greater adaptability of Algorithm 1 in handling non-independent homogeneous distributions. Additionally, on the CIFAR-10 dataset, Algorithm 1 demonstrates differences in non-independent homogeneous distribution of 2.38% and 1.06% at varying levels of α, which are notably lower than the 4.07% and 5.68% observed for DP-FedAvg, and significantly less than the 11.24% and 2.75% exhibited by DP-FedSAMtopk. These results confirm that, despite the challenges posed by heterogeneous data, Algorithm 1 remains resilient and exhibits enhanced robustness and stability.

C. Ablation Experiment

We carried out a number of ablation tests to clarify the role that each element of our strategy had in the overall performance. Our proposed method encompasses two key components: adaptive gradient cropping (AGC) and model similarity-based aggregation (MSA). To assess their individual impacts, we explored several variant configurations, including the removal of both adaptive gradient cropping and model similarity-based aggregation, the removal of adaptive gradient cropping while retaining model similarity-based aggregation, and the removal of model similarity-based aggregation while utilizing only adaptive gradient cropping.

- In the first variant, we eliminated both adaptive gradient cropping and model similarity-based aggregation, opting for the commonly employed differential privacy and simple average weighted aggregation methods based on fixed gradient cropping. This configuration serves as the baseline model for our comparative analysis.
- In the second variant, adaptive gradient cropping was removed, and we employed the conventional differential privacy method utilizing fixed gradient cropping for training, while maintaining model similarity-based aggregation for server-side aggregation. This setup allows us to evaluate the effectiveness of the model similarity-based aggregation method.
- The third variant involved the removal of the model similarity-based aggregation component, utilizing a simple average weighted aggregation method to assess

the effectiveness of the adaptive gradient cropping technique.

In this context, for the first variant, the differential privacy aspect implemented fixed gradient cropping with a cropping threshold c set to 0.4. All other parameter settings remained consistent with those outlined in Section V. The specific experimental results are presented in Table II.

TABLE II. ABLATION STUDY WITH DIFFERENT PRIVACY BUDGETS

Data	ACG	MSA	ε=2	ε=4	ε=8
Fminst			57.74	60.21	62.40
		✓	58.34	61.14	62.67
	✓		70.21	72.82	74.01
	✓	✓	72.31	73.62	74.78
SVHN			72.19	73.17	74.51
		✓	73.59	75.24	77.51
Cifar10			45.77	48.31	49.46
		✓	47.44	49.09	49.74
	✓		61.12	63.84	65.73
	✓	✓	64.62	65.88	67.26

The experimental data reveal that the removal of both modules resulted in a notable decrease in model accuracy, thereby underscoring the importance of each component in the modeling framework. Adaptive Gradient Cropping. The implementation of adaptive gradient cropping significantly enhances accuracy across various privacy budgets. Adaptive gradient cropping significantly improves accuracy across all privacy budgets when used in customized federated learning (as seen in the third row of each dataset in Table 2 as opposed to the baseline model (first row of each dataset). This finding validates the effectiveness of adaptive gradient cropping in bolstering the model's performance.

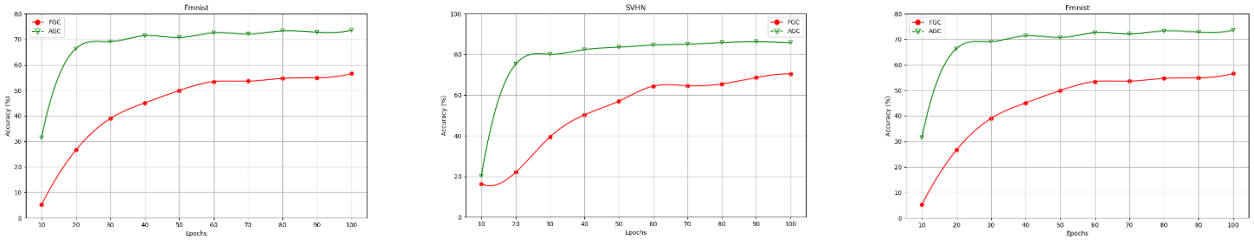


Fig. 3. Fixed gradient cropping and adaptive gradient cropping accuracy plots.

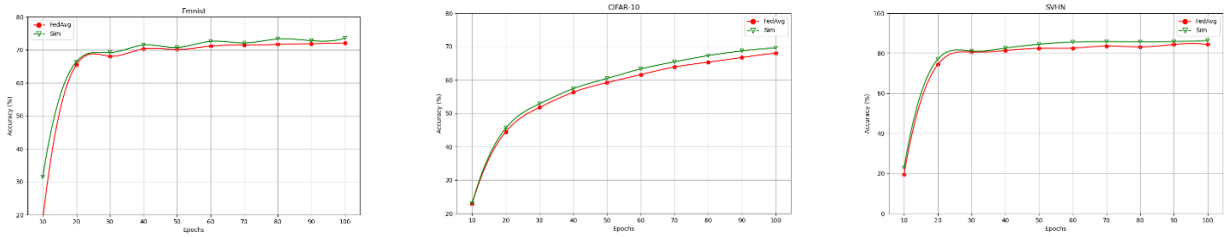


Fig. 4. Weighted average aggregation and model similarity based model accuracy plots.

Model Similarity-Based Aggregation. Similarly, when personalized federated learning relies solely on model similarity-based aggregation (as illustrated in the second row of each dataset in Table 2, there is a marked improvement in accuracy across the privacy budgets relative to the baseline model. This result further corroborates the effectiveness of model similarity-based aggregation within this framework.

Moreover, when both adaptive gradient cropping and model similarity-based aggregation are utilized in tandem (as shown in the fourth row of each dataset in Table 2), the accuracy demonstrates improvement across different privacy budgets compared to the baseline model, the model utilizing adaptive gradient cropping alone, and the model employing model similarity-based aggregation alone. The results clearly indicate that both adaptive gradient cropping and model similarity-based aggregation significantly contribute to the overall performance of the model under varying privacy budgets. The synergistic combination of these two components yields optimal results, thereby enhancing the effectiveness of our proposed algorithm.

D. Hyperparametric Analysis

In experiments concerning personalized federated learning with differential privacy, the selection of hyperparameters significantly influences both the performance and training efficiency of the model. This paper specifically examines the impact of the hyperparameter related to the number of clients on model performance, conducting tests across two distinct datasets.

As illustrated in Figures 5, when the number of clients is set to 5, 10, and 20 for both the Fashion-MNIST and CIFAR-10 datasets, the experimental results indicate a positive correlation between the number of clients and the overall accuracy of the model. The participation of a larger number of clients enables the system to leverage a broader range of local data, thereby enhancing the global model's generalization capability. Furthermore, an increased client count results in a data distribution that more closely reflects real-world scenarios, which helps mitigate the adverse effects of individual client data biases on the model's performance.

However, this increase in client numbers is accompanied by a significant rise in training time. This phenomenon is primarily attributed to the augmented participation of clients in local computations and model aggregation during each training round, leading to increased communication overhead and computational demands. Notably, with the implementation of differential privacy mechanisms, as the number of clients rises and the volume of data per client decreases, the number of communication rounds necessary to achieve a comparable level of convergence may increase, thereby exacerbating the overall training time.

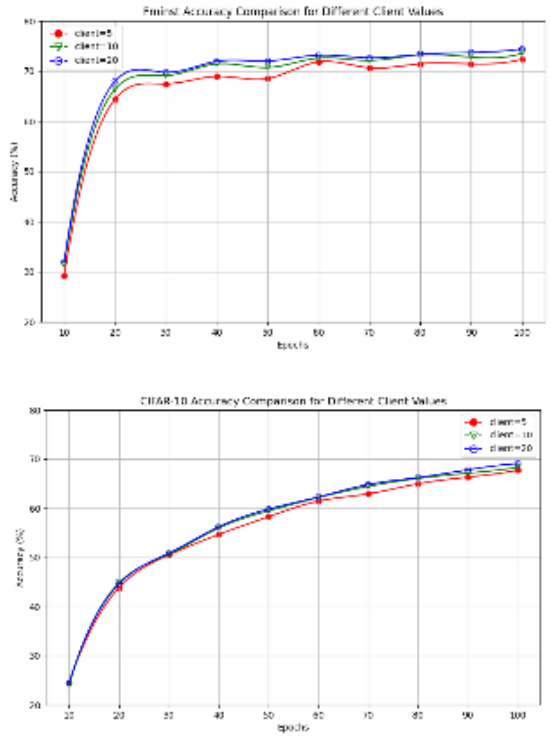


Fig. 5. Accuracy curve for different number of clients.

Consequently, selecting the optimal number of clients necessitates a careful balance between model performance and training efficiency. In scenarios where accuracy is paramount, increasing the number of clients can substantially enhance the model's generalization ability. Conversely, in time-sensitive training contexts, it is imperative to regulate the number of clients to mitigate computation and communication overhead. In practical applications, the number of clients can be adjusted according to specific requirements to achieve an optimal trade-off between performance and efficiency.

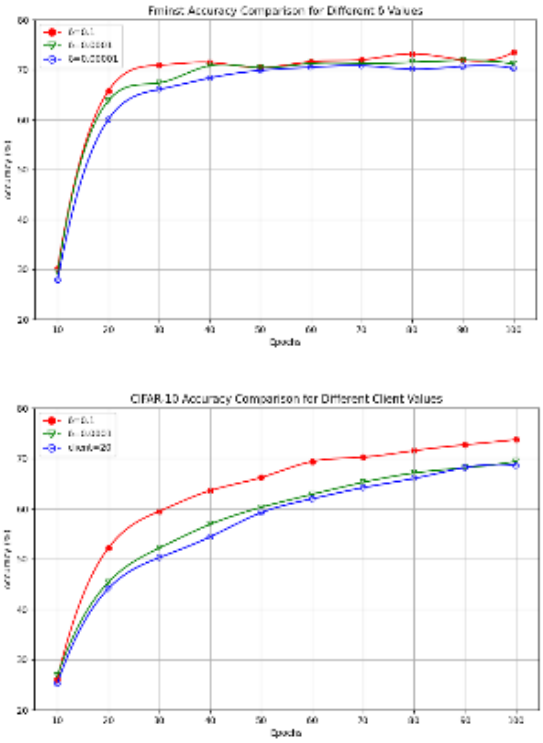


Fig. 6. Accuracy curve for different values of δ.

In Figures 6, we investigate the impact of varying δ-values (0.1, 0.0001, and 0.00001) on the performance of our model using the Fashion-MNIST and CIFAR-10 datasets. The experimental results demonstrate that model accuracy improves as the δ-value increases. This observation can be attributed to the fact that a larger δ-value signifies weaker privacy protection, resulting in reduced noise interference during the training process. Consequently, the model is able to extract useful information from the data more effectively, thereby enhancing its overall accuracy.

However, while a higher δ-value may yield performance benefits, it is crucial to acknowledge that it cannot be set excessively high within the framework of differential privacy. According to the principles of differential privacy, the δ-value represents the probability that the algorithm may violate the privacy budget. A large δ-value consequently diminishes the security of the differential privacy mechanism. If δ is set too high, the efficacy of privacy protection becomes questionable, potentially exposing sensitive data to the risk of leakage.

Therefore, in practical applications, the choice of parameter δ needs to be based on the differential privacy framework, which is a trade-off between model accuracy and privacy protection strength. Differential privacy achieves privacy protection by adding noise or data perturbation, and its privacy budget parameter ε and relaxation parameter δ together determine an upper bound on the risk of privacy leakage. Specifically, δ denotes the probability threshold that the algorithm cannot satisfy strict ε-differential privacy; a smaller

value of δ enhances the privacy guarantee but may lead to a decrease in the model's utility; conversely, increasing δ may enhance the model's performance but increase the likelihood of sensitive information exposure. For example, a decrease in the noise scale will reduce the perturbation to the training data distribution, but will weaken the strictness of the privacy boundaries. Therefore, it is recommended to experimentally quantify the effects of different (ε,δ) combinations on the model metrics according to the sensitivity requirements of the application scenarios, and ultimately choose the optimal parameter configurations that can satisfy the privacy authentication criteria while maintaining the model usability.

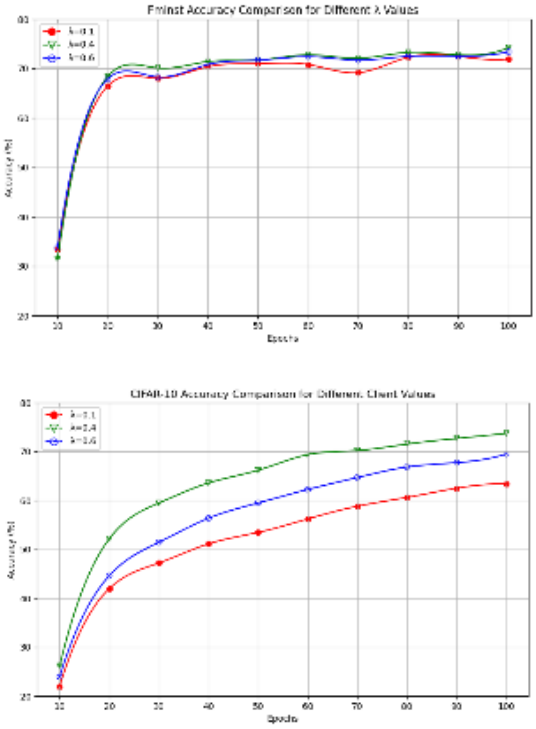


Fig. 7. Accuracy curve for different values of λ.

The value λ is a crucial hyperparameter in customized federated learning, affecting the weight balance between the global and local models in the experiment shown in Figure 7. By adjusting λ, the system can regulate the extent of fusion between the global model and the personalized model during the aggregation process. In our experiments, we set λ to values of 0.1, 0.4, and 0.6 and evaluated the model's performance across different datasets.

The results indicate that setting λ to 0.4 yields the best performance, achieving high accuracy on both the Fashion-MNIST and CIFAR-10 datasets. Specifically, when λ is 0.4, the model strikes an optimal balance between global generalization and local personalization, effectively maintaining a degree of personalization while retaining the shared knowledge encapsulated in the global model. This finding suggests that a moderate fusion of global and local models can enhance the overall performance of personalized federated learning, highlighting the importance of carefully selecting λ to achieve the desired model efficacy.

VI. Conclusion

DP-FedSim is a customized federated learning system with adaptive differential privacy that we present in this paper. This framework effectively addresses the limitations of traditional personalized federated learning, particularly its inflexibility in handling data heterogeneity, while also mitigating the adverse effects of additive noise associated with differential privacy on model performance. DP-FedSim leverages the properties of Fisher's information entropy matrix to accurately quantify the significance of model parameters, allowing for the retention of parameters with larger Fisher values. This strategy reduces the detrimental impact of noise addition on model efficacy. From the perspective of differential privacy, we introduce a hierarchical adaptive gradient cropping method that enables the system to automatically adjust the cropping threshold based on current privacy protection requirements and the real-time state of the model. During the model aggregation phase, the server evaluates the similarity between model parameters by computing metrics such as cosine similarity and dynamically modifies the contribution of each client model to the global model update. This adaptive approach enhances the framework's ability to accommodate the diverse data distributions and model qualities present among different clients. We apply the proposed algorithm to the Fashion-MNIST, SVHN, and CIFAR-10 datasets, demonstrating that our model achieves superior accuracy compared to other differential privacy algorithms, as evidenced by comparative experiments against state-of-the-art models. Furthermore, we conduct ablation experiments to analyze the contribution of each component to the overall performance of the model, while also discussing the rationale behind our hyperparameter settings through detailed hyperparameter analysis.

References

[1] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). PMLR.

[2] Tan, Y., Liu, Y., Long, G., Jiang, J., Lu, Q., & Zhang, C. (2023, June). Federated learning on non-iid graphs via structural knowledge sharing. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 8, pp. 9953-9961).

[3] Ye, M., Fang, X., Du, B., Yuen, P. C., & Tao, D. (2023). Heterogeneous federated learning: State-of-the-art and research challenges. ACM Computing Surveys, 56(3), 1-44.

[4] Huang, W., Ye, M., Shi, Z., & Du, B. (2023). Generalizable heterogeneous federated cross-correlation and instance similarity learning. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[5] Fang, X., Ye, M., & Yang, X. (2023). Robust heterogeneous federated learning under data corruption. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5020-5030).

[6] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and trends® in machine learning, 14(1–2), 1-210.

[7] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE signal processing magazine, 37(3), 50-60.

[8] Fang, X., & Ye, M. (2022). Robust federated learning with noisy and heterogeneous clients. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10072-10081).

[9] Huang, W., Ye, M., & Du, B. (2022). Learn from others and be yourself in heterogeneous federated learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10143-10153).

[10] Arivazhagan, M. G., Aggarwal, V., Singh, A. K., & Choudhary, S. (2019). Federated learning with personalization layers. arxiv preprint arxiv:1912.00818.

[11] Liang, P. P., Liu, T., Ziyin, L., Allen, N. B., Auerbach, R. P., Brent, D., ... & Morency, L. P. (2020). Think locally, act globally: Federated learning with local and global representations. arxiv preprint arxiv:2001.01523.

[12] Ma, X., Zhang, J., Guo, S., & Xu, W. (2022). Layer-wised model aggregation for personalized federated learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10092-10101).

[13] Tan, A. Z., Yu, H., Cui, L., & Yang, Q. (2022). Towards personalized federated learning. IEEE transactions on neural networks and learning systems, 34(12), 9587-9603.

[14] Fredrikson, M., Jha, S., & Ristenpart, T. (2015, October). Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security (pp. 1322-1333).

[15] Melis, L., Song, C., De Cristofaro, E., & Shmatikov, V. (2019, May). Exploiting unintended feature leakage in collaborative learning. In 2019 IEEE symposium on security and privacy (SP) (pp. 691-706). IEEE.

[16] Wen, Y., Geiping, J., Fowl, L., Goldblum, M., & Goldstein, T. (2022). Fishing for user data in large-batch federated learning via gradient magnification. arXiv preprint arXiv:2202.00580.

[17] Huang, Y., Gupta, S., Song, Z., Li, K., & Arora, S. (2021). Evaluating gradient inversion attacks and defenses in federated learning. Advances in neural information processing systems, 34, 7232-7241.

[18] Geyer, R. C., Klein, T., & Nabi, M. (2017). Differentially private federated learning: A client level perspective. arxiv preprint arxiv:1712.07557.

[19] Cheng, A., Wang, P., Zhang, X. S., & Cheng, J. (2022). Differentially private federated learning with local regularization and sparsification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10122-10131).

[20] Shi, Y., Liu, Y., Wei, K., Shen, L., Wang, X., & Tao, D. (2023). Make landscape flatter in differentially private federated learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 24552-24562).

[21] Wei, K., Li, J., Ding, M., Ma, C., Su, H., Zhang, B., & Poor, H. V. (2021). User-level privacy-preserving federated learning: Analysis and performance optimization. IEEE Transactions on Mobile Computing, 21(9), 3388-3401.

[22] Ma, X., Zhang, J., Guo, S., & Xu, W. (2022). Layer-wised model aggregation for personalized federated learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10092-10101).

[23] Liang, P. P., Liu, T., Ziyin, L., Allen, N. B., Auerbach, R. P., Brent, D., ... & Morency, L. P. (2020). Think locally, act globally: Federated learning with local and global representations. arxiv preprint arxiv:2001.01523.

[24] T Dinh, C., Tran, N., & Nguyen, J. (2020). Personalized federated learning with moreau envelopes. Advances in neural information processing systems, 33, 21394-21405.

[25] Li, Y., Yang, S., Ren, X., Shi, L., & Zhao, C. (2023). Multi-Stage Asynchronous Federated Learning with Adaptive Differential Privacy. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[26] Qi, P., Chiaro, D., Guzzo, A., Ianni, M., Fortino, G., & Piccialli, F. (2024). Model aggregation techniques in federated learning: A comprehensive survey. Future Generation Computer Systems, 150, 272-293.

[27] Pillutla, K., Kakade, S. M., & Harchaoui, Z. (2022). Robust aggregation for federated learning. IEEE Transactions on Signal Processing, 70, 1142-1154.

[28] Jhunjhunwala, D., Wang, S., & Joshi, G. (2024, April). FedFisher: Leveraging Fisher Information for One-Shot Federated Learning. In International Conference on Artificial Intelligence and Statistics (pp. 1612-1620). PMLR.

[29] Bietti, A., Wei, C. Y., Dudik, M., Langford, J., & Wu, S. (2022, June). Personalization improves privacy-accuracy tradeoffs in federated learning. In International Conference on Machine Learning (pp. 1945-1962). PMLR.

[30] Oh, J., Kim, S., & Yun, S. Y. (2021). Fedbabu: Towards enhanced representation for federated image classification. arxiv preprint arxiv:2106.06042.

[31] Li, X., Jiang, M., Zhang, X., Kamp, M., & Dou, Q. (2021). Fedbn: Federated learning on non-iid features via local batch normalization. arxiv preprint arxiv:2102.07623.

[32] Zhang, J., Hua, Y., Wang, H., Song, T., Xue, Z., Ma, R., ... & Guan, H. (2023). Gpfl: Simultaneously learning global and personalized feature information for personalized federated learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5041-5051).

[33] Huang, Y., Chu, L., Zhou, Z., Wang, L., Liu, J., Pei, J., & Zhang, Y. (2021, May). Personalized cross-silo federated learning on non-iid data. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 9, pp. 7865-7873).

[34] Li, T., Zaheer, M., Reddi, S., & Smith, V. (2022, June). Private adaptive optimization with side information. In International Conference on Machine Learning (pp. 13086-13105). PMLR.

[35] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems, 2, 429-450.

[36] Wang, J., Liu, Q., Liang, H., Joshi, G., & Poor, H. V. (2020). Tackling the objective inconsistency problem in heterogeneous federated optimization. Advances in neural information processing systems, 33, 7611-7623.

[37] Duan, J. H., Li, W., Zou, D., Li, R., & Lu, S. (2023). Federated learning with data-agnostic distribution fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8074-8083).

[38] Dwork, C. (2006, July). Differential privacy. In International colloquium on automata, languages, and programming (pp. 1-12). Berlin, Heidelberg: Springer Berlin Heidelberg.

[39] Caldas, S., Duddu, S. M. K., Wu, P., Li, T., Konečný, J., McMahan, H. B., & Talwalkar, A. (2018). Leaf: A benchmark for federated settings. arxiv preprint arxiv:1812.01097.

[40] Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., & Ng, A. Y. (2011, December). Reading digits in natural images with unsupervised feature learning. In NIPS workshop on deep learning and unsupervised feature learning (Vol. 2011, No. 2, p. 4).

[41] Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.

[42] McMahan, H. B., Ramage, D., Talwar, K., & Zhang, L. (2017). Learning differentially private recurrent language models. arxiv preprint arxiv:1710.06963.

Smart Night-Vision Glasses with AI and Sensor Technology for Night Blindness and Retinitis Pigmentosa

Shaheer Hussain Qazi, M. Batumalay
Faculty of Data Science and Information Technology, INTI International University, Nilai, Malaysia

***Abstract*—This paper presents the conceptualization of Smart Night-Vision Glasses, an innovative assistive device aimed at individuals with night blindness and Retinitis Pigmentosa (RP). These conditions, characterized by significant difficulty in seeing in low-light or dark environments, currently have no effective medical solution. The proposed glasses utilize advanced sensor technologies such as LiDAR, infrared, and ultrasonic sensors, combined with artificial intelligence (AI), to create a real-time, visual representation of the surroundings. Unlike conventional camera-based systems, which require light to function, this device relies on non-visible, non-harmful rays to detect environmental data, making it suitable for use in pitch-dark conditions. The AI processes the sensor data to generate a simplified, user-friendly view of the environment, outlined with clear, cartoon-like visuals for easy identification of objects, obstacles, and surfaces. The glasses are designed to look like regular prescription eyewear, ensuring comfort and discretion, while a button or trigger can switch them to "night mode" for enhanced vision in low-light settings. This concept aims to improve the independence, safety, and quality of life for individuals with night blindness and RP, offering a transformative solution where no medical alternatives currently exist. However, challenges such as sensor miniaturization, power consumption, and AI integration must be addressed for successful implementation. Beyond its direct benefits for users, the device could have broader societal and economic impacts by enhancing accessibility, reducing nighttime accidents, and fostering technological innovation in assistive wearables. The paper also discusses future directions for research and refinement of the technology while supporting the Process Innovation.**

Keywords*—Night-vision; glasses; night blindness; Retinitis Pigmentosa (RP); IoT; assistive technology; sensor technology; AI; data processing; low-light navigation; wearable devices; process innovation

I. Introduction

Night blindness and Retinitis Pigmentosa (RP) are significant visual impairments that severely restrict an individual's ability to navigate and function in low-light or nighttime environments. Night blindness, also known as nyctalopia, is a condition that reduces the eyes' ability to adjust to dim lighting, making it challenging for affected individuals to see clearly in conditions such as dusk, poorly lit areas, or complete darkness. Similarly, Retinitis Pigmentosa is a progressive degenerative disorder that damages the retina's photoreceptor cells, leading to tunnel vision, difficulty in perceiving peripheral objects, and eventual loss of night vision. Studies estimate that RP affects approximately 1 in 4,000 people globally, making it one of the leading causes of inherited blindness. Additionally, night blindness is a widespread symptom of various eye conditions, including Vitamin A deficiency and congenital disorders, impacting millions worldwide. These visual impairments significantly limit an individual's independence, mobility, and overall quality of life, underscoring the urgent need for effective assistive solutions.

Despite advancements in medical science, no definitive cure exists for either night blindness or RP. Current interventions focus on managing symptoms or slowing the progression of RP through dietary supplements, light therapy, or experimental gene therapy. However, these approaches do not restore functional vision or significantly improve navigation ability in low-light conditions. As a result, individuals with these conditions often rely on assistive devices or human aid to perform everyday tasks.

Technological solutions, such as night-vision goggles and camera-based systems, have been developed to assist individuals in low-visibility scenarios. However, these technologies come with inherent limitations. Night-vision goggles, widely used in military and industrial applications, amplify existing light but are often bulky, expensive, and unsuitable for daily civilian use. Camera-based systems, while more compact, require some amount of ambient light to function effectively, making them unreliable in complete darkness. They are also prone to latency issues, glare interference, unidentifiable environment in sight and privacy concerns, limiting their practicality for visually impaired individuals. These shortcomings highlight the pressing need for an innovative, accessible, and efficient solution tailored to individuals with night blindness and RP.

The primary objective of this research is to conceptualize a novel assistive device: Smart Night-Vision Glasses. These glasses are designed to empower individuals with night blindness and RP by enabling them to perceive their surroundings in low-light and no-light conditions. Unlike existing solutions, this concept utilizes advanced sensor technologies and artificial intelligence (AI) to create a real-time representation of the user's environment. The device would rely on non-visible, non-harmful rays, such as LiDAR or infrared, to gather environmental data without depending on visible light. This data would then be processed by AI

algorithms to generate a simplified, cartoon-like visualization of the surroundings, displayed directly on the glasses' lenses.

The glasses aim to be lightweight, discreet, and easy to use, closely resembling conventional eyewear. A user-friendly interface would allow individuals to switch between regular and night-vision modes with the press of a button, ensuring seamless integration into daily life. This approach not only offers functional assistance in darkness but also caters to aesthetic and comfort considerations for regular use.

The proposed smart glasses represent a significant leap forward in assistive technology for low-vision individuals. Unlike traditional camera-based night-vision systems, which rely on ambient light or emit visible light, this concept leverages non-visible rays such as LiDAR or infrared, which are both safe and effective in complete darkness. These sensors can create a detailed map of the environment by detecting surfaces, objects, and movement, irrespective of lighting conditions.

The innovation lies in the integration of AI-powered processing to transform raw sensor data into a cartoon-like outlined visual representation that can be projected on lenses of glasses. By simplifying complex environmental data into outlines and clear distinctions, the glasses can provide users with an intuitive and easily interpretable view of their surroundings. This approach is inspired by the visualization techniques used in video games and animated media, where objects and characters are outlined with thick lines for better differentiation.

Moreover, the glasses are designed to prioritize accessibility and practicality. Their discreet appearance and lightweight design address the stigma and inconvenience often associated with assistive devices. By incorporating real-time processing and a seamless mode-switching mechanism, the glasses aim to provide users with an enhanced sense of independence and confidence in various scenarios, from nighttime walks to driving and navigating unfamiliar environments.

This concept aligns with the United Nations Sustainable Development Goal (SDG) 9, which emphasizes fostering innovation and developing sustainable infrastructure. By focusing on Internet of Things (IoT) integration, advanced sensor technologies, and AI, this project promotes the creation of inclusive, forward-thinking solutions that enhance quality of life and address unmet needs in the assistive technology sector. The smart glasses not only represent a technological innovation but also hold the potential to transform how we approach visual impairments, particularly in conditions where no medical solutions currently exist.

II. Literature Review

The evolution of assistive technologies for visually impaired individuals has led to the development of various smart glasses and wearable devices aimed at improving mobility, navigation, and environmental perception. These technologies integrate artificial intelligence (AI), computer vision, and Internet of Things (IoT) components to assist users in performing daily activities independently. Despite these advancements, significant limitations persist, particularly concerning cost, usability, and effectiveness in low-light or nighttime conditions. This section critically examines existing assistive technologies, their shortcomings, and potential areas for improvement.

A highly sophisticated AI-based smart glasses design is presented by [1] that integrates multiple vision correction functionalities along with advanced navigation and security features. The proposed glasses dynamically adjust for far vision, computer use, and reading through ultrasound-based mechanisms while incorporating night-vision-enabled cameras for obstacle detection and environmental awareness. The inclusion of motion detection, explosive-trace identification, and blind-corner monitoring through strategically placed cameras and sensors highlights the potential of AI-driven wearable technology in enhancing both accessibility and security. While our conceptual design does not rely on cameras or auditory feedback, [1]'s work reinforced the importance of real-time environmental awareness and adaptive visual augmentation. His approach to integrating AI-based obstacle detection helped refine our understanding of non-visual sensing methods, particularly how sensor-based data interpretation could be leveraged to provide a seamless navigation aid for individuals with night blindness and RP.

The Internet of Things (IoT) continues to emerge as a transformative technology, facilitating solutions that improve human life. A significant portion of the global population faces disabilities that complicate daily activities, particularly those with visual impairments. While various assistive tools have been developed to aid individuals with blindness, many of these solutions fall short in terms of accessibility and efficiency. Integrating artificial intelligence (AI) into such assistive devices emphasized by [2] significantly enhances their utility by offering users a simulated sense of vision, ultimately promoting independence. However, despite advancements, challenges persist regarding affordability, compact design, and the incorporation of essential functionalities.

Assistive technology (AT) plays a crucial role in improving the daily lives of individuals with disabilities by enabling greater independence. It encompasses a wide range of devices and services tailored to specific needs, including mobility aids, vision-enhancing products, and cognitive assistance tools. The importance of AT in bridging the accessibility gap for visually impaired individuals is underscored by [2], highlighting how modern innovations in AI-driven object detection and real-time voice feedback are redefining the scope of such technologies.

One of the primary challenges faced by visually impaired individuals is navigation in public spaces. According to [2], IoT-based solutions are increasingly addressing these challenges, introducing a variety of navigation aids designed specifically for users with limited or no vision. The development of smart glasses equipped with advanced sensors and deep learning algorithms presents a promising avenue for enhancing mobility and independence. The functionality of such smart glasses was explored by [3], emphasizing the role of adaptive algorithms and deep learning techniques in object recognition and real-time environmental interpretation. The ability of such devices to process visual data accurately enables

visually impaired users to interact with their surroundings more effectively.

Moreover, [3] highlights the growing importance of functional analysis and statistical mechanics in assessing the strengths and limitations of smart glasses. These analytical methods provide a deeper understanding of how AI-driven assistive devices operate in real-world conditions, paving the way for future advancements. The practical application of smart glasses extends beyond functional performance, as [3] notes that usability, ergonomic design, and social acceptability are equally vital considerations. A seamless integration into daily life, coupled with a lightweight and comfortable design, enhances the practicality of these devices for both indoor and outdoor use.

Beyond navigation, AI-powered image processing algorithms are instrumental in improving object recognition and text-to-speech conversion capabilities. Convolutional Neural Networks (CNNs) and its ability to enable smart glasses to detect and classify objects in real time is detailed by [3]. Similarly, Haar Cascade Classifiers, which rely on machine learning for object detection, are particularly effective in identifying specific patterns such as faces and gestures. These technological advancements underscore the growing precision of AI-driven visual assistance tools.

Obstacle detection and avoidance remain fundamental features of smart glasses designed for visually impaired individuals. Additionally, [3] discusses how these systems leverage a combination of sensors, algorithms, and feedback mechanisms to enhance navigation safety. Technologies such as LiDAR (Light Detection and Ranging) and time-of-flight (ToF) sensors provide real-time depth measurements, allowing users to navigate around obstacles more effectively. Additionally, computer vision techniques, including edge recognition and object segmentation, further improve the glasses' ability to analyze surroundings and guide users through audio or tactile feedback. Machine learning algorithms, such as CNNs, contribute to distinguishing obstacles from the background, enhancing accuracy and minimizing the risk of collisions.

The development of AI-powered smart glasses aimed at enhancing the independence and social integration of visually impaired individuals was explored by [4]. Their design incorporates multiple assistive functionalities, including text reading, currency recognition, color differentiation, obstacle detection, and facial recognition. By focusing on discreet and user-friendly features, the authors address the psychological and practical concerns of individuals who may prefer less conspicuous assistive devices. While our conceptual design does not involve cameras or facial recognition, this work provided valuable insight into the broader scope of smart glasses as accessibility tools. In particular, their emphasis on seamless obstacle detection reinforced the need for intuitive, real-time feedback, shaping our approach to integrating non-visual environmental awareness through LiDAR and ultrasonic sensing.

A comprehensive analysis of the integration of Artificial Intelligence (AI) and Visible Light Communications (VLC) in assistive technologies for visually impaired individuals was provided by [5]. Their work explores how VLC, an emerging communication technology using modulated light signals, can be leveraged for navigation and environmental awareness, complementing AI-driven assistance. By reviewing existing solutions and outlining a roadmap for AI-VLC integration, the authors highlight the transformative potential of these technologies in enhancing accessibility. While our concept does not incorporate VLC, this study broadened our perspective on alternative sensing methods beyond traditional camera-based or infrared solutions. The discussion on AI-driven early disease detection was particularly insightful, reinforcing the importance of optimizing assistive devices to cater to diverse visual impairments.

A smart glasses system designed to assist blind individuals by converting written English text into audio output using artificial intelligence was presented by [6]. Their approach integrates Optical Character Recognition (OCR), the EAST text detector, and ultrasonic sensors to capture and interpret text efficiently. Additionally, the use of motion sensors and RFID technology enables indoor navigation, particularly in structured environments like classrooms and lecture halls. While their device focuses on reading assistance rather than environmental awareness, it provided valuable insights into sensor-based guidance and real-time AI processing. This study reinforced the importance of designing intuitive and responsive assistive devices, which influenced our decision to prioritize real-time sensor data processing for enhanced navigation in low-light conditions. Our conceptual design doesn't have such features as our concept focusses on Night Blindness and Rp patients specifically.

SMART_EYE, a smart assistive technology was introduced by [7], aimed at helping visually impaired individuals navigate unfamiliar environments and detect obstacles using AI and sensor integration. Their system leverages a mobile application for image classification, while ultrasonic sensors handle real-time obstacle detection, providing auditory feedback via voice commands. A key contribution of their work is the emphasis on cost-effectiveness, addressing the affordability barriers that often limit the adoption of assistive devices. While our concept diverges by focusing on night vision enhancement rather than image-based classification, their research reinforced the importance of integrating real-time sensor feedback for seamless navigation. Their findings also highlighted the necessity of lightweight, user-friendly solutions, which further validated our approach of ensuring ergonomic and intuitive design in our Smart Night-Vision Glasses.

A comprehensive overview of how artificial intelligence, particularly deep learning, is transforming both the diagnosis of eye diseases and the development of assistive visual aids was conducted by [8]. Their research underscores the dual role of AI in early disease detection and the enhancement of everyday accessibility tools for visually impaired individuals. While our concept does not focus on AI-driven diagnostics, their discussion on smart devices reinforced the potential of AI in wearable assistive technologies. Their work highlighted the rapid advancements in deep learning applications, which further supported our decision to integrate AI-driven sensor processing for real-time navigation assistance. Additionally, their insights into the future directions of AI-assisted visual

technologies provided a broader perspective on how such innovations could evolve, aligning with our goal of leveraging AI to enhance spatial awareness for individuals with night blindness.

The challenges of object detection in low-light environments were explored by [9], emphasizing the limitations of traditional surveillance methods and the advantages of deep learning-based approaches using thermal infrared imaging. Their study highlights the difficulties of maintaining high detection accuracy at night due to poor illumination, a challenge that aligns with the core problem our concept aims to address. While their research is focused on security and surveillance, the insights on leveraging thermal imaging for object recognition reinforced the feasibility of using infrared sensors in our conceptual design of Smart Night-Vision Glasses. Their findings validated our approach of relying on non-visible light sources rather than conventional cameras, demonstrating the effectiveness of alternative sensing technologies for enhancing visibility in dark environments. Additionally, their discussion on deep learning's role in feature extraction and classification provided useful perspectives on potential future enhancements for intelligent obstacle recognition in assistive wearables.

An innovative approach was presented by [10] to self-powered sensor systems by leveraging triboelectric nanogenerators (TENG) for energy harvesting and signal processing, addressing key challenges in wearable technology, such as power sustainability and deployment flexibility. Their discussion on TENG's ability to convert mechanical stimuli into electrical signals offers valuable insights into potential advancements in assistive technology. While our proposed Smart Night-Vision Glasses do not incorporate sound recognition or energy harvesting, their research highlights the growing trend of integrating self-driven sensors with AI for real-time data processing, which could be relevant for future iterations of assistive wearables. The study reinforces the importance of low-power, intelligent sensor networks, aligning with our concept's focus on developing a compact and efficient solution for individuals with night blindness. Moreover, their exploration of machine learning applications in sensor signal processing provides a broader perspective on how AI can enhance the accuracy and responsiveness of wearable devices.

A comprehensive review of the key enabling technologies that drive advancements in autonomous vehicles (AVs) is provided by [11], focusing on the integration of IoT, edge intelligence, 5G, and blockchain to enhance safety, security, and efficiency. Their discussion of sensor networks and real-time data processing is particularly relevant to assistive technologies like our proposed Smart Night-Vision Glasses, which also rely on sensor fusion and AI-driven decision-making. While our concept does not involve vehicular automation, the paper highlights the broader potential of IoT-enabled systems in improving real-time navigation and obstacle detection—principles that align with our device's goal of enhancing spatial awareness for individuals with night blindness. Furthermore, their exploration of edge intelligence emphasizes the benefits of decentralized processing, reinforcing the importance of efficient, low-latency data handling, a crucial aspect for wearable assistive devices.

The role of computer vision and AI algorithms in autonomous driving was explored by [12], highlighting their application in scene perception, obstacle detection, and intelligent decision-making. Their discussion on sensor integration and real-time image processing resonates with the conceptual framework of our Smart Night-Vision Glasses, which also rely on AI-driven perception for enhanced visibility in low-light conditions. While our design does not incorporate computer vision for autonomous navigation, the paper underscores the significance of real-time data acquisition and preprocessing, key principles that inform our approach to sensor-based environmental awareness. The emphasis on obstacle detection and avoidance further reinforces the importance of adaptive assistive technologies, validating our decision to integrate multi-sensor fusion for enhanced spatial perception.

Some papers were considered beyond the scope of our concept, in order to get as many intuitive ideas as possible while designing a concept of future technology in human assistance. The role of digital technologies in optimizing energy grid integration is discussed by [13], emphasizing predictive analytics, monitoring, and control systems that enhance stability and efficiency. While our proposed Smart Night-Vision Glasses do not directly relate to energy management, the concept of real-time data processing and intelligent decision-making aligns with our approach to sensor-driven environmental awareness. The paper's insights into predictive analytics reinforce the importance of proactive adaptation in assistive technologies, much like how our design leverages AI to interpret sensor data and provide real-time feedback for improved night vision and obstacle detection.

By integrating cutting-edge AI technologies with wearable assistive devices, researchers continue to push the boundaries of accessibility and independence for visually impaired individuals. The growing body of literature in this domain highlights the potential of smart solutions to revolutionize how users interact with their environment, offering not only enhanced perception but also a deeper sense of autonomy.

A. Existing Assistive Technologies

Several smart solutions designed for visually impaired users have emerged in recent years, incorporating features such as object recognition, text reading, and navigation assistance. Notable products include SMART_EYE, OrCam, eSight, and Aira smart glasses. These devices leverage AI-driven image processing to enhance the user's understanding of their surroundings. For example, OrCam MyEye utilizes a small, camera-equipped module that attaches to traditional eyewear, allowing users to receive real-time audio descriptions of their environment. eSight employs high-resolution cameras to capture images and magnify them for individuals with low vision, while Aira connects users to remote human assistants who provide navigation guidance.

In addition to camera-based solutions, some assistive technologies integrate advanced sensors such as LiDAR (Light Detection and Ranging) and ultrasonic sensors. These technologies are particularly beneficial in low-light environments, as they generate depth maps and detect obstacles regardless of ambient lighting. However, despite their

potential, they are not widely adopted in wearable devices for visually impaired users.

B. Limitations of Current Systems

While the aforementioned technologies have contributed significantly to improving mobility for visually impaired individuals, they exhibit critical shortcomings that hinder widespread adoption. One of the primary issues is the over-reliance on light-dependent cameras. Most existing smart glasses function optimally in well-lit environments but fail to perform adequately in darkness or poor lighting conditions. Since night blindness and Retinitis Pigmentosa predominantly affect individuals' ability to see in low-light conditions, these devices do not fully address their needs.

Another major limitation is the reliance on audio-based feedback. Many smart glasses provide auditory cues to guide users, which can be problematic in noisy environments or situations where the user must remain aware of external sounds, such as traffic. Overloading the auditory senses can reduce situational awareness and pose safety risks.

Cost is another prohibitive factor. Many commercially available smart glasses are expensive, often exceeding several thousand dollars. This restricts accessibility for a large portion of the visually impaired population, particularly in developing countries where affordability is a significant concern.

Moreover, many of these devices are cumbersome, requiring additional hardware such as handheld controllers, smartphones, or external batteries. This lack of seamless integration into daily life reduces user adoption rates. In addition, devices that rely on cloud-based AI processing introduce latency issues, making real-time navigation less effective and frustrating for users.

C. Why Certain Technologies have Failed to Gain Widespread Adoption

Despite the promise of AI and IoT in assistive technology, several solutions have struggled to achieve mainstream adoption. One reason is the lack of personalization in existing systems. Current technologies often take a one-size-fits-all approach, failing to consider the varying degrees of visual impairment among users. Additionally, user interfaces are often unintuitive, requiring extensive training before users can fully benefit from the technology.

Another critical reason for limited adoption is the failure to address social and psychological barriers. Many visually impaired individuals prefer discreet assistive devices that blend seamlessly into everyday life. Bulky or conspicuous designs contribute to social stigma and discourage users from adopting the technology.

Battery life also remains a concern. Many smart glasses consume significant power due to continuous image processing and AI computations. Frequent recharging requirements and limited operational hours reduce their practicality for all-day use.

Furthermore, privacy concerns associated with camera-based systems deter users. Devices that continuously capture and process visual data may be perceived as intrusive in social or professional settings, raising ethical and legal considerations regarding data security and consent.

D. Opportunities for Improvement

The shortcomings of current assistive technologies highlight several areas where innovation is necessary. The most pressing issue is the need for improved night-vision capabilities. Future solutions should prioritize the integration of non-visible ray-emitting sensors, such as infrared (IR) and LiDAR, which can accurately detect objects and depth in complete darkness. LiDAR, in particular, has shown remarkable success in autonomous vehicles, where real-time spatial mapping is critical for navigation. Implementing similar sensor-based approaches in wearable assistive devices would significantly enhance their reliability in low-light environments.

Another area for improvement is the development of multi-modal feedback systems. Instead of relying solely on audio cues, future assistive devices should incorporate haptic feedback and visual overlays to guide users. Augmented reality (AR) displays that outline obstacles and objects in a simplified, high-contrast format could provide a more intuitive navigation experience while reducing dependence on sound-based instructions.

Reducing hardware bulk and enhancing user comfort should also be prioritized. Advances in miniaturization and power-efficient AI processing could lead to the creation of lightweight, discreet smart glasses that resemble traditional eyewear. Wireless charging and extended battery life would further enhance their usability.

Furthermore, affordability must be a key consideration in the development of next-generation assistive devices. Open-source software and cost-effective hardware solutions could make high-quality assistive technology more accessible to a broader audience.

Finally, privacy concerns must be addressed through on-device processing rather than cloud-based computation. Edge AI technologies, which allow real-time data analysis without transmitting information to external servers, could enhance security and alleviate users' concerns about continuous surveillance.

E. Conclusion

The review of existing literature highlights the progress made in assistive technologies for visually impaired individuals, particularly through AI-driven smart glasses. However, critical gaps remain, particularly regarding low-light performance, reliance on auditory feedback, high costs, and limited user adoption due to social and usability concerns. Future innovations should focus on sensor-based night-vision solutions, multi-modal feedback systems, ergonomic designs, and affordability to maximize the impact of assistive technologies.

By leveraging LiDAR, infrared, and other advanced sensors, wearable assistive devices can evolve beyond their current limitations and provide an effective, all-encompassing solution for individuals with night blindness and Retinitis Pigmentosa. This research aims to bridge these gaps by

conceptualizing an intelligent, sensor-driven smart glasses system that enhances mobility and independence while maintaining user comfort and discretion.

III. Conceptual Design of Smart Night-Vision Glasses

A. System Architecture

The proposed Smart Night-Vision Glasses present a conceptual solution aimed at enhancing mobility and spatial awareness for individuals with night blindness and Retinitis Pigmentosa (RP). By leveraging advanced sensor technology and AI-driven data processing, the system is designed to provide users with a real-time, visually intuitive representation of their surroundings, even in complete darkness. Unlike traditional night-vision solutions, which rely on cameras or amplified light, this design is entirely sensor-based, ensuring reliable performance across varying environmental conditions.

The core architecture of the glasses integrates multiple sensor modalities, including LiDAR (Light Detection and Ranging), infrared (IR), and ultrasonic transducers. These sensors can collaboratively construct a detailed environmental map, with LiDAR providing precise depth perception, IR capturing heat signatures, and ultrasonic sensors supplementing obstacle detection. The AI-powered processing unit interprets this data to generate a simplified, cartoon-like outline of the surroundings, which is then projected onto the glasses' lenses. The visual output on the lens of glasses can mirror the style of video games or cartoons, where objects and characters are outlined with bold, thick lines, ensuring easy differentiation. For example, walls might appear as clearly outlined flat surfaces, trees as simple silhouettes, and people or animals as distinctive shapes with identifiable boundaries. This approach ensures that users can perceive objects, obstacles, and spatial boundaries through an intuitive, high-contrast visual representation, enhancing navigation and situational awareness.

A key challenge in such an architecture is real-time processing, as sensor data must be collected, analyzed, and displayed with minimal latency to ensure a seamless user experience. The AI system must handle vast streams of depth, thermal, and ultrasonic data, transforming them into an intelligible format without noticeable delay. To achieve this, edge AI processing techniques are considered, allowing computations to be performed directly within the glasses rather than relying on external cloud-based processing. This not only reduces latency but also ensures uninterrupted functionality in all environments, including remote or offline settings.

B. Sensor Technology and Power Efficiency

The reliance on LiDAR, infrared, and ultrasonic sensors eliminates the need for traditional camera-based systems, addressing common issues such as sensitivity to glare, poor weather performance, and privacy concerns. LiDAR emits laser pulses that reflect off objects, allowing the system to construct a 3D depth map, while infrared detects heat-emitting entities, making it particularly useful for identifying living beings. Ultrasonic sensors provide an additional safety layer, especially in detecting obstacles in close proximity where LiDAR and IR may have limitations.

Fig. 1 shows how LIDAR is being used in today's Autonomous Vehicles and how it detects its surrounding obstacles. Whereas Fig. 2 shows what an Infrared vision looks like, with High Heat signature (Red-Orange) being mostly Living Creatures like humans, and Low Heat Signature (Blue-Green) being surrounding objects. If the sensors work together to collect data, it can be quite an efficient strategy to merge their data and get combined result using cutting-edge AI processing.

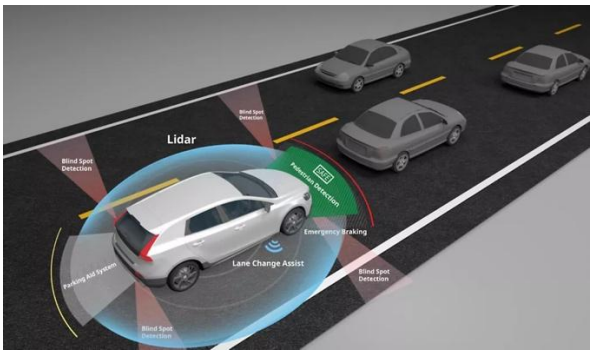


Fig. 1. LIDAR working in self-driving vehicles.

Fig. 2. Infrared vision.

However, one of the primary concerns in implementing these sensor technologies in a wearable device is power consumption. LiDAR systems, while highly accurate, can be power-intensive, depending on the scanning frequency and resolution. Infrared sensors also require continuous energy input, particularly when detecting thermal variations over a wide field of view. Ultrasonic sensors, although relatively low-power, still contribute to overall battery drain. To ensure prolonged battery life without compromising performance, the Smart Night-Vision Glasses would incorporate several power management strategies, such as:

- Adaptive Sensor Activation: Instead of running all sensors continuously, the system dynamically activates specific sensors based on the detected environment. For instance, LiDAR may operate at full resolution in complex spaces but switch to a lower-power mode in familiar environments.
- Efficient AI Processing: The AI model is optimized to minimize redundant computations, using compressed

depth mapping techniques and selective rendering to process only the most relevant visual elements.

- Low-Power Display Technology: The projection system within the lenses could employ energy-efficient micro-OLED or e-paper technology to reduce power consumption while maintaining high visibility.
- Rechargeable, High-Capacity Batteries: Advanced lithium-polymer batteries with optimized energy density would provide extended usage, supported by rapid charging mechanisms for convenience.

By incorporating these strategies, the Smart Night-Vision Glasses aim to strike a balance between high-performance sensing and practical battery life, ensuring users can rely on the device throughout their daily activities without frequent recharging.

C. AI Implementation and Data Processing

At the core of the proposed system lies an advanced AI model designed to process multimodal sensor data and generate an intuitive, real-time visual output. Unlike conventional AI-driven vision systems that rely on image recognition, this model could function exclusively on depth, thermal, and ultrasonic data, reconstructing a scene based purely on environmental contours and object boundaries. The AI processing pipeline can follow several key steps:

- Data Fusion and Preprocessing: The system first aggregates raw data from LiDAR, IR, and ultrasonic sensors. Noise reduction techniques, such as Kalman filtering and temporal smoothing, are applied to enhance signal clarity and reduce measurement inconsistencies.
- Edge Detection and Scene Reconstruction: Using deep learning-based edge extraction techniques, the AI can identify object contours and spatial structures. This can be achieved through modified Convolutional Neural Networks (CNNs) trained specifically on depth and thermal datasets. Unlike traditional image-based edge detection, this method could operate on geometric point clouds and thermal differentials.
- Cartoon-like Rendering and Display Optimization: The extracted object boundaries can then be stylized into a high-contrast visual format, resembling bold outlines in a simplified 3D space. This representation can be rendered using lightweight GPU-based processing or custom-designed FPGA hardware for minimal latency.
- Latency Optimization and Edge AI Deployment: Given that real-time responsiveness is critical, the AI model can be deployed on an embedded system featuring an ARM-based neural processing unit (NPU). This will reduce computational overhead and ensures that the glasses can function autonomously without reliance on external computing resources.

A significant challenge in this implementation can be to maintain real-time performance while handling large volumes of sensor data. LiDAR alone can generate millions of data points per second, and without efficient processing, the system could introduce delays. To counter this, the AI can employ hierarchical processing, prioritizing essential objects and filtering out irrelevant background data. Additionally, by leveraging sparsity-aware algorithms, the system could ensure that only critical edge information is rendered, thereby reducing computational load without sacrificing accuracy.

D. Visual Representation of System Design

To further clarify the conceptual framework of the Smart Night-Vision Glasses, the following simple diagrams illustrate the data flow, operational process, and system's architecture. These visual aids provide a structured understanding of how the proposed technology integrates various components to enhance night vision capabilities for individuals with night blindness and Retinitis Pigmentosa (RP).

The data flow diagram in Fig. 3 illustrates the sequence in which data moves through the system, ensuring real-time processing and visualization. When night vision mode is active, the Sensors continuously scan the environment, collecting depth, obstacle, and ambient light information. This data is transmitted to the Microcontroller, which aggregates and formats the input before passing it to the AI Processing Unit. Here, advanced machine learning algorithms analyze the data, filter out noise, and generate a simplified yet accurate spatial representation of the surroundings. The processed information is then transmitted to the Display Lenses, enabling users to visualize objects and obstacles in their path. Additionally, User Controls provide input that alters processing settings, allowing for adaptive visual representation based on user preferences or environmental conditions.

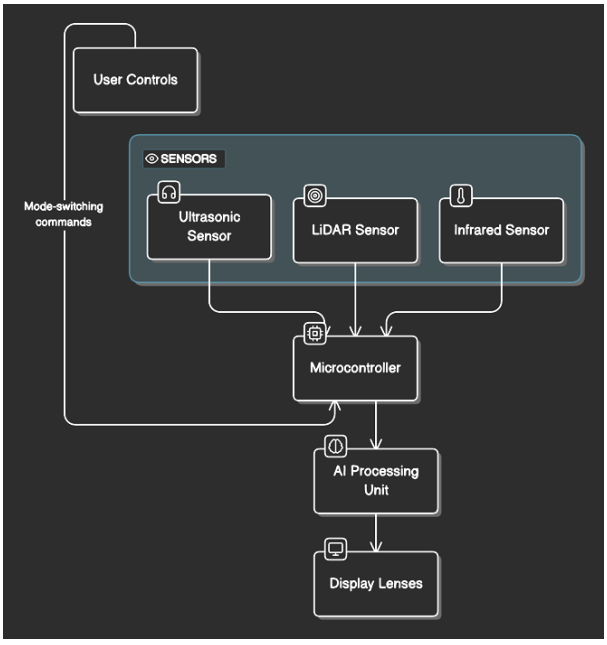


Fig. 3. Conceptual data flow diagram.

The flowchart in Fig. 4 details the step-by-step operational process of the Smart Night-Vision Glasses, from activation to data visualization and user interaction. The user can manually switch between normal and night vision modes via User Controls. When the user activates the night vision, the system initiates sensor data collection, where LiDAR, infrared, and

ultrasonic sensors gather environmental depth and object proximity information. The data is then processed by the AI-driven computational unit, which generates a structured visual overlay displayed on the lenses. This structured flow ensures efficient operation with minimal latency.

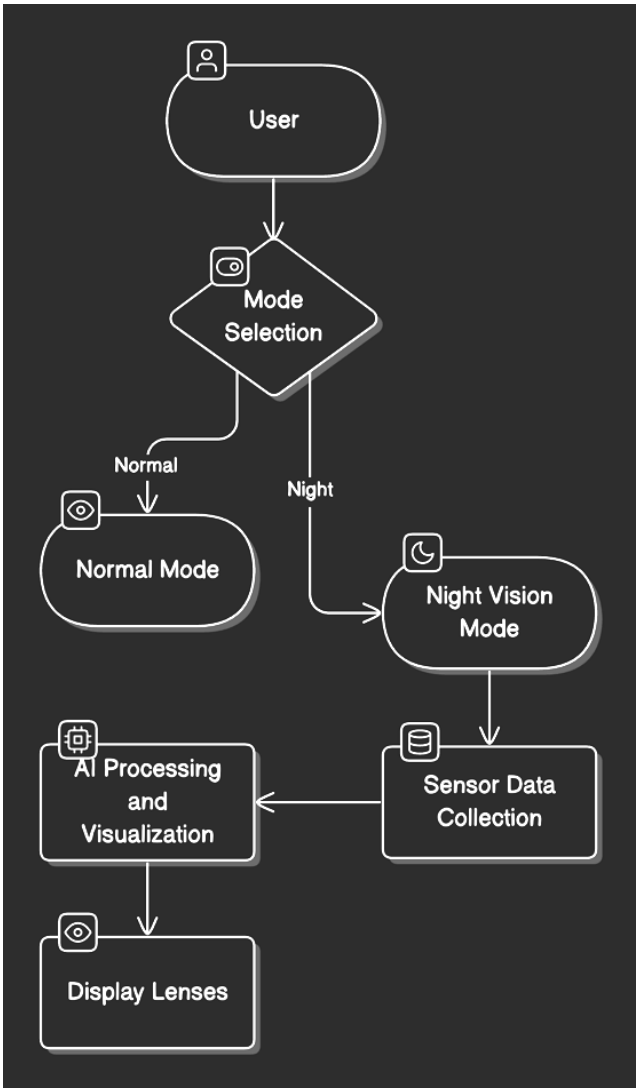


Fig. 4. Conceptual operational flow diagram.

The system architecture diagram in Fig. 5 presents a high-level overview of the key hardware and functional components of the Smart Night-Vision Glasses. The device consists of multiple sensors, including LiDAR, infrared, and ultrasonic sensors, which capture environmental data in low-light conditions. This data is transmitted to a Microcontroller, which serves as the central processing hub, relaying raw information to the AI Processing Unit for real-time interpretation and visualization. The processed data is then displayed on transparent Lenses of the glasses, allowing users to perceive a structured representation of their surroundings. A Power Supply Unit ensures continuous operation, while User Controls facilitate switching between normal and night vision modes. This modular architecture ensures a seamless and efficient user experience while maintaining a compact and lightweight form factor.

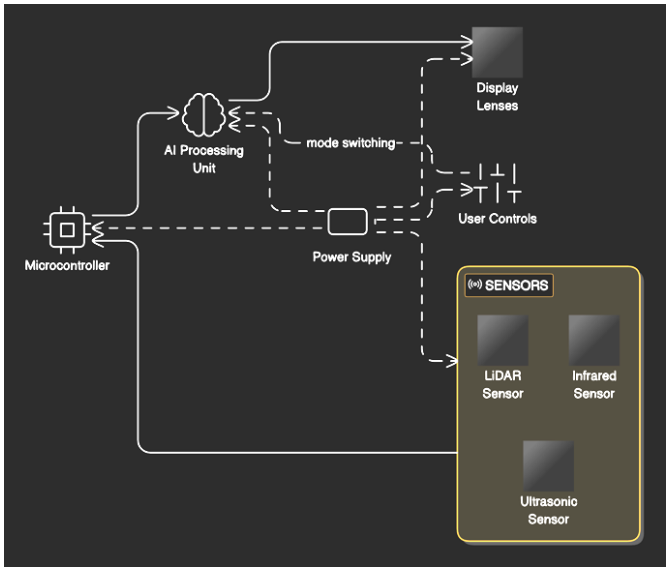


Fig. 5. Conceptual system architecture diagram.

These conceptual visual representations provide a comprehensive overview of the system's functionality, technical integration, and user interaction, reinforcing the feasibility of the proposed concept while acknowledging the need for future research and development.

E. Usability and Experience

A key design consideration for the Smart Night-Vision Glasses is their ease of use and seamless integration into daily life. The glasses are conceptualized to closely resemble standard prescription eyewear, ensuring that users feel comfortable and confident wearing them in any social or professional setting. The frame is designed to be lightweight yet durable, allowing for extended use without discomfort or fatigue.

A critical feature that enhances usability is the mode-switching mechanism, which enables users to seamlessly transition between normal vision and AI-enhanced night vision. This transition can be facilitated by an ergonomically positioned button or touch-sensitive trigger embedded in the frame. With a simple press, the device shifts from standard eyewear mode to night vision mode, activating the sensor suite and AI processing system. This instantaneous switch ensures adaptability to various lighting environments, such as moving from a well-lit indoor space to a completely dark outdoor setting.

In normal mode, the glasses are to function as standard eyewear, which can accommodate prescription lenses if required. The transparent display remains inactive in this mode, ensuring that the glasses look and feel like conventional spectacles. In night vision mode, the LiDAR, infrared, and ultrasonic sensors begin scanning the surroundings, while the AI system processes this data to generate an outlined, cartoon-like representation of objects and surfaces. This visual output can then be projected onto the lenses in real-time, providing users with an intuitive spatial awareness of their environment.

To enhance accessibility, the glasses could integrate haptic feedback mechanisms to alert users of mode changes or

provide additional contextual awareness about their surroundings. For instance, subtle vibrations in the frame could indicate the presence of nearby moving objects or sudden changes in terrain. Additionally, users could have the ability to fine-tune the displayed visuals through adjustable settings, such as contrast levels, edge thickness, or depth emphasis, ensuring an optimized and personalized viewing experience.

Simplicity and user-friendliness remain central to the design philosophy. The system can be envisioned to operate autonomously, without requiring manual calibration or technical adjustments. The goal will be to create a device that enhances mobility and independence for individuals with night blindness or RP, enabling them to navigate their surroundings with confidence and ease.

F. User Testing and Feedback Mechanisms

While the technological advancements behind the Smart Night-Vision Glasses are crucial, their real-world effectiveness depends on user experience and adaptability. Given the conceptual nature of this design, ensuring that it meets the needs of individuals with night blindness and RP would require iterative testing and feedback-driven refinements. Potential user testing approaches would involve the following stages:

- Simulated Virtual Testing: Before physical prototypes are developed, a virtual simulation environment could be used to evaluate the AI's rendering capabilities. Individuals with visual impairments could interact with the simulated display to provide feedback on contrast levels, object clarity, and usability.
- Controlled User Trials with Functional Prototypes: Once a working prototype is available, selected users could test the glasses in controlled environments such as indoor pathways, urban streets, and dimly lit areas. These trials would focus on ease of navigation, reaction time, and cognitive load assessment.
- Iterative Feedback and Design Optimization: Continuous feedback loops would allow developers to refine edge thickness, object emphasis, and dynamic adjustments based on user preferences. Customization options could be introduced, allowing users to fine-tune aspects like contrast intensity and depth sensitivity.
- Longitudinal Studies on Adaptation and Learning Curve: Since adapting to a new visual representation takes time, long-term studies would track how users adjust to the glasses over weeks or months, identifying patterns in usage and potential areas for enhancement.

Another critical factor is social acceptability. The design must not only be functional but also aesthetically discreet, ensuring that users feel comfortable wearing the glasses in various social settings. Unlike bulky assistive devices that may draw unwanted attention, the Smart Night-Vision Glasses aim to resemble conventional eyewear, reinforcing confidence and normalizing their usage in everyday life.

IV. Potential Applications and Benefits

A. Everyday Life

The Smart Night-Vision Glasses have the potential to transform the daily lives of individuals with night blindness and Retinitis Pigmentosa (RP) by addressing the fundamental challenge of navigating in low-light or completely dark environments. Everyday tasks that most people take for granted such as walking through a dimly lit room, taking an evening stroll, or finding their way through a darkened parking lot can be hazardous and stressful for individuals with these conditions. These glasses could serve as a life-changing solution by enabling users to perceive their surroundings clearly, regardless of ambient lighting.

Consider a real-world scenario where an individual with night blindness needs to walk through a dark alley to reach a bus stop after work. Without assistance, they might struggle to detect obstacles such as curbs, trash bins, or uneven pavement, increasing the risk of falls or injury. With the Smart Night-Vision Glasses, the path ahead would be outlined in a clear, cartoon-like format, making obstacles and surface transitions instantly recognizable. This enhanced spatial awareness could instill confidence and encourage greater independence.

Another case study could involve a person with RP navigating a crowded subway station at night. Since RP often reduces peripheral vision, individuals with the condition may have difficulty detecting people or objects approaching from the sides. The glasses, by outlining and emphasizing moving objects in their field of view, could help them avoid collisions and safely navigate through dense crowds.

Driving at night is another area where these glasses could offer significant benefits. Individuals with night blindness often struggle with reduced contrast sensitivity and difficulty perceiving road markings, pedestrians, and other vehicles. The glasses could enhance nighttime driving safety by clearly outlining road edges, lane markings, and potential hazards, reducing anxiety and improving reaction time. While this would not replace standard vehicle lighting or existing driver-assist technologies, it could provide an additional layer of visual clarity for those with night vision impairments.

Beyond mobility and transportation, the glasses could enhance nighttime social engagement. Attending concerts, dining in dimly lit restaurants, or even engaging in outdoor activities such as hiking or camping often presents challenges for individuals with night blindness. With the aid of the glasses, they could experience a newfound sense of inclusion, participating in nighttime activities with greater ease and confidence.

B. Specialized Applications

Beyond personal use, the Smart Night-Vision Glasses could be invaluable in various specialized fields that require enhanced vision in low-light conditions.

For instance, in firefighting, visibility is often compromised by thick smoke and darkness, making it difficult for first responders to locate victims and navigate burning structures.

The glasses, using LiDAR and infrared sensors, could highlight structural outlines, detect heat-emitting bodies, and improve situational awareness. Firefighters could move more efficiently and safely through hazardous environments, increasing the chances of successful rescues while reducing the risk of injury.

Similarly, in military and law enforcement applications, the glasses could serve as a lightweight, energy-efficient alternative to traditional night-vision goggles. Unlike existing solutions that rely on image intensification technology, the Smart Night-Vision Glasses would generate a high-contrast, simplified outline of the environment, enabling soldiers or officers to identify threats, maneuver through unfamiliar terrain, and conduct surveillance with greater precision.

Search-and-rescue operations, particularly those conducted at night or in disaster-stricken areas, could also benefit from this technology. Rescuers navigating through collapsed buildings or forests at night could detect obstacles, locate missing individuals, and assess environmental hazards more effectively, potentially saving lives.

In addition to emergency response and defense, professionals working in remote or extreme environments such as deep-sea researchers, cave explorers, or Arctic scientists could use the glasses to navigate terrain where traditional lighting solutions are ineffective. The ability to perceive surroundings clearly without reliance on external illumination could enhance safety and efficiency in these challenging environments.

C. Social and Economic Impact

The Smart Night-Vision Glasses could have a profound societal impact, particularly for individuals with night blindness and RP. By enabling safer and more independent mobility, they could drastically improve the quality of life for users and reduce reliance on caregivers or mobility assistance programs.

One of the most significant potential benefits is a reduction in nighttime accidents. Falls, collisions, and other visibility-related incidents could be significantly decreased, leading to fewer emergency room visits and hospitalizations. This reduction in injury rates could, in turn, lower healthcare costs associated with treating fractures, concussions, or other accident-related conditions.

From an economic standpoint, these glasses could help individuals with night blindness and RP maintain employment opportunities that require nighttime mobility. Workers in industries such as transportation, security, and emergency response could remain active in their fields without being restricted by their visual impairments. The ability to work more independently could also lead to increased earnings and reduced dependency on disability benefits, contributing to economic self-sufficiency.

Employers could benefit as well by fostering a more inclusive workforce. With assistive technologies like these glasses, businesses might find it easier to accommodate employees with vision impairments, reducing workplace accessibility barriers and promoting diversity.

Educational institutions could also see advantages. Students with night blindness or RP might struggle with evening classes, fieldwork, or extracurricular activities held in low-light conditions. The Smart Night-Vision Glasses could enable them to participate more fully, ensuring that their academic experience is not limited by their condition.

Overall, the potential applications and benefits of the Smart Night-Vision Glasses extend far beyond personal convenience. Whether empowering individuals with visual impairments, enhancing public safety, supporting professionals in high-risk fields, or reducing healthcare costs, this concept represents a significant step toward greater accessibility and inclusion. By leveraging cutting-edge sensor and AI technologies, this innovation envisions a future where vision impairments no longer dictate the boundaries of human potential.

D. Simulated Visual Examples Demonstrating the Potential Benefits

The following examples highlight how smart glasses with advanced sensor integration can significantly improve visibility in dimly lit environments. These conceptual illustrations showcase enhanced edge detection, offering a clear, outlined view of surroundings to aid individuals with night blindness or low vision.

1) Example 1: Staircase in a Dimly Lit Lounge

Fig. 6 depicts a round staircase located in the lounge of a house. The environment is either dimly lit or captured during nighttime, creating significant challenges for individuals with difficulty perceiving in low-light conditions. In such settings, the edges of the stair steps are nearly indistinguishable, posing a risk of missteps or even falls. Additionally, any objects or hazards on the ground blend into the dimly lit surroundings, further exacerbating the risk of accidents.

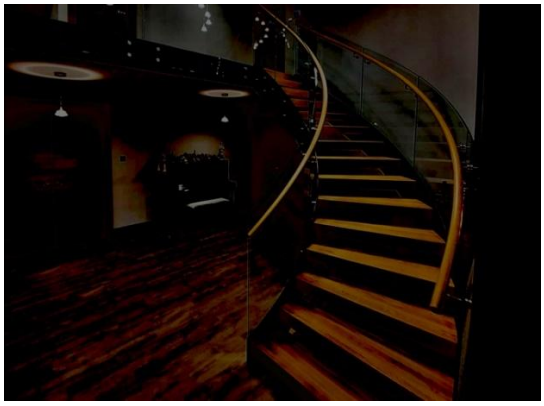


Fig. 6. Staircase in a dimly lit lounge.

Fig. 7 illustrates the same lounge but with a proposed outlined visualization generated by the smart glasses. In this outlined version, the edges of the staircase, furniture, and any other objects in the room are clearly highlighted. The high-contrast outlines allow users to discern each step and object with precision, even in near-complete darkness. This visualization demonstrates how the project's concept could significantly enhance safety and mobility in poorly lit environments, ensuring that users navigate their homes with confidence and ease.

Fig. 7. Conceptual view of the staircase in a dimly lit lounge.

2) Example 2: Sidewalk with Broken Path-Blocks

Fig. 8 captures a section of a sidewalk partially shaded by a tree and an overhanging shop canopy. The broken path-blocks in this area are hard to detect due to the interplay of light and shadows, creating a potential tripping hazard for anyone, especially individuals with reduced night vision or low-light perception.

Fig. 8. Sidewalk with broken path-blocks.

In Fig. 9, the same sidewalk is shown with the concept's outlined visualization. Here, the broken path-blocks, as well as the edges of nearby curbs and other elements, are distinctly outlined, offering a detailed, high-contrast representation of the environment. This enhanced visual information would allow users to easily identify and avoid hazards, improving their safety while walking on shaded or poorly illuminated paths. By providing a clear and real-time depiction of obstacles, the smart glasses concept can empower users to navigate public spaces with greater autonomy and reduced risk of injury. It's something that camera-based solution can't do.

Fig. 9. Conceptual view of the sidewalk with broken path-blocks.

3) Example 3: Dark-Colored Staircase Against a Wall

Fig. 10 shows another staircase situated against a wall within a home. The deep, dark brown color of the staircase steps makes them difficult to visually distinguish from one another, especially under dim lighting conditions. This scenario presents a significant tripping hazard, as users may struggle to gauge the height and depth of each step, increasing the likelihood of accidents.

Fig. 10. Dark-colored staircase against a wall.

Fig. 11 demonstrates the same scene but with the outlined visualization that could be generated by the proposed system. The outlines clearly separate each step and highlight the edges of the surrounding environment, providing a vivid and easily interpretable view of the staircase. This outlined representation ensures that users can confidently ascend or descend the stairs without the fear of losing their footing, even in low-light settings. Such an enhancement highlights the practical benefits of the project, emphasizing how its implementation could transform challenging environments into navigable spaces for individuals with night blindness or similar conditions.

Fig. 11. Conceptual view of the dark-colored staircase against a wall.

These outlined images, while simulated, effectively demonstrate the potential of the proposed solution. By transforming real-world environments into easily distinguishable visual outlines, the concept showcases a promising direction for improving safety and independence for visually impaired individuals.

E. Alignment with SDG Goal 9: Innovation, Industry, and Infrastructure

The concept of Smart Night-Vision Glasses exemplifies a commitment to advancing innovation within the realm of assistive technologies by leveraging cutting-edge developments in IoT, AI, and sensor technology. The concept represents a significant departure from conventional camera-based solutions, adopting a more advanced approach that integrates non-visible, non-harmful sensors such as LiDAR and infrared with sophisticated AI-driven data processing. This combination creates a user-friendly, real-time representation of the environment that is both intuitive and practical, showcasing how emerging technologies can be adapted to meet specific needs in the field of accessibility.

This innovation directly aligns with Sustainable Development Goal (SDG) 9, which emphasizes the role of technology in fostering inclusive and sustainable industrial development. By addressing a critical gap in assistive devices for individuals with night blindness and RP, the glasses contribute to the creation of accessible technologies that empower marginalized groups. Their potential to enhance independence, safety, and mobility underscores the importance of designing solutions that are both innovative and equitable.

Moreover, the proposed system's reliance on efficient, scalable sensor technologies like those used in autonomous vehicles demonstrates its potential for broader applications and integration into existing infrastructure. This versatility positions the concept as a steppingstone toward more comprehensive IoT ecosystems, where assistive devices are seamlessly interconnected with other smart technologies, promoting inclusivity and sustainability.

V. Challenges and Future Directions

The development of Smart Night-Vision Glasses presents a range of challenges spanning technical, user adoption, and economic considerations. Addressing these hurdles will require multidisciplinary advancements in sensor technology, AI optimization, material science, and human-computer interaction.

A. Technological Challenges

One of the primary technical challenges is sensor miniaturization. Embedding LiDAR, infrared, or ultrasonic sensors into a form factor as small and lightweight as regular glasses requires advanced engineering. These sensors must not only be compact but also maintain high accuracy in various environmental conditions, including fog, rain, and extreme temperatures. Current LiDAR and infrared modules used in autonomous vehicles or industrial applications are too large for wearable integration. A potential solution lies in developing micro-electromechanical systems (MEMS)-based LiDAR or solid-state LiDAR, which could offer the same functionality in a smaller footprint.

Another significant challenge is real-time AI processing and latency reduction. The glasses need to process sensor data instantly to generate a clear and intuitive visual representation without noticeable lag. AI models for object detection, edge enhancement, and depth perception must operate at high speed while consuming minimal power. Solutions may involve algorithmic optimizations, such as employing efficient neural network quantization techniques to reduce the computational burden. Additionally, specialized AI hardware accelerators, such as edge TPU (Tensor Processing Unit) chips, could enable efficient on-device processing without excessive energy consumption.

Power efficiency is another critical issue. Continuous operation of sensors and AI computations demands substantial energy, yet integrating large batteries would compromise the device's weight and comfort. Potential solutions include low-power AI chips, optimized power management algorithms, and energy-harvesting technologies (e.g., thermoelectric generators or solar cells embedded in the frame). Additionally, modular battery packs that allow users to swap or recharge batteries externally could provide extended usability.

Finally, data fusion and calibration pose challenges. The glasses rely on multiple sensor inputs to construct a meaningful visual representation, requiring precise sensor synchronization and dynamic calibration algorithms that adapt to environmental conditions in real-time. Advances in sensor fusion algorithms and adaptive AI models will be key in ensuring accurate perception under varying lighting conditions.

B. User Adoption Challenges

Even if the technical barriers are overcome, widespread adoption depends on user comfort, ease of use, and adaptability.

Ergonomic design and comfort are paramount, as users may need to wear the glasses for extended periods. Unlike traditional night-vision goggles, which are bulky and used intermittently, these glasses must be lightweight, aesthetically appealing, and customizable for different face shapes and prescription lens requirements. Research into advanced lightweight materials, such as graphene-reinforced polymers or titanium alloys, could help reduce weight while maintaining durability.

Learning curve and usability present another hurdle. Users with night blindness or RP may need time to adapt to the AI-generated visual output. The glasses must provide an intuitive and natural visual experience that does not overwhelm or confuse the user. A possible solution is an adaptive AI system that gradually customizes the display output based on the user's preferences and past behavior. In addition, incorporating a guided onboarding process (e.g., interactive tutorials or gradual exposure modes) could ease the transition for new users.

Additionally, public perception and stigma surrounding assistive technologies may impact adoption. If the glasses appear too conspicuous or resemble medical devices, some users may feel self-conscious wearing them in social settings. A sleek, inconspicuous design that mimics standard eyewear will be crucial in normalizing their use and encouraging widespread adoption.

C. Economic and Market Challenges

The affordability of the glasses is a significant factor in market viability. Advanced sensors and AI hardware are expensive, and integrating them into a consumer-grade wearable device may result in high manufacturing costs.

Cost-effective production methods will be essential to making the glasses accessible to a broad audience. Potential strategies include:

- Scaling manufacturing through mass production, which could drive down costs.
- Exploring alternative materials that offer a balance between affordability and performance.
- Leveraging modular hardware designs to allow users to upgrade specific components without purchasing a completely new device.

Market positioning and funding also present challenges. Unlike traditional night-vision goggles, which are typically aimed at military or industrial markets, these glasses target individuals with night blindness and RP. This is a relatively niche consumer base, meaning financial incentives such as insurance coverage, government subsidies, or assistive technology grants may be necessary to ensure widespread adoption. Partnering with healthcare providers, vision impairment advocacy groups, and public health agencies could facilitate market entry.

D. Future Research

To transform this concept into a viable product, a structured research and development (R&D) roadmap is necessary. The following timeline outlines key phases of development:

1) Short-Term (0–2 Years): Foundational Research and Prototyping:

- Conduct feasibility studies on sensor integration and power-efficient AI processing.
- Develop early-stage prototypes focusing on real-time data visualization and sensor fusion.
- Collaborate with materials scientists to design a lightweight, ergonomic frame.
- Begin small-scale user testing to gather feedback on usability and comfort.

2) Mid-Term (3–5 Years): Iterative Development and Field Testing:

- Optimize miniaturized LiDAR/infrared technology for compact, consumer-friendly use.
- Enhance AI models for low-latency, edge-device processing.
- Implement personalized AI adaptation, allowing the glasses to tailor their visual output to individual users.
- Expand user trials, partnering with vision impairment organizations to refine functionality.
- Conduct regulatory assessments and seek approval from assistive technology regulatory bodies.

3) Long-Term (6–10 Years): Commercialization and Widespread Deployment:

- Scale up manufacturing processes for mass production.
- Secure insurance coverage and reimbursement options for affordability.
- Expand the market by integrating cross-platform compatibility (e.g., optional smartphone connectivity for advanced settings).
- Continue AI refinement through real-world data collection to enhance adaptability and accuracy.

Beyond this timeline, future iterations could explore additional innovations such as haptic feedback integration, augmented reality (AR) overlays, or brain-computer interface advancements to further improve accessibility and user experience.

The development of Smart Night-Vision Glasses presents a compelling opportunity to enhance the mobility, safety, and quality of life for individuals with night blindness and RP. However, significant challenges must be addressed, including sensor miniaturization, real-time AI processing, and affordability. Through an iterative R&D approach, leveraging advancements in wearable technology, machine learning, and human-centered design, these challenges can be systematically tackled.

By outlining a structured roadmap for future research and development, this concept moves beyond theoretical discussion into the realm of practical feasibility. If successfully developed, Smart Night-Vision Glasses could redefine accessibility, enabling individuals with vision impairments to navigate the world with confidence and independence.

VI. Conclusion

The Smart Night-Vision Glasses concept represents a transformative leap in assistive technology, offering a non-invasive, AI-driven solution for individuals with night blindness and Retinitis Pigmentosa (RP). By integrating advanced sensor technologies such as LiDAR and infrared, the design proposes an innovative method for enhancing vision in

low-light and dark environments—one that does not rely on traditional camera-based systems or visible light sources. This unique approach not only prioritizes user privacy and security but also ensures optimal functionality in all lighting conditions without being affected by glare, reflections, or light pollution.

A. Key Contributions of this Work

- Conceptualization of a novel assistive device that leverages sensor-based depth perception rather than camera-based vision, addressing privacy concerns and low-light visibility challenges.
- Integration of LiDAR, infrared, and ultrasonic sensing technologies to create real-time, AI-enhanced visual representations of the environment, improving spatial awareness for individuals with night blindness and RP.
- Development of an AI-driven visualization framework that simplifies complex sensory input into clear, user-friendly imagery, allowing for intuitive navigation.
- Exploration of power-efficient, compact hardware designs suitable for a wearable, lightweight form factor, ensuring comfort and prolonged usability.
- Discussion of technical, user adoption, and market challenges, offering practical solutions for sensor miniaturization, AI processing, and affordability to facilitate real-world implementation.
- Proposal of a structured research and development roadmap, outlining future directions in prototyping, AI optimization, field testing, and large-scale deployment over the next decade.

B. Call to Action

While this paper presents a conceptual framework rather than a functional prototype, it lays the foundation for an entirely new class of sensor-based assistive eyewear. To bring this vision to reality, collaboration across multiple disciplines is essential. Researchers, engineers, AI specialists, medical professionals, and assistive technology advocates are encouraged to explore the feasibility of this design, contribute to prototype development, and refine the system through user testing.

Furthermore, industry stakeholders, wearable tech manufacturers, healthcare institutions, and policymakers should consider funding and supporting research into next-generation assistive devices. Addressing the mobility challenges of individuals with vision impairments is not just a technological endeavor; it is a step toward fostering greater accessibility, inclusivity, and independence for millions worldwide.

By advancing this research, we move closer to a future where vision impairments no longer dictate limitations, and individuals with night blindness or RP can navigate their world with confidence, safety, and autonomy.

REFERENCES

[1] Kamal, S. A. A Design of AI-based-Smart Glasses, which Offer Navigation in Addition to Correcting Vision. In Proceedings of the International Conference on Engineering, Natural Sciences and Technological Developments(ICENSTED 2024) (pp. 39-5).

[2] Tarik, H., Hassan, S., Naqvi, R. A., Rubab, S., Tariq, U., Hamdi, M., ... & Cha, J. H. (2023). Empowering and conquering infirmity of visually impaired using AI-technology equipped with object detection and real-time voice feedback system in healthcare application. CAAI Transactions on Intelligence Technology.

[3] Gugulothu, S. (2024). Functional Analysis and statistical Mechanics for exploring the Potential of Smart Glasses: An Assessment of Visually Impaired Individuals. Communications on Applied Nonlinear Analysis, 31(2s), 402-421.

[4] Badawi, M., Al Nagar, A. N., Mansour, R., Mansour, R., Ibrahim, K., Ibrahim, K., ... & Elaskary, S. (2024). Smart Bionic Vision: An Assistive Device System for the Vis-ually Impaired Using Artificial Intelligence. International Journal of Telecommunications, 4(01), 1-12.

[5] Lavric, A., Beguni, C., Zadobrischi, E., Căilean, A. M., & Avătămăniței, S. A. (2024). A comprehensive survey on emerging assistive technologies for visually impaired persons: lighting the path with visible light communications and artificial intelligence innovations. Sensors, 24(15), 4834.

[6] Gollagi, S. G., Bamane, K. D., Patil, D. M., Ankali, S. B., & Akiwate, B. M. (2023). An innovative smart glass for blind people using artificial intelligence. Indonesian Journal of Electrical Engineering and Computer Science, 31(1), 433-439.

[7] Pydala, B., Kumar, T. P., & Baseer, K. K. (2023). Smart_Eye: a navigation and obstacle detection for visually impaired people through smart app. Journal of Applied Engineering and Technological Science (JAETS), 4(2), 992-1011.

[8] Wang, J., Wang, S., & Zhang, Y. (2023). Artificial intelligence for visually impaired. Displays, 77, 102391.

[9] Bhabad, D., Kadam, S., Malode, T., Shinde, G., & Bage, D. (2023). Object detection for night vision using deep learning algorithms. International Journal of Computer Trends and Technology, 71(2), 87-92.

[10] Wan, W., Sun, W., Zeng, Q., Pan, L., & Xu, J. (2024, January). Progress in artificial intelligence applications based on the combination of self-driven sensors and deep learning. In 2024 4th International Conference on Consumer Electronics and Computer Engineering (ICCECE) (pp. 279-284). IEEE.

[11] Biswas, A., & Wang, H. C. (2023). Autonomous vehicles enabled by the integration of IoT, edge intelligence, 5G, and blockchain. Sensors, 23(4), 1963.

[12] Tan, K., Wu, J., Zhou, H., Wang, Y., & Chen, J. (2024). Integrating Advanced Computer Vision and AI Algorithms for Autonomous Driving Systems. Journal of Theory and Practice of Engineering Science, 4(01), 41-48.

[13] Leong, W. Y. (2023, August). Digital technology for ASEAN energy. In 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT) (pp. 1480-1486). IEEE.

Comparative Analysis of Cardiac Disease Classification Using a Deep Learning Model Embedded with a Bio-Inspired Algorithm

Nandakumar Pandiyan[1], Subhashini Narayan[2]*
Research Scholar, School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, India[1]
Associate Professor, School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, India[2]*

***Abstract*—Cardiac disease classification is a crucial task in healthcare aimed at early diagnosis and prevention of cardiovascular complications. Traditional methods such as machine learning models often face challenges in handling high-dimensional and noisy datasets, as well as in optimizing model performance. In this study, we propose and compare a novel approach for heart disease prediction using deep learning models embedded in bioinspired algorithms. The integration of deep learning techniques allows for automatic feature learning and complex pattern recognition from raw data, while bioinspired algorithms provide optimization capabilities for enhancing model accuracy and generalization. Specifically, the cuckoo search algorithm and elephant herding optimization algorithm are employed to optimize the architecture and hyperparameters of deep learning models, facilitating the exploration of diverse model configurations and parameter settings. This hybrid approach enables the development of highly effective predictive models by efficiently leveraging the complementary strengths of deep learning and bioinspired optimization. Experimental results on benchmark heart disease datasets demonstrate the superior performance of the proposed method compared to conventional approaches, achieving higher accuracy and robustness in predicting heart disease risk. The proposed framework holds significant promise for advancing the state-of-the-art in heart disease prediction and facilitating personalized healthcare interventions for at-risk individuals.**

***Keywords*—Cardiac disease; heart disease; bio-inspired; machine learning; deep learning; prediction; classification**

I. Introduction

Over the years, people all around the world have fought many terrible illnesses, with heart disease [1] receiving the most attention and is widely recognized in medical studies. It's a well-known illness that strikes a lot of people in their middle or advanced years and can sometimes result in life-threatening complications. Men are more prone to heart disease than women are. A cardiac disease diagnosis is crucial to the prognostication process. Reducing health risks and averting cardiac arrests depends heavily on the early detection of heart disease. According to the World Health Organization (WHO) [2], 17.9 million people worldwide pass away from cardiovascular disorders each year, making it the top cause of mortality worldwide. One significant risk faced by healthcare institutions, including hospitals and clinics, is the provision of quality services at reasonable prices.

Medical professionals are currently using data mining and analytical modeling to learn about various unknown diseases in the future. The key challenges for deep learning include data collection, efficient and interpretable work, integration with other traditional approaches, maintaining imbalanced labels with multimodal data, and the release of new models with other multidisciplinary studies. In the future, this kind of work can be achieved by using the deep learning approach to either feature extraction, classification, or a combination of the two.

A subfield of machine learning known as "Deep Learning" [3] gained prominence in the world, particularly for the classification of diseases. The development of deep learning models has shown promising results in the classification and diagnosis of various cardiac conditions, including arrhythmia, myocardial infarction, and coronary artery disease. Convolutional neural networks (CNNs) [4], which are significant and potent deep learning models, can be improved with additional layers to create neural architecture designed especially for medical applications. The deep learning model can be further enhanced and we can obtain additional network performance in terms of learning rate and dropout rate by using novel data pre-processing, augmentation techniques, and hyperparameter optimization. These are the essential components of effective deep learning techniques in healthcare. The proposed concept focuses on evaluating how different deep-learning models predict and classify cardiac conditions. Several outcomes in the cardiac disease prediction process were discussed and examined in this article, along with their performance using different deep learning techniques. However, the performance of these models can be further enhanced through the integration of bio-inspired algorithms, which can optimize the model's architecture and hyperparameters, leading to improved accuracy and efficiency. This research paper aims to conduct a comparative analysis of cardiac disease classification using deep learning models embedded with bio-inspired algorithms.

The motivation behind this research lies in addressing key challenges in cardiac disease classification using AI-driven techniques. The conventional machine learning and deep learning approaches, while powerful, often require extensive hyperparameter tuning, suffer from model generalization issues, and may lack robustness when applied to diverse datasets. Moreover, high-dimensional medical data, such as ECG signals and cardiac imaging, introduce complexity in feature extraction and classification. Bio-inspired algorithms, including genetic

algorithms (GA), particle swarm optimization (PSO), ant colony optimization (ACO), and artificial immune systems (AIS), mimic natural evolutionary and adaptive processes to optimize model parameters, improve feature selection, and enhance classification accuracy. By embedding a bio-inspired optimization technique within a deep learning framework, the proposed approach aims to enhance model performance, improve computational efficiency, increase generalization ability, enable automated feature selection, and reduce dependence on large labeled datasets.

The integration of a bio-inspired optimization algorithm with deep learning for cardiac disease classification presents several benefits, including higher classification accuracy, efficient feature engineering, scalability and adaptability, reduction in false positives/negatives, and enhanced decision support in clinical settings.

The following is the proposed research work: First, in Section II, the study will review the current state of deep learning techniques in the field of cardiovascular disease diagnosis, highlighting the challenges and limitations of existing approaches. The review will also explore the potential of bio-inspired algorithms to enhance the performance of deep learning models. Next, in Section III, the paper will present the development of a novel deep-learning model integrated with a bio-inspired algorithm for the classification of cardiac diseases. The model will be trained and evaluated on a comprehensive dataset of electrocardiogram signals and other relevant clinical data. The comparative analysis will be conducted by evaluating the performance of the proposed model against other state-of-the-art deep learning architectures, both with and without the integration of bio-inspired algorithms is discussed in Section IV. The difficulties and potential future paths of the suggested effort are covered in Section V. The comparative analysis work is concluded in Section VI with recommendations for further research and use in real-time applications.

II. Literature Review

A review of the literature on machine learning-based heart disease prediction indicates that there is increasing interest in utilizing cutting-edge computational methods to improve the precision and effectiveness of heart disease detection and prediction. Here is a summarized overview of key findings and trends:

Electronic health records (EHRs) [5], clinical databases [6], and public repositories [7] are among the most often used datasets in research. Large-scale datasets make it easier to create solid machine-learning models. Feature selection and extraction strategies [8] are critical for improving model performance while lowering dimensionality. Demographic information, clinical characteristics, and medical history are among the most common components. Various machine learning algorithms are used, ranging from basic methods like logistic regression to more complex techniques like decision trees, support vector machines, and ensemble methods. Deep learning models, particularly neural networks, are becoming popular due to their capacity to automatically learn hierarchical representations. Model performance is often assessed using metrics such as accuracy, sensitivity, specificity, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC). Cross-validation on independent datasets is a standard way to check generalization performance.

Ensemble approaches [9], such as random forests and gradient boosting, are commonly used to increase forecast accuracy and handle imbalanced datasets. Addressing class imbalance in heart disease datasets is a recognized concern, and researchers use strategies such as oversampling, undersampling, and cost-sensitive learning to address it. Some research emphasizes the necessity of using domain knowledge and clinical competence when selecting features, developing models, and interpreting results. With the introduction of wearable technologies and continuous health monitoring, there is a growing emphasis on building models for real-time cardiovascular disease prediction and monitoring [10]. As machine learning models are used in clinical settings, there is a greater emphasis on making algorithms interpretable and offering reasons for predictions, which improves confidence and acceptability among healthcare practitioners.

While machine learning and deep learning approaches have shown promise in heart disease prediction, each has its own set of demerits. The following enumerates the disadvantages of utilizing conventional machine learning techniques for heart disease prediction in contrast to deep learning:

Traditional machine learning frequently relies on human feature engineering, which necessitates domain knowledge to choose appropriate features. This can be time-consuming and may overlook complex patterns in the data. Machine learning methods may struggle to automatically learn complicated hierarchical characteristics from raw data, perhaps missing subtle patterns critical to good heart disease prediction. Scalability issues may arise for traditional machine learning methods when working with big and high-dimensional datasets. This shortcoming may impair their ability to manage the huge amount of heterogeneous data available in healthcare settings. Machine learning methods may struggle to detect nonlinear correlations between features in data. Heart disease prediction frequently involves complicated relationships, which typical machine-learning methods may fail to capture adequately. Traditional machine learning algorithms may fail to appropriately capture the temporal features of heart disease progression. Sequential patterns and time-dependent interactions may be required for reliable forecasts, although they are frequently difficult for these models.

Deep learning models, with their several layers, are excellent at automatically learning hierarchical characteristics from data. This enables them to record complicated linkages and representations, which may lead to better prediction performance. Deep learning models frequently require large amounts of labeled data for training, and performance may degrade if data is constrained. In contrast, traditional machine learning models can sometimes outperform smaller datasets. Deep learning models are computationally expensive, necessitating sophisticated hardware and substantial computational resources. This intricacy can be an impediment, particularly in resource-constrained contexts. Deep learning models are typically regarded as "black boxes," making it difficult to explain their decisions. In healthcare, interpretability is critical for building confidence with clinicians and patients,

and older machine learning models may be more interpretable. Deep learning models, particularly those with many parameters, are susceptible to overfitting, especially when training data is restricted. Regularization techniques are necessary, although they may not eliminate the risk.

The literature survey highlights the promising role of machine learning in heart disease prediction, with a focus on model performance, interpretability, and real-time monitoring. In summary, while deep learning has shown advantages in automatically learning complex features, it comes with challenges such as interpretability issues, computational complexity, and the need for large amounts of labeled data. Traditional machine learning methods may have limitations in capturing intricate patterns but offer advantages in terms of interpretability and scalability with smaller datasets. The healthcare application's criteria for interpretability, the resources at hand, and the characteristics of the dataset will determine which of the two techniques is best. Continued advancements in this field have the potential to significantly impact cardiovascular healthcare. This resulted in the addition of a bioinspired metaheuristic algorithm that powers the suggested work's feature selection and optimization processes. Bio-inspired algorithms [11], such as genetic algorithm, cuckoo search algorithm, particle swarm optimization, elephant herding optimization, simulated annealing, and ant colony optimization, among others, are computational techniques that mimic the behavior of natural systems or processes. These algorithms have been increasingly applied in various fields, including optimization problems in healthcare, such as predicting heart diseases. Here's how bio-inspired algorithms can help optimize dataset features for better accuracy in predicting heart diseases:

One of the key steps in building predictive models is selecting the most relevant features from the dataset. Bio-inspired algorithms can help identify the most informative features by mimicking natural selection processes. To evolve a population of possible feature subsets, for instance, one can utilize genetic algorithms [12], where each subset is a potential solution to the feature selection problem. The algorithm iteratively evaluates and evolves these subsets based on their performance in predicting heart diseases. High-dimensional datasets can lead to overfitting and increased computational complexity. Bio-inspired algorithms can be used for dimensionality reduction by finding a low-dimensional representation of the data that preserves its essential characteristics. Techniques like particle swarm optimization or ant colony optimization can optimize the selection of features or combinations of features that best represent the dataset while minimizing redundancy and noise. Bio-inspired algorithms can also be employed to optimize the parameters of machine learning models used for predicting heart diseases. For instance, genetic algorithms can search the parameter space of complex models such as neural networks or support vector machines to find the combination of parameters that maximizes prediction accuracy. Bio-inspired algorithms can facilitate the creation of diverse ensemble models [13] for predicting heart diseases. These algorithms can be used to generate multiple models with different subsets of features or different parameter settings. Ensemble approaches frequently outperform individual models in terms of prediction accuracy by aggregating the predictions of these several models.

Bio-inspired algorithms can automate the process of hyperparameter tuning, which involves selecting the optimal values for parameters that control the learning process of machine learning algorithms. By exploring the hyperparameter space efficiently, these algorithms can help optimize the performance of predictive models for heart disease prediction. Overall, bio-inspired algorithms provide powerful tools for optimizing dataset features and improving the accuracy of predictive models for heart diseases. By leveraging principles from nature, these algorithms can efficiently explore complex solution spaces and identify high-quality solutions that lead to more accurate predictions. The proposed model for cardiac disease prediction will be analyzed and compared with the deep learning model incorporating a bio-inspired algorithm in the following section.

III. Proposed Model of Cardiac Disease Prediction

Heart disease is one of the most significant ailments that must be treated early on. Humans are encouraged to investigate the AI-driven industry by advances in cardiac and medical technology, which are supported by deep learning models. Based on the statistics presented in the study, deep learning has become the new preferred option for researchers. Researchers are encouraged to explore novel ways of classifying heart diseases using deep learning models. The approach for diagnosing heart disease is based on numerous well-known datasets, including the Cardiovascular Health Study (CHS), Framingham Heart Study, Statlog Heart Disease, and UCI (Cleveland, Hungary, Switzerland). According to this study [14], bio-inspired algorithms can improve the accuracy of deep learning models that include heart disease. Bioinspired computing has the potential to grow and spread throughout several research communities. Future computing generations will surely be greatly impacted by the algorithms covered in the proposed work. It has been demonstrated that biocomputing offers a very natural answer to the issue that many wild animals and humans face. Fig. 1 shows the schematic diagram for the proposed task. It has four layers. The first layer analyses data through collection, preprocessing, and feature selection. The second layer consists of implementing a model, such as a deep learning model, training the model, and optimizing it using the bio-inspired algorithm. The third layer uses metrics including accuracy, precision, recall, and F1-score to assess performance through validation and testing. The fourth layer connects the generated results to the human interaction model, healthcare support systems, and user interfaces.

A. Comparative Analysis of Proposed Work

In this article, we examine two results for heart illness prediction. The first paper discusses how deep belief networks can embed the cuckoo search algorithm, while the second paper discusses how to use a convolutional neural network with the inception-resnet-v2 model. In the first work, the authors [19] employed the feature selection approach, which was followed by a hamming distance-based data-cleaning procedure. After being collected and evaluated, datasets on cardiac disorders are delivered to the cleaning process. The hamming distance feature

selection strategy is used during the data pre-processing stage to manage missing values and eliminate falsified, and unrelated features. Hamming distance is also used to calculate the distance between features from the heart disease dataset. Cross-validating the dataset using test and training data is the next stage. Machine learning methods for classification including logistic regression, Naïve Bayes, decision trees, and support vector machines are used to process the training data. A bioinspired cuckoo search technique combined with an optimized deep learning model classification like deep belief networks handles the testing set. Cuckoo search [21] refers to a significant population-based optimization technique developed by Xin-she Yang et al. in 2009. It works by permeating their oocytes inside the protection of various nourishing birds. Breeding behavior inspires cuckoo search, which is used to tackle a variety of optimization challenges. A cuckoo search is another nature-inspired method that is commonly used to solve various optimization problems across several engineering fields. Larochelle et al. developed deep belief networks [22] as probabilistic generative models acquired by stacked restricted Boltzmann machines in 2007. These networks offer an alternative to the discriminative character of classic neural networks. Two key features of deep belief networks are their ability to encode higher-request network topologies and their quick induction. It uses two probabilities to calculate yields autonomously. They have both coordinated and undirected layers and are made up of double inert components. Unlike previous models, deep belief networks learn all information at each layer. It can be used to record signal data, process images, create video sequels, identify clusters, and train nonlinear autoencoders. Fig. 2 depicts cardiac illness prediction utilizing deep belief networks and embedded cuckoo search.

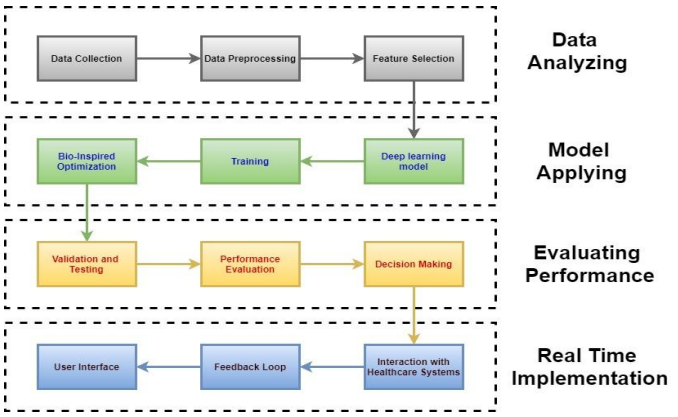


Fig. 1. The schematic diagram of the proposed work.

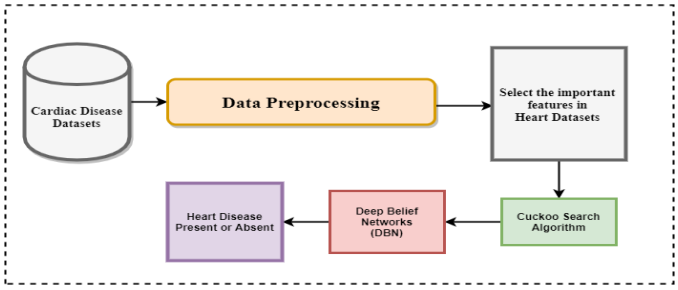


Fig. 2. Cuckoo search algorithm and deep belief networks for the identification of cardiac illness.

The mathematical model of deep belief networks is as follows:

A DBN with l hidden layers contains l weight matrices: $W^{(1)}, \ldots, W^{(l)}$

It also contains $l+1$ bias vectors: $b^{(0)}, \ldots, b^{(l)}$ where $b^{(0)}$ provide the biases for the visible layer.

The probability distribution for DBN is given by,

$$P(h^{(l)}, h^{(l-1)}) \propto exp(b^{(l)^T} h^{(l)} + b^{(l-1)^T} h^{(l-1)} + h^{(l-1)^T} W^{(l)} h^{(l)}) \quad (1)$$

$$P\left(h_i^{(k)} = 1 \middle| h^{(k+1)}\right) = \sigma\left(b_i^{(k)} + W_{:,i}^{(k+1)^T} h^{(k+1)}\right) \quad (2)$$

Where $\forall i, \forall k \in 1, \ldots, l-2$

$$P(v_i = 1 \mid h^{(1)}) = \sigma(b_i^{(0)} + W_{:,i}^{(1)^T} h^{(1)}) \quad \forall i. \quad (3)$$

In the case of real-valued visible units, substitute

$$v \sim N(b^{(0)} + W^{(1)^T} h^{(l)} \beta^{-1}), \quad (4)$$

With β diagonal for tractability $\sigma(x) = 1/(1+exp(-x))$

The weights from the trained DBN can be used as the initialized weights of a DNN,

$$h^{(1)} = \sigma(b^{(1)} + v^T W^{(1)}), \quad (5)$$

$$h^{(l)} = \sigma(b_i^{(l)} + h^{(l-1)^T} W^{(l)}), \forall l \in 2, \ldots, m \quad (6)$$

also, then, at that point, the entirety of the loads is tweaked by applying backpropagation or other discriminative models to enhance the efficiency of the entire network.

The mathematical models of the cuckoo search algorithm are as follows:

When generating new solutions $x^{(t+1)}$ for, say, a cuckoo i, a Lévy flight is defined by equation (7) as follows,

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \oplus Levy(\lambda) \quad (7)$$

Where $x_i^{(t)}$ is the current location of the cuckoo, α is a step size and positive constant tuned according to the dimensions of the search space, $\oplus$ is the entry-wise multiplication and λ is the levy exponent.

The above equation is essentially the stochastic equation for a random walk. The random step length is drawn from a Lévy distribution defined in equation (8) as follows,

$$Levy(\lambda) \sim u = t^{-\lambda}, (1 < \lambda \leq 3) \quad (8)$$

Where λ is the levy exponent that defines the decay of the probability density function (PDF) with t. In most cases, $\alpha = 1$ and $\lambda = 1.5$.

Lévy-flight has the unique virtue of increasing population variety in stages, allowing the algorithm to efficiently exit the local optimum. The following equation (9) is calculated as Levy random numbers,

$$Levy\ (\lambda) \sim \frac{\phi \times u}{|v|^{1/\lambda}} \quad (9)$$

Where u and v are both standard normal distributions, then ϕ is defined in equation (10) as follows:

$$\phi = \left[\frac{\gamma\,(1+\lambda) \times sin\,(\pi \times \lambda/2)}{\gamma\,(((1+\lambda)/2) \times \lambda \times 2^{(\lambda-1)/2})}\right]^{1/\lambda} \quad (10)$$

where, γ is a standard gamma function and ϕ is a random angle used for convergence.

The suggested method outperforms alternative deep learning and machine learning models. Table I presents the latest research conducted by multiple authors on the use of a deep learning model integrated with bio-inspired algorithms to predict heart illness. The heart disease review study analyses the accuracy rates of numerous classification strategies used to detect or forecast it using deep learning models.

TABLE I. ANALYZING AND CONTRASTING DIFFERENT DEEP LEARNING MODELS FOR PREDICTING CARDIAC ILLNESS WITH DATASET INFORMATION

Author Name	Year	Models	Dataset	Accuracy
Girish S. Bhavekar, Agam Das Goswami [15]	2022	RNN, LSTM	UCI Cleveland	95.10%
A Bhardwaj et al., [16]	2023	DCNN	Physionet PCG data	93.07%
AL Golande, T Pavankumar [17]	2023	CNN, LSTM	PTB Diagnostic ECG Data	95.89%
Yunqing Liu et al., [18]	2023	EfficientNet-based network	ECG dataset	73.33%
Nandakumar P, Subhashini R [19]	2022	DBN-CSA	UCI Statlog	91.26%
Nandakumar P, Subhashini R [20]	2024	CNN-Inception-ResNet-v2-EHO	UCI Cleveland	98.77%

In this study [19], the authors employed Euclidean distance to preprocess data. This method is used to sanitize data. For feature selection, a metaheuristic technique such as elephant herding optimization is used. After selecting the best features, a convolutional neural network with the Inception-ResNet-v2 model is used to classify the output based on the chosen features. Elephant herding optimization [23] was used to choose features. Feature selection is viewed as a pre-processing stage in machine learning. One of the most difficult tasks is determining which feature subset in a large or complex dataset is the most relevant. It is now crucial to find important information or hidden patterns in vast amounts of data. It has been demonstrated that feature selection effectively eliminates superfluous features. It can also cut computing costs, increase storage capacity, and improve classifier performance. A convolutional neural network called Inception-ResNet-v2 [24] combines the Inception structure with Residual connections. Multiple convolutional filters of different sizes are combined with residual connections in the Inception-ResNet block. In addition to avoiding the degradation issue with deep structures, adding residual connections speeds up training. Using an Inception-ResNet-v2 model, this study classified the UCI heart disease dataset into two categories: present and missing, as shown in Fig. 3.

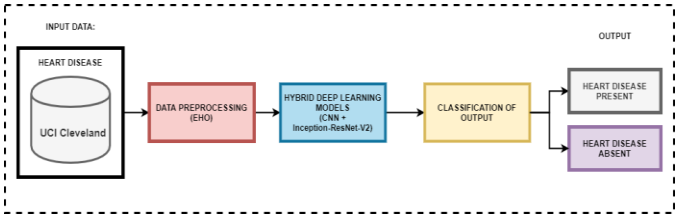


Fig. 3. The combined deep learning model for predicting heart disease.

The mathematical models of EHO begin with two basic notions, then introduce the velocity and separation strategies, and finally apply the elitism strategy to the entire algorithm.

As a result, for the elephant j in clan ci, the position can be updated as:

$$\mathrm{X_{new,ci,j}} = \mathrm{X_{ci,j}} + \alpha \times \left(\mathrm{X_{best,ci}} - \mathrm{X_{ci,j}}\right) \times \mathrm{r} \quad (11)$$

Where $X_{new,ci,j}$ and $X_{ci,j}$ are new and old positions for elephant j in clan ci, respectively. $\alpha \in [0,1]$ is a scale factor. $X_{best,ci}$ represents the best position in clan ci. $r \in [0,1]$ is a commonly distributed random number.

In Eq. (11), the matriarch's position has not changed. It can be updated as follows for the fittest individual:

$$\mathrm{X_{new,ci,j}} = \beta \times \mathrm{X_{center,ci}}, \quad (12)$$

$$\mathrm{X_{center,ci}} = \frac{1}{\mathrm{n_{ci}}} \times \sum_{\mathrm{j=1}}^{\mathrm{n_{ci}}} \mathrm{X_{ci,j}}, \quad (13)$$

where, $\beta \in [0,1]$ is a scale factor. $X_{center,ci}$ is the middle place in clan ci. n_{ci} is the number of elephants in clan ci. The positions of all the clan members may be observed to be modified by the matriarch.

Male elephants live alone after being separated from their group. The EHO method is designed to replace the worst elephant in each clan with a separation operator. The procedure can be shown in Eq. (14):

$$\mathrm{X_{worst,ci}} = \mathrm{X_{min}} + (\mathrm{X_{max}} - \mathrm{X_{min}} + 1) \times \mathrm{r}, \quad (14)$$

where, $X_{worst,ci}$ symbolizes the lousiest elephant in clan ci. The elephant position's upper and lower boundaries are X_{max} and X_{min}, respectively.

In this work, the input consists of one-dimensional data with $x = (x_1, x_2, x_3, \ldots\ldots, x_{n-1}, x_n, clabel)$ where $x_n \in R^d$ denotes the heart disease features and $clabel \in R$ denotes a class label used for the output of either heart disease present or absent. Conv1D is used to generate a feature map Fm. The convolution operation is then applied to the heart disease input data with filtering of $w \in R^{\mathrm{F}d}$ where F represents the intrinsic properties of the input data that will produce the final output after feeding it into the next input block.

From the set of features, the new feature map Fm is obtained as follows:

$$h\ell_i^{\mathrm{F}m} = tanh(w^{\mathrm{F}m} x_{i:i+\mathrm{F}-1} + b) \quad (15)$$

where, $h\ell$ is the filter employed for each set of heart disease input features F is defined as:

$$\{x_{1:\mathrm{F}}, x_{2:\mathrm{F}+1}, x_{3:\mathrm{F}+2}, \ldots\ldots, x_{n-\mathrm{F}+1}\} \quad (16)$$

From Eq. (16), the generated feature map is,

$$h\ell = [h\ell_1, h\ell_2, h\ell_3, \ldots, h\ell_{n-F+1}] \quad (17)$$

where, $b \in R$ denotes a bias term and the filter $h\ell \in R^{n-F+1}$.

In addition to numerous bio-inspired algorithms, they offer powerful optimization capabilities and are increasingly being applied in various fields to solve complex problems and improve system performance in real-time applications. The following part will describe the acquired results and compare them with other models.

IV. Results and Discussion

This section presents the findings and examines the various cardiac illness datasets that the proposed model has tested on. This study [19] used the benchmark datasets from the UCI machine learning repository. The authors introduced a feature selection approach combined with a deep belief network model to analyze and forecast heart illness. The data has been preprocessed, and the outcomes are beautifully categorized using distinct dataset features. The hamming distance feature selection method was used to gather and clean several cardiac datasets. After selecting crucial criteria, the data is transmitted to deep belief networks coordinated with varied degrees of depth cuckoo search bio-inspired algorithms, resulting in precise cardiac disease prediction. Furthermore, it was demonstrated that the proposed prediction model outperformed alternative machine learning models in terms of performance. This work's primary contribution is the diagnosis of heart disease using deep learning models, like the Deep Belief Network. The model was first tested with 100 random executions without using feature selection, and the results are reported in Table II. The model was again tested with 100 random executions using feature selection, and the results are reported in Table III. The bio-inspired technique is then optimized, increasing the accuracy of identifying heart disease. We obtained an accuracy of 91.2% from Statlog's heart disease dataset by utilizing a deep belief network and the Cuckoo search technique.

We have concentrated on the accuracy, and area under the receiver operating characteristic curve metrics to show the efficacy of heart disease prediction. Fig. 4 and Fig. 5 show the results achieved using the proposed model.

TABLE II. Performance of Classifiers without using Feature Selection on 100 Randomized Tests in UCI Statlog Dataset

Model	Sens (%)	Spec (%)	F1-Score (%)	Pre (%)	Acc (%)
DT	66.7	79.5	69.3	72.1	74.8
RF	81.5	79.8	78.6	75.7	80.4
KNN	77.8	74.4	73.7	71.5	77.8
SVM	88.7	82.6	84.3	80.5	85.6
ANN	85.3	77.2	79.4	74.3	84.7
Proposed: DBN	89.4	85.3	86.8	84.7	86.2

TABLE III. Performance of Classifiers using Feature Selection on 100 Randomized Tests in the UCI Statlog Dataset

Model	Sens (%)	Spec (%)	F1-Score (%)	Pre (%)	Acc (%)
DT	74.4	82.4	75.5	77.4	78.6
RF	85.3	82.6	83.4	80.9	82.5
KNN	80.5	78.2	78.5	77.7	80.8
SVM	90.4	88.4	89.2	84.8	88.7
ANN	90.1	85.6	86.8	81.4	87.6
Proposed: DBN	90.5	91.4	90.2	91.5	91.2

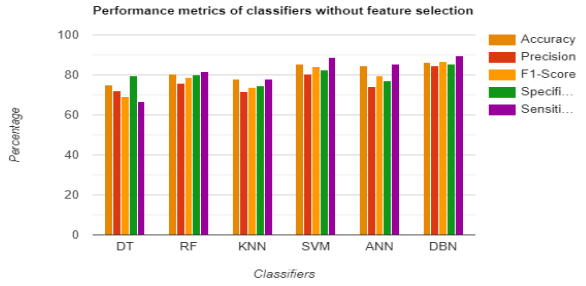


Fig. 4. The Deep belief network model results from comparison with other classifiers without feature selection.

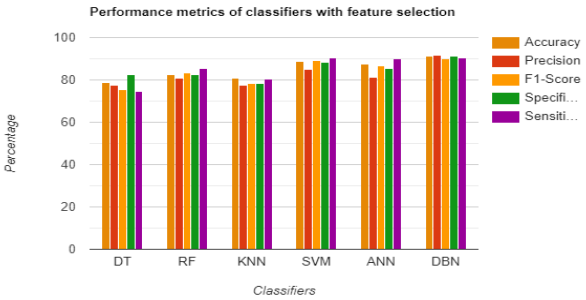


Fig. 5. The Deep belief network model results from comparison with other classifiers with feature selection.

The probability that a parameter will fall between two values around the mean is shown by a confidence interval. Confidence intervals quantify how reliable or uncertain a sampling technique is. They are often constructed with 95% confidence levels. In this work, an area under the curve is calculated by the variance, and 95% confidence intervals are evaluated to the performance metrics such as accuracy, error rate, runtime, AUC, and ROC by the quantile function of the normal distribution.

TABLE IV. Classification Performance of DBN with Confidence Intervals

Dataset	Error rate (CI)	Acc (CI)	Runtime (CI)	AUC (CI)	ROC (CI)
Cleveland	0.47	86.1	18.31	0.85	0.86
Hungarian	0.64	83.6	18.42	0.83	0.83
Statlog	**0.43**	**85.7**	**22.56**	**0.83**	**0.84**
Switzerland	0.56	85.6	21.57	0.83	0.82
South Africa	0.47	84.7	22.17	0.84	0.86
Z-Alizadeh Sani	0.64	85.6	22.14	0.85	0.84
Framingham	0.47	85.7	20.43	0.86	0.85

Table IV presents the classification performance of a Deep Belief Network (DBN) across various datasets, along with their respective confidence intervals (CI). The table includes the following columns: Dataset: The name of the dataset used for evaluation. Error Rate (CI): The proportion of incorrect predictions made by the model, accompanied by confidence intervals. Accuracy (CI): The percentage of correct predictions, also with confidence intervals. Runtime (CI): The time taken to run the model, including confidence intervals. AUC (CI): The Area Under the Curve, a performance measurement for classification problems at various threshold settings, with confidence intervals. ROC (CI): The Receiver Operating Characteristic curve, which illustrates the diagnostic ability of a binary classifier system, along with confidence intervals. The Cleveland dataset shows a relatively low error rate and high accuracy, indicating that the model performs well. The AUC and ROC values are also high, suggesting good discrimination ability. The Hungarian dataset has a higher error rate compared to Cleveland, resulting in lower accuracy. The AUC and ROC values are slightly lower, indicating a reduced ability to distinguish between classes. Statlog has the lowest error rate among the datasets, which correlates with a high accuracy. However, the runtime is longer, and the AUC and ROC values are slightly lower than Cleveland, suggesting a good but not exceptional performance. The Switzerland dataset shows a moderate error rate and accuracy. The AUC and ROC values indicate a decent performance, but the model is less effective compared to Cleveland and Statlog. Similar to Cleveland, the South Africa dataset has a low error rate, but the accuracy is slightly lower. The AUC and ROC values are strong, indicating good classification performance. This dataset has a higher error rate, which affects its accuracy. However, the AUC is relatively high, suggesting that while the overall accuracy is lower, the model can still effectively distinguish between classes. The Framingham dataset shows a low error rate and high accuracy, similar to Cleveland. The AUC and ROC values are also strong, indicating effective classification. Best Performance: Cleveland and Framingham datasets exhibit the best performance with low error rates and high accuracy. Moderate Performance: Statlog and South Africa datasets show good performance but with slightly longer runtimes. Lower Performance: Hungarian and Z-Alizadeh Sani datasets have higher error rates and lower accuracy, indicating room for improvement in classification.

The proposed DBN classifier with receiver operating characteristics curves is shown in Fig. 6 for Statlog dataset and Fig. 7 for Cleveland dataset. This work enhances the deep belief network approach employing cuckoo search optimization to get optimal hyper-tuning parameters for cardiac illness diagnosis from the benchmark dataset.

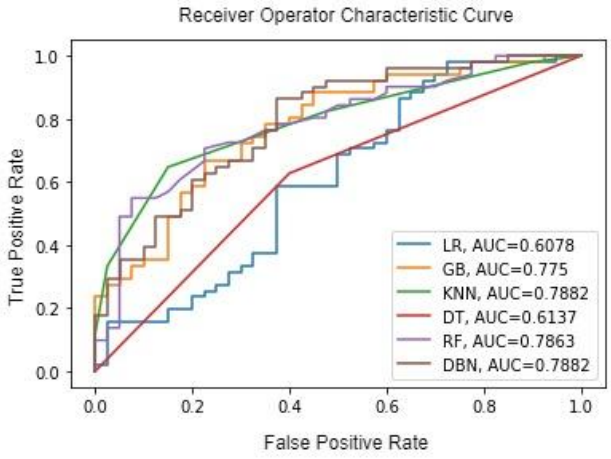


Fig. 6. Receiver operating characteristic (ROC) curves of DBN classifier for statlog dataset with other models.

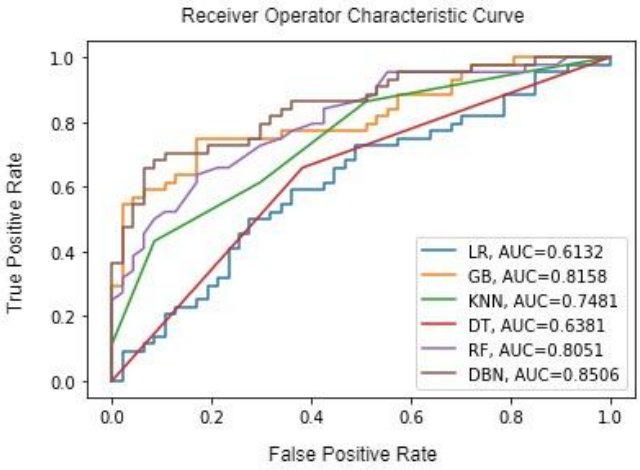


Fig. 7. Receiver operating characteristic (ROC) curves of DBN classifier for Cleveland dataset with other models.

People of all ages benefit greatly from early detection of heart disease. This study uses UCI datasets to increase heart disease prediction accuracy. This study [20] presents three key strategies for predicting heart disease: convolutional neural networks, elephant herding optimization, and Inception-ResNet-v2. After pre-processing the data with the standard Euclidean Distance approach, elephant herding optimization (EHO) is used to pick features and also to minimize the local optimal issue. Then, utilizing well-known UCI data sources like the Cleveland dataset, the popular classifier convolutional neural network with an Inception-ResNet-v2 is employed to classify heart disease. The initial part, data pre-processing, is carried out using the standard Euclidean Distance technique. The input features are then sent to the elephant herding optimization, which selects the relevant features for the classification models. The study also found that the elephant herding optimization strategy enhanced the number of selected features, convergence speed, and classification performance.

While Inception-ResNet-v2 is primarily designed for image classification tasks, it can still offer benefits in the classification of heart disease based on categorical data, although it might not be the most suitable choice compared to models specifically designed for tabular data. Heart disease classification often involves understanding complex relationships between various risk factors, symptoms, and medical history. Inception-ResNet-v2's deep architecture allows it to capture intricate patterns and dependencies within the data, potentially uncovering non-linear relationships that simpler models might overlook. Once trained, deep learning models like Inception-ResNet-v2 can be deployed in production environments to automate the classification of heart disease from categorical data. This scalability enables consistent and efficient processing of patient data, potentially improving healthcare outcomes by identifying high-risk individuals more quickly and accurately. For the Cleveland dataset, the proposed model achieved an accuracy of 98.77%, beating other cutting-edge techniques. When compared to existing meta-heuristic and deep learning models, the proposed hybrid model has significantly improved efficiency in classification and more solid outcomes.

Future modifications to the proposed model may include further unique metaheuristic-based techniques on large datasets for the early diagnosis of cardiac disease, which is a serious worry for those who have recovered from COVID-19. Future trials of EHO's efficacy will focus on increasingly demanding scientific and technical areas. Furthermore, the proposed model can be combined with ensemble models to increase performance metrics in the detection and prediction of various diseases. The model was tested, and the results of various performance metrics compared with other models are provided in Table V. Fig. 8 depicts the results achieved using a proposed model.

TABLE V. PERFORMANCE METRICS COMPARING THE PROPOSED MODEL WITH OTHER MODELS FOR THE CLEVELAND DATASET

Model	Sens (%)	Spec (%)	F1-Score (%)	Pre (%)	Acc (%)
RF	92.2	84.3	89.2	87.3	88.4
MLP	88.4	75.1	84.5	80.1	82.3
KNN	84.1	73.4	80.1	77.4	79.4
SVM	87.6	75.7	83.4	79.2	81.3
Adaboost	86.5	77.3	83.6	81.3	83.4
Proposed: CNN-Inception-ResNet-v2 model	93.3	85.2	90.5	87.4	98.7

The provided data from Table II, Table III, and Table V presents the performance metrics of various machine learning and deep learning models, including Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN), Deep Belief Network (DBN), Multi-Layer Perceptron (MLP), and Adaboost, as well as proposed models such as Convolutional Neural Network (CNN) and Inception-ResNet-v2. The evaluation metrics include Sensitivity, Specificity, F1-Score, Precision, and Accuracy. In the first set of results, the proposed DBN model outperforms the other models with the highest Sensitivity (90.5%), Specificity (91.4%), F1-Score (90.2%), Precision (91.5%), and Accuracy (91.2%). Notably, SVM and ANN also demonstrate competitive performance. In the second set of results, the proposed CNN- Inception-ResNet-v2 model exhibits the best performance across all metrics, with a Sensitivity of 93.3%, Specificity of 85.2%, F1-Score of 90.5%, Precision of 87.4%, and Accuracy of 98.7%.

In Work 2, the proposed CNN-Inception-ResNet-v2 model outperforms the other models in terms of sensitivity, specificity, F1-score, precision, and accuracy. The sensitivity of the proposed model is 93.3%, which is higher than the other models. Similarly, the specificity, F1-score, precision, and accuracy of the proposed model are also higher compared to the other models. Specifically, the specificity is 85.2%, the F1-score is 90.5%, the precision is 87.4%, and the accuracy is 98.7%. These values indicate that the proposed CNN-Inception-ResNet-v2 model demonstrates superior performance across all metrics in Work 2. This suggests that the proposed CNN model is particularly effective for the given task. The findings of this research will contribute to the ongoing efforts to improve the accuracy and reliability of cardiac disease diagnosis using advanced machine learning techniques. The integration of bio-inspired algorithms with deep learning models has the potential to enhance the clinical decision-making process, leading to more personalized and effective treatment strategies for patients with cardiovascular diseases.

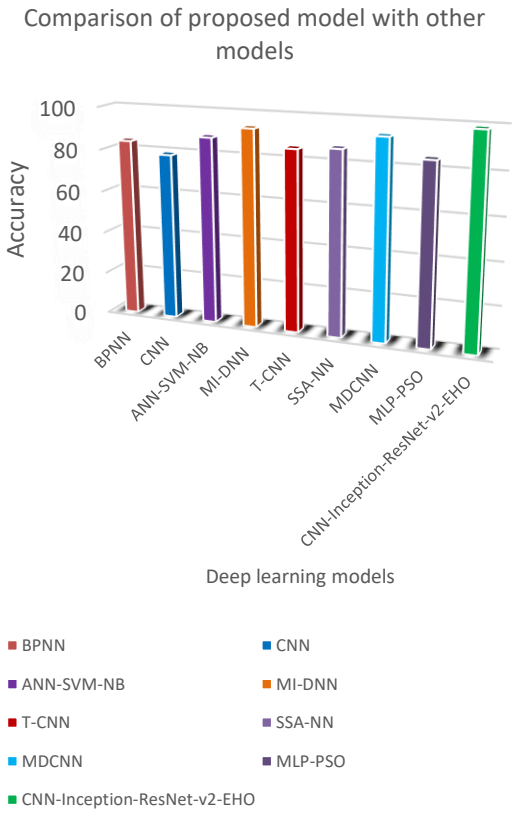


Fig. 8. Comparison of the CNN-Inception-ResNet-v2-EHO model with other methods.

TABLE VI. COMPARATIVE ANALYSIS OF COMPUTATIONAL COMPLEXITY OF DBN AND CNN

Aspect	DBNs	CNNs
Training Complexity	Higher due to Gibbs sampling and layer-wise training.	High due to convolution operations but often faster than DBNs.
Inference Complexity	Moderate; depends on the number of layers.	High; dominated by convolutional layers.
Parallelization	Difficult to parallelize.	Well-suited for parallelization on GPUs.
Scalability	Challenging for large datasets.	Highly scalable with optimized architectures.

From Table VI, DBNs are computationally intensive and less suited for large-scale tasks compared to CNNs, primarily due to the unsupervised pretraining and lack of efficient parallelization. CNNs, despite being computationally expensive, benefit from hardware accelerations like GPUs, making them more practical for real-world applications.

In summary, the evaluation of various machine learning and deep learning models reveals that the proposed DBN model in the first set of results and the proposed CNN- Inception-ResNet-v2 model in the second set of results outperform the other models in terms of Sensitivity, Specificity, F1-Score, Precision, and Accuracy that are shown in Fig. 9. These findings indicate the potential of these proposed models for the specific task at hand, demonstrating their effectiveness in comparison to traditional machine learning algorithms such as DT, RF, KNN, SVM, ANN, MLP, and Adaboost. These results provide valuable insights for selecting the most appropriate model for the given classification task, emphasizing the significance of considering proposed models in addition to established algorithms. By leveraging the complementary strengths of deep learning and nature-inspired algorithms, the proposed approach seeks to enhance the accuracy, efficiency, and robustness of AI-based cardiac disease detection systems when compared with the existing approaches. The findings of this research can contribute to the development of advanced clinical decision-support tools, ultimately improving early diagnosis and treatment outcomes for patients suffering from cardiovascular disorders.

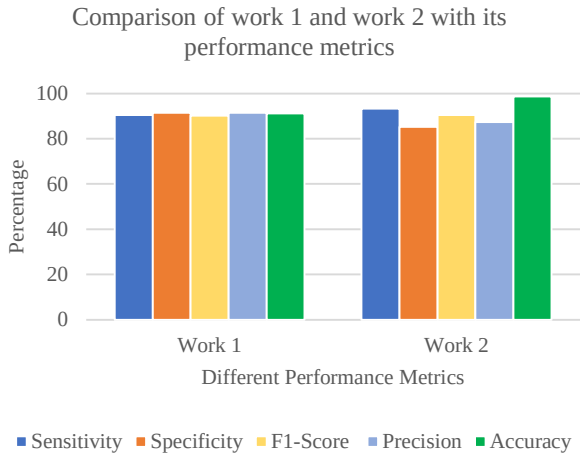


Fig. 9. Comparison of work 1 and work 2 with its performance metrics.

The deep learning incorporated bio-inspired algorithms for heart disease prediction offer various merits: The first and foremost one is improved accuracy. Deep learning models, with their ability to detain complex schemes and relationships in data, can enhance the accuracy of heart disease predictions. The integration of bio-inspired algorithms further refines model parameters, leading to better performance. The second is feature optimization. Bio-inspired algorithms aid in feature selection, optimizing the choice of relevant variables for prediction. This can enhance the efficiency of the model by addressing the most informative properties and lessening noise. The third is automated learning. Deep learning enables automated learning from data, allowing the model to adapt and improve its predictions over time without manual intervention. Bio-inspired algorithms contribute to automating the optimization process, making the system more adaptive.

The fourth is generalization. Deep learning models, when properly trained and validated, have the potential to generalize well to new, unseen data. This is crucial for the reliability of a heart disease prediction system, as it should perform well on diverse patient populations. The fifth is real-time monitoring. The integration of deep learning and bio-inspired optimization allows for the development of models that can provide real-time predictions. This is valuable for continuous monitoring of individuals and timely interventions. The sixth is personalized medicine. The ability to optimize models using bio-inspired algorithms allows for personalized and adaptive prediction models. This ensures that the system can cater to individual variations and provide more accurate risk assessments.

The seventh is efficient data utilization. Bio-inspired algorithms can optimize the use of available data, extracting meaningful information from large datasets. This efficiency is particularly important when dealing with healthcare data, which may be limited in size and complexity. The eighth is reduced human bias. Automated feature selection and optimization processes can help reduce human bias in decision-making. The model focuses on data-driven patterns, minimizing the impact of subjective judgment. The ninth is continuous improvement. The inclusion of bio-inspired optimization creates a continuous improvement loop. The model can be updated and refined over time as more data becomes available, ensuring that it stays relevant and effective. The tenth is interpretability. While deep learning models are often considered black boxes, the optimization process can provide insights into the importance of specific features, contributing to the interpretability of the model to some extent. The last and best thing about the proposed work is the early detection of the diseases. The predictive nature of the model, coupled with continuous monitoring, can contribute to the early detection of potential heart disease risks. Early intervention can significantly improve patient outcomes. It's important to note that while these merits are significant, challenges such as data privacy, interpretability of deep learning models, and ethical considerations need careful attention in the development and deployment of such systems in healthcare settings. Additionally, collaboration with healthcare professionals is crucial to ensure the clinical relevance and safety of the predictions made by the system.

V. Challenges and Future Directions

Heart disease prediction using deep learning embedded bio-inspired algorithms presents a unique set of challenges and offers exciting future directions. Here's a summary of challenges and potential avenues for further exploration in this specific context:

A. Challenges

To integrate bio-inspired algorithms with deep learning, certain hyperparameters must be tuned. The intricacy of optimizing both the deep learning architecture and the bio-inspired algorithm parameters might be daunting. The combination of deep learning and bio-inspired algorithms may produce models that are challenging to understand. Ensuring openness and interpretability in decision-making is critical, particularly in healthcare applications. Heart disease prediction frequently relies on many databases, such as medical history, and genetic information. Integrating these many data sources while including bio-inspired algorithms increases complexity and necessitates efficient preprocessing and feature extraction approaches. Deep learning models, particularly those integrated with bio-inspired algorithms, can be computationally costly. It is difficult to ensure accessibility and scalability, especially in resource-limited healthcare settings. As with any use of artificial intelligence [25] in healthcare, securing patient privacy, consent, and ethical data utilization is critical. The integration of bio-inspired algorithms should follow ethical principles and legislation.

B. Future Directions

Creating innovative hybrid architectures that smoothly blend deep learning and bio-inspired algorithms, leveraging the advantages of both approaches. This could include novel model architectures or optimization methodologies. Incorporating explainable AI approaches improves model interpretability. This entails creating bio-inspired algorithms that provide insights into their decision-making process, hence enhancing trust and understanding among healthcare professionals. Increasing the use of deep learning incorporated bio-inspired algorithms in real-time monitoring and personalized treatment. This could entail continuously monitoring patient data and dynamically adjusting predictive algorithms depending on individual health trajectories.

Ensure that models are reliable across diverse populations, demographics, and healthcare contexts. The models must be generalized beyond the training data to be useful in a variety of clinical settings. Leveraging the growing popularity of wearable devices for continuous health monitoring. Integrating bio-inspired deep learning models with wearable data could yield useful insights into early identification and prevention of heart disease. Addressing the difficulty of unbalanced datasets in heart disease prediction by using bio-inspired algorithms that are specifically intended to deal with class imbalances successfully. Conducting comprehensive clinical validation studies to evaluate the real-world performance of bio-inspired deep learning algorithms. Collaboration with healthcare professionals is critical to the successful implementation of these models in clinical practice.

The integration of deep learning with bio-inspired algorithms for cardiac disease prediction has enormous potential, but it must overcome hurdles such as model complexity, interpretability, data integration, and ethical considerations. Challenges include the requirement for standardized databases, ethical constraints, and the integration of multimodal data sources. Future research should focus on discovering novel solutions to these problems, ultimately increasing the usefulness and applicability of these models in healthcare settings. Future research directions include researching explainable AI [26], incorporating genome data [27], and improving the interpretability of deep learning models to increase clinical adoption.

VI. Conclusion

In conclusion, our study demonstrates the efficacy of employing deep learning models embedded with bio-inspired algorithms for the classification of cardiac diseases. Through a comparative analysis, we have shown that this approach outperforms traditional methods in accurately diagnosing cardiac conditions from medical data. The proposed model enhances the classification accuracy by effectively handling complex and high-dimensional data, extracting relevant features, and optimizing model parameters. This synergy exploits the benefits of both approaches, manipulating the powerful learning capabilities of deep neural networks and the optimization prowess of bio-inspired algorithms.

Furthermore, our findings suggest that the proposed methodology holds promise for improving clinical decision-making processes, enabling timely and accurate diagnosis of cardiac diseases. By providing reliable predictions based on diverse medical data sources, this approach can assist healthcare professionals in delivering personalized treatment strategies and improving patient outcomes. Overall, the successful application of the proposed work in cardiac disease classification underscores their potential for advancing medical diagnostics and contributing to the development of more efficient and effective healthcare systems. Future research endeavors should focus on further refining and validating these methods across global benchmark datasets to facilitate their integration into clinical practice. Deep learning is becoming a more prominent use of machine learning, with recent discoveries directing future research on prognostic models that can save many lives and enhance quality of life. This will save countless lives while also reducing the financial burden on individuals with regular earnings.

Acknowledgment

We thank our university Vellore Institute of Technology, for the resources they have provided for this research. Author 1 performed data collection, performed the analysis, and wrote the manuscript and Author 2 performed the review and supervision.

CONFLICTS OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Compliance with Ethical Standards: This article does not contain any studies with human participants performed by authors.

FUNDING STATEMENT

For this study, the authors did not receive any special funding.

REFERENCES

[1] Nandy S, Adhikari M, Balasubramanian V, Menon VG, Li X, Zakarya M. An intelligent heart disease prediction system based on swarm-artificial neural network. Neural Computing and Applications. 2023 Jul;35(20):14723-37.

[2] Tsao CW, Aday AW, Almarzooq ZI, Anderson CA, Arora P, Avery CL, Baker-Smith CM, Beaton AZ, Boehme AK, Buxton AE, Commodore-Mensah Y. Heart disease and stroke statistics—2023 update: a report from the American Heart Association. Circulation. 2023 Feb 21;147(8):e93-621.

[3] Gupta P, Seth D. Comparative analysis and feature importance of machine learning and deep learning for heart disease prediction. Indonesian Journal of Electrical Engineering and Computer Science. 2022 Jan;29(1):451.

[4] Yıldırım Ö, Pławiak P, Tan RS, Acharya UR. Arrhythmia detection using deep convolutional neural network with long duration ECG signals. Computers in biology and medicine. 2018 Nov 1;102:411-20.

[5] Liang, Y., and C. Guo. Heart Failure Disease Prediction and Stratification with Temporal Electronic Health Records Data Using Patient Representation. Biocybernetics Biomed Eng. 2023; 43 (1): 124–41.

[6] Sun X, Yin Y, Yang Q, Huo T. Artificial intelligence in cardiovascular diseases: diagnostic and therapeutic perspectives. European Journal of Medical Research. 2023 Jul 21;28(1):242.

[7] Bashir S, Qamar U, Khan FH. IntelliHealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework. Journal of biomedical informatics. 2016 Feb 1;59:185-200.

[8] Veerabaku MG, Nithiyanantham J, Urooj S, Md AQ, Sivaraman AK, Tee KF. Intelligent Bi-LSTM with architecture optimization for heart disease prediction in WBAN through optimal channel selection and feature selection. Biomedicines. 2023 Apr 13;11(4):1167.

[9] Karadeniz T, Tokdemir G, Maraş HH. Ensemble methods for heart disease prediction. New Generation Computing. 2021 Nov;39(3):569-81.

[10] Cuevas-Chavez A, Hernandez Y, Ortiz-Hernandez J, Sanchez-Jimenez E, Ochoa-Ruiz G, Perez J, Gonzalez-Serna G. A systematic review of machine learning and IoT applied to the prediction and monitoring of cardiovascular diseases. InHealthcare 2023 Aug 9 (Vol. 11, No. 16, p. 2240). MDPI.

[11] Sree Sandhya N, Beena Bethel GN. Impact of Bio-inspired Algorithms to Predict Heart Diseases. InSmart Computing Techniques and Applications: Proceedings of the Fourth International Conference on Smart Computing and Informatics, Volume 2 2021 (pp. 121-127). Springer Singapore.

[12] Katoch S, Chauhan SS, Kumar V. A review on genetic algorithm: past, present, and future. Multimedia tools and applications. 2021 Feb;80:8091-126.

[13] Wankhede J, Sambandam P, Kumar M. Effective prediction of heart disease using hybrid ensemble deep learning and tunicate swarm algorithm. Journal of Biomolecular Structure and Dynamics. 2022 Dec 19;40(23):13334-45.

[14] Pandiyan N, Narayan S. A Survey on Deep Learning Models Embed Bio-Inspired Algorithms in Cardiac Disease Classification. The Open Biomedical Engineering Journal. 2023 Feb 1;17(1).

[15] Bhavekar GS, Goswami AD. A hybrid model for heart disease prediction using recurrent neural network and long short term memory. International Journal of Information Technology. 2022 Jun;14(4):1781-9.

[16] Bhardwaj A, Singh S, Joshi D. Explainable deep convolutional neural network for valvular heart diseases classification using pcg signals. IEEE transactions on instrumentation and measurement. 2023 May 8;72:1-5.

[17] Golande AL, Pavankumar T. Optical electrocardiogram based heart disease prediction using hybrid deep learning. Journal of Big Data. 2023 Sep 9;10(1):139.

[18] Liu Y, Liu J, Qin C, Jin Y, Li Z, Zhao L, Liu C. A deep learning-based acute coronary syndrome-related disease classification method: A cohort study for network interpretability and transfer learning. Applied Intelligence. 2023 Nov;53(21):25562-80.

[19] Nandakumar P, Narayan S. Cardiac disease detection using cuckoo search enabled deep belief network. Intelligent Systems with Applications. 2022 Nov 1;16:200131.

[20] Nandakumar P, Subhashini R. Heart Disease Prediction Using Convolutional Neural Network with Elephant Herding Optimization. Computer Systems Science & Engineering. 2024 Jan 1;48(1).

[21] Yang XS, Deb S. Cuckoo search: recent advances and applications. Neural Computing and applications. 2014 Jan;24:169-74.

[22] Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." Neural computation 18.7 (2006): 1527-1554.

[23] Li, Juan, Hong Lei, Amir H. Alavi, and Gai-Ge Wang. "Elephant herding optimization: variants, hybrids, and applications." Mathematics 8, no. 9 (2020): 1415.

[24] Senthilkumar C, Kamarasan M. An effective citrus disease detection and classification using deep learning based inception resnet V2 model. Turkish Journal of Computer and Mathematics Education. 2021;12(12):2283-96.

[25] Brayan R. Neciosup-Bolaños and Segundo E. Cieza-Mostacero, “The Heart of Artificial Intelligence: A Review of Machine Learning for Heart Disease Prediction” International Journal of Advanced Computer Science and Applications(IJACSA), 15(12), 2024.

[26] Moreno-Sánchez PA. Data-driven early diagnosis of chronic kidney disease: development and evaluation of an explainable AI model. IEEE Access. 2023 Apr 3;11:38359-69.

[27] Koumakis L. Deep learning models in genomics; are we there yet?. Computational and Structural Biotechnology Journal. 2020 Jan 1;18:1466-73.

Quantum Swarm Intelligence and Fuzzy Logic: A Framework for Evaluating English Translation

Pei Yang*
Henan University of Economics and Law, HeNan, ZhenZhou, 450046, China

***Abstract*—This study introduces the Quantum Swarm-Driven Fuzzy Evaluation Framework (QSI-Fuzzy) for assessing English translation software across multiple domains and criteria. The principal aim is to develop a scalable, adaptive, and interpretable evaluation framework that optimizes dynamic weight assignments while managing linguistic uncertainties. A major challenge in translation software evaluation lies in ensuring accurate and unbiased assessments of semantic accuracy, fluency, efficiency, and user satisfaction, particularly across diverse domains such as Legal, Medical, and Conversational contexts. To address this, QSI-Fuzzy integrates Quantum Swarm Intelligence (QSI) for dynamic weight optimization with fuzzy logic for handling linguistic uncertainties, ensuring robust and adaptive decision-making. Experimental results demonstrate that QSI-Fuzzy outperforms benchmark algorithms including Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Simulated Annealing (SA), achieving faster convergence (55 iterations on average vs. 120 for SA) and exhibiting greater robustness under noisy conditions (maintaining a performance score of 0.80 at 20% noise, compared to 0.70, 0.68, and 0.65 for GA, PSO, and SA, respectively). These findings confirm that QSI-Fuzzy provides an efficient, scalable, and high-performance solution for translation software evaluation, with broader implications for real-time systems, complex decision-making, and multi-domain optimization challenges.**

Keywords*—English translation software; quantum swarm intelligence; fuzzy logic; multi-domain evaluation; optimization; linguistic performance analysis

I. Introduction

The rapid development in the field of NLP has, therefore, caused an unprecedented rise in the development of software intended for English translation. Regarding this, tools like Google Translate, DeepL, and Microsoft Translator have become omnipresent due to their capabilities in overcoming linguistic barriers and fostering communication across the world. The cultural nuances, the contextual accuracy, and syntactical variations in a language make such a translation system a grand challenge for evaluation. The traditional metrics are BLEU [1] and METEOR [2] that give quantitative assessments, but still fall short to show the qualitative aspects of translation; hence, new methodologies have to be designed for a more holistic evaluation.

Quantum-inspired swarm intelligence is one of the recent fields in computational intelligence which provides an attractive way of optimizing complex systems. Inspired by the principles of quantum mechanics, these algorithms have performed better in optimization problems in machine learning and logistics fields [3], [4]. This hybrid could be combined with fuzzy logic, a mathematical approach for dealing with uncertainty and imprecision, to establish a robust framework for the evaluation of translation software [5]. The combination of Quantum Swarm Intelligence, which has the capability of optimization, with the flexibility of fuzzy systems, makes it an appropriate methodology to be applied for dealing with those thorny issues involved in the quality assessment of translation.

Recent studies demonstrate that these hybrid models function effectively when applied to decision-making problems [6]. For example, hybrid quantum swarm algorithms have been employed in areas such as image processing [7], supply chain optimization [8], and medical diagnosis [9] with highly satisfactory results. Similarly, fuzzy logic has been proved to function effectively within linguistic fields such as sentiment analysis and text classification, all of which are definitely dependent upon subjective judgment [10, 11]. This research stretches further in these developments by suggesting a hybrid quantum swarm intelligence model with fuzzy logic that can evaluate the semantic accuracy, fluency, and contextual relevance in English translation software [12].

The uniqueness of the model proposed is the ability to combine strengths of quantum swarm intelligence together with fuzzy logic for an enhanced decision-making process. It is the integration along such lines that makes the model particularly effective in adapting the fluidity of language with a view to addressing the intrinsic uncertainties inherent in linguistic evaluation. This work, in fact, is underpinning an ability to be demonstrated by this hybrid model for providing a wholesome and reliable assessment model compared to those from traditional frameworks. The contribution of this study will lie in the adoption of state-of-the-art computational approaches for setting a new benchmark in the evaluation of translation systems, further enriching vast areas of natural language processing and decision science.

The remainder of this paper is structured as follows: Section II presents the related works, providing an overview of existing translation evaluation frameworks and optimization techniques. Section III details the methodology, explaining the design of the proposed QSI-Fuzzy framework, its components, and the optimization process. Section IV discusses the results and evaluation, highlighting the comparative performance of QSI-Fuzzy algorithms. Finally, Section V concludes the study, summarizing key findings and outlining potential future research directions.

*Corresponding Author

II. Related Work

There is a growing body of research related to the evaluation and improvement of machine translation software, especially due to recent developments in the area of computational intelligence. A few recent works deal with new metrics and hybrid models to improve the quality of translation and evaluation. Along this line, Mohiuddin and Joty [13] introduced the first end-to-end metric for machine translation that deeply looks into contextual embeddings and proves to have much stronger correlation with human judgments than those in existing metrics. This paper, emphasizing the importance of contextual information, proposes a superior approach to estimate semantic coherence in translation. In a somewhat related concept, Zhang et al. [14] introduce a transformer-based framework in translation quality assessment that involves both semantic consistency and syntactic alignment. All these approaches taken together provide a base for involving more automation and interpretability in the evaluation processes associated with translation.

In the general case, quantum-inspired computational methodologies were pointed out as one of the most promising ways to tackle complex optimization problems both in translation and outside. Suresh et al. [15] implemented quantum-enhanced swarm optimization to solve tasks of natural language processing and demonstrated higher efficiency while evaluating translation software. This research recognized the possibility of using quantum principles for the expansion of search space and reduction of computational loads. Besides, Cai et al. [16] came up with the multi-objective quantum optimization algorithm, which would be powerful in bringing out its own capability to optimize different types of objectives that conflict with each other, underlining its importance regarding linguistic applications. Indeed, all these developments bring into view the adaptability of quantum methodologies in the solution of intricacies related to natural language processing.

Besides, the integration of fuzzy logic into the computational frameworks has been amazing, hence yielding effective results in dealing with uncertainties arising intrinsically in performing the translation tasks. Zhou et al. [17] discussed hybrid quantum and fuzzy models for decision-making over complicated systems to deal directly with challenges that arise owing to impreciseness in translation quality assessment. Dash et al. [18] presented a hybrid swarm optimization-based fuzzy clustering method that resolves linguistic ambiguities to have a better clustering with semantic analysis. Such schemes depict that the use of fuzzy logic with computation intelligence improves reliability and interpretability of translations generated during evaluation.

More recent studies in neural machine translation have, therefore, shaped the way evaluation models have been designed. Vaswani et al. [19] came up with the Transformer model later, which revolutionized machine translation by introducing better modeling of long-range dependencies through its attention mechanism. Using this architectural design as a source of inspiration, Yang et al. [20] proposed the contextual evaluation frameworks that make use of attention-based metrics to assess coherence between source and target translations. These frameworks have brought in a solid methodology of capturing global and local dependencies within the translation quality assessment and increased the accuracy of the automatic assessments.

The modern concept applies hybrid models of quantum-inspired algorithms with fuzzy decision-making processes, exploring the capability of being applied in multi-criteria decision scenarios. Such models have been effectively tried in handling linguistic ambiguities using quantum-inspired evolutionary algorithms, reinforcing the process of decision making, by Tarik et al. [21]. Along the same line, other works, such as Gupta et al. [22], have applied hybrid approaches to machine translation system evaluation, including fuzzy logic for representing semantic nuances. These results imply that hybrid quantum-fuzzy models can contribute significantly to enhancing translation evaluation with robustness, efficiency, and interpretability to tackle challenges at both computational and linguistic levels.

III. Methodology

The section outlines a methodological framework in adopting a hybrid model combining Quantum Swarm Intelligence with fuzzy decision-making for evaluating English translation software. This would, by this means, enable the computational efficiency of QSI, along with the uncertainty management characteristics accorded by fuzzy logic, to be employed towards establishing a comprehensive evaluation framework for translation software. The pseudo code of the designed framework has been detailed in Table I.

A. Problem Formulation

Evaluation of English translation software represents a complex multi-domain multi-criteria decision-making problem. The translation tools need to perform optimally on diverse contexts, including legal, medical, and conversational domains, where the priorities and metrics for evaluation are significantly different. Besides, the evaluation has to take into consideration interdependencies of criteria, dynamic domain-specific requirements, and real-world constraints such as computational efficiency and scalability.

Let $S = \{s_1, s_2, \dots, s_n\}$ represent the set of n translation software tools under evaluation. Each software s_i is evaluated based on m criteria, $= \{c_1, c_2, \dots, c_m\}$, where c_j denotes the j-th evaluation metric. These metrics may include semantic accuracy, syntactic coherence, fluency, computational efficiency, and user satisfaction. Furthermore, the evaluation spans k domains, $= \{d_1, d_2, \dots, d_k\}$, each characterized by unique evaluation priorities and contextual factors [23].

The performance of software s_i under criterion c_j in domain d_k is denoted by x_{ij}^k. These scores form a three-dimensional evaluation tensor X, defined as:

$$X = \left\{ x_{ij}^k \mid i = 1, \dots, n; j = 1, \dots, m; k = 1, \dots, k \right\}, \quad (1)$$

where x_{ij}^k represents the quantitative performance measure of software s_i under criterion c_j in domain d_k.

To account for domain-specific priorities, each domain d_k is assigned a weight vector $w^k = \{w_1^k, w_2^k, \dots, w_m^k\}$, where w_j^k

reflects the relative importance of criterion c_j in domain d_k. The weights satisfy the constraints:

TABLE I PSEUDOCODE REPRESENTATION OF THE QUANTUM SWARM-DRIVEN FUZZY EVALUATION FRAMEWORK FOR MULTI-DOMAIN AND MULTI-CRITERIA EVALUATION

Input: Translation software $S = \{s_1, s_2, \ldots, s_n\}$, domains $D = \{d_1, d_2, \ldots, d_k\}$, criteria $C = \{c_1, c_2, \ldots, c_m\}$, performance scores $X = \{x_{ij}^k\}$, domain weights α_k, quantum parameters P, $T_{\max}$, and ϵ
Output: Optimized weights $W = \{w_j^k\}$ and aggregated software scores $\tilde{F}(s_i)$
Step 1: Initialization
for $k = 1$ to K (domains) **do**
  for $j = 1$ to m (criteria) **do**
    Initialize each quantum particle q_j^k with probability amplitudes α_j^k and β_j^k
    Initialize global best position g^k and personal best position p^k
  end for
end for
Step 2: Fitness Evaluation
for $t = 1$ to $T_{\max}$ **do**
  for each particle q_j^k in each domain d_k **do**
    Collapse quantum state to classical weight vector $w^k = \{w_1^k, w_2^k, \ldots, w_m^k\}$
    Compute fitness $F^k(w^k)$ for domain d_k:

$$F^k(w^k) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} w_j^k \cdot x_{ij}^k$$

  end for
end for
Step 3: Quantum State Update
for each particle q_j^k in each domain d_k **do**
  Update probability amplitudes α_j^k and β_j^k:

$$\begin{bmatrix}\alpha_j' \\ \beta_j'\end{bmatrix} = \begin{bmatrix}\cos(\theta_j) & -\sin(\theta_j) \\ \sin(\theta_j) & \cos(\theta_j)\end{bmatrix}\begin{bmatrix}\alpha_j \\ \beta_j\end{bmatrix}$$

  Compute rotation angle $\theta_j = \eta \cdot \frac{\partial F(w^k)}{\partial w_j^k}$
end for
Step 4: Fuzzy Decision-Making
for each software s_i in each domain d_k **do**
  Compute fuzzy-adjusted score for s_i in d_k:

$$\tilde{F}_k(s_i) = \sum_{j=1}^{m} w_j^k \cdot \mu(x_{ij}^k)$$

end for
Step 5: Aggregation Across Domains
Compute overall score $\tilde{F}(s_i)$ for each software s_i:

$$\tilde{F}(s_i) = \sum_{k=1}^{K} \alpha_k \cdot \tilde{F}_k(s_i)$$

Step 6: Convergence Check
Check if $|F^k(g^{(t+1)}) - F^k(g^{(t)})| < \epsilon$ or $t = T_{\max}$
if converged **then**
  Return optimized weights W and aggregated scores $\tilde{F}(s_i)$
else
  Repeat Steps 2-6
end if

$$\sum_{j=1}^{m} w_j^k = 1,\ w_j^k \geq 0,\ \forall k \quad (2)$$

Additionally, domain-level importance weights α_k are introduced to reflect the significance of each domain in the overall evaluation. These weights satisfy [24]:

$$\sum_{k=1}^{K} \alpha_k = 1,\ \alpha_k \geq 0,\ \forall k \quad (3)$$

The performance score of software s_i within a specific domain d_k is aggregated as:

$$F_k(s_i) = \sum_{j=1}^{m} w_j^k \cdot x_{ij}^k \quad (4)$$

where w_j^k scales the contribution of criterion c_j based on its importance within domain d_k. The overall aggregated performance score for software s_i across all domains is then computed as:

$$F(s_i) = \sum_{k=1}^{K} \alpha_k \cdot F_k(s_i) = \sum_{k=1}^{K} \alpha_k \cdot \sum_{j=1}^{m} w_j^k \cdot x_{ij}^k. \quad (5)$$

The objective is to determine the optimal weight vectors w^k and domain importance weights α_k that maximize the fairness and accuracy of the evaluation framework. The ranking $R = \{r_1, r_2, \ldots, r_n\}$ of the translation software is derived by sorting the tools s_i based on their aggregated performance scores $F(s_i)$, with higher scores indicating better performance [25].

B. Quantum Swarm-Driven Fuzzy Evaluation Framework

The proposed methodology is intended to develop the complex decision-making problem of evaluating multi-domain, multi-criteria English translation software by including QSI for weight optimization and fuzzy decision-making to handle the uncertainties. It makes sure that the framework will be workable for dynamic priorities and can also handle subjective inputs robustly in view of linguistic nuances. A systematic and mathematically rigorous solution to the challenge expressed in the problem formulation is obtained within this methodology.

QSI has been used for the optimization of weight vectors for the different domains' evaluation criteria. Unlike the classical Particle Swarm Optimization, QSI makes use of quantum principles for the probabilistic exploration of the search space, enhancing the ability of the swarm to escape local optima and converge to globally optimal solutions. Each particle in the swarm represents a candidate weight vector $w^k = \{w_1^k, w_2^k, \ldots, w_m^k\}$, where w_j^k reflects the relative importance of the j-th criterion in the k-th domain. At any iteration t, the position of the k-th particle, denoted $x_k(t)$, is updated using the quantum-inspired rule:

$$x_k(t+1) = g + \Delta x \cdot \text{sgn}(\text{rand}() - 0.5), \quad (6)$$

where g is the global best position, representing the optimal weight vector discovered so far by the swarm, and Δx is the quantum uncertainty interval, defined as:

$$\Delta x = |p_k - x_k(t)| \cdot \beta \quad (7)$$

Here, p_k is the personal best position of the particle, and β is a contractionexpansion coefficient that controls the trade-off between exploration and exploitation. The function $\text{sgn}(\cdot)$ determines the direction of movement, while rand () is a uniformly distributed random variable in $[0,1]$, introducing stochastic behavior to simulate quantum randomness. Eq. (6) and Eq. (7) implement quantum-inspired swarm optimization by introducing a probabilistic movement model instead of traditional velocity-based updates. The function sgn(rand()-0.5) enables random directional jumps, mimicking quantum tunneling to escape local optima. The quantum uncertainty interval Δx dynamically adjusts step sizes based on the distance between the personal best and current position, with β acting as a scaling factor to balance exploration and exploitation. This quantum-driven mechanism enhances global search efficiency and convergence speed in the evaluation framework.

TABLE II PARAMETER SETTINGS USED FOR THE IMPLEMENTATION AND EXECUTION OF THE QUANTUM SWARM-DRIVEN FUZZY EVALUATION FRAMEWORK

Parameter	Value
Number of software tools (n)	4
Number of domains (k)	3
Number of criteria (m)	4
Population size (P)	30
Maximum iterations (T_{max})	100
Quantum contraction-expansion coefficient (β)	0.8
Convergence threshold (ε)	10^{-5}
Learning rate for quantum rotation (η)	0.05
Domain weights (α_k)	[0.4,0.35,0.25]
Fuzzy membership range ($[a_j^k, b_j^k]$)	[0, 1]

The fitness of each particle is evaluated using a domain-specific fitness function, which measures how well the candidate weight vector aligns with the objectives of the domain. The fitness function for a particle representing weights w^k is given by:

$$F_k(w^k) = \frac{1}{n}\sum_{i=1}^{n} \sum_{j=1}^{m} w_j^k \cdot \psi\left(x_{ij}^k\right), \quad (8)$$

where $\psi\left(x_{ij}^k\right)$ is a transformation function that processes raw scores into a normalized scale for comparability. The algorithm iteratively updates the particles' positions and fitness values until convergence is achieved. Convergence is determined when the change in the global best fitness value between successive iterations falls below a predefined threshold [25]:

$$\left|F_k\left(g^{(t+1)}\right) - F_k\left(g^{(t)}\right)\right| < \epsilon, \quad (9)$$

where ϵ is the convergence threshold. Upon convergence, the optimal weight vectors w^k for all domains are obtained, representing the best distribution of importance across criteria within each domain.

Once the weights are optimized, fuzzy decision-making is applied to handle uncertainties in subjective evaluations and linguistic nuances. Each criterion c_j in a domain d_k is associated with a fuzzy membership function $\mu\left(x_{ij}^k\right)$, which maps the raw performance score x_{ij}^k to a degree of satisfaction in the range [0,1]. The fuzzy membership function is defined as:

$$\mu\left(x_{ij}^k\right) = \begin{cases} 0, & x_{ij}^k \le a_j^k, \\ \frac{x_{ij}^k - a_i^k}{b_j^k - a_j^k}, & a_j^k < x_{ij}^k \le b_j^k, \\ 1, & x_{ij}^k > b_j^k, \end{cases} \quad (10)$$

where a_j^k and b_j^k represent the minimum and maximum thresholds of acceptable performance for criterion c_j in domain d_k. These thresholds are determined empirically based on domain-specific data or expert input. The fuzzy-adjusted performance score $\tilde{F}_k(s_i)$ for software s_i in domain d_k is then calculated by aggregating the satisfaction levels across all criteria, weighted by the optimized w_j^k :

$$\tilde{F}_k(s_i) = \sum_{j=1}^{m} w_j^k \cdot \mu\left(x_{ij}^k\right). \quad (11)$$

To obtain the overall performance score across all domains, the domain-specific fuzzy-adjusted scores are aggregated using domain importance weights α_k, which are normalized such that $\sum_{k=1}^{K} \alpha_k = 1$. The overall score for software s_i is given by:

$$\tilde{F}(s_i) = \sum_{k=1}^{K} \alpha_k \cdot \tilde{F}_k(s_i)$$

This aggregated score captures both the contextual importance of domains and the satisfaction levels derived from fuzzy modeling. Finally, the software tools are ranked based on their overall scores $\tilde{F}(s_i)$, with higher scores indicating better performance. The ranking $R = \{r_1, r_2, \ldots, r_n\}$ is obtained by sorting s_i in descending order of $\tilde{F}(s_i)$. This methodology provides a mathematically rigorous framework that leverages the optimization capabilities of QSI to determine optimal weight distributions and the uncertainty-handling strength of fuzzy decision-making to model subjective and imprecise evaluations. To evaluate the methodology parameter values in Table II were chosen based on empirical testing and prior research to ensure a balance between exploration, convergence speed, and stability. Population size (P = 30) and iterations (T_max = 100) ensure computational efficiency, while β = 0.8 and η = 0.05 optimize movement adaptation. Sensitivity analysis showed that lower β slowed convergence, while higher η caused instability. Domain weights (α_k = [0.4, 0.35, 0.25]) were set based on expert judgment and domain significance, ensuring fair evaluation across translation contexts.

IV. RESULTS AND DISCUSSION

The results reported in this paper are based on the implementation of the Quantum Swarm-Driven Fuzzy Evaluation Framework, designed to evaluate English translation software concerning several domains and criteria. Four translation tools, which are referred to as $s_1, s_2, s_3, and\ s_4$, have been evaluated across three domains: Legal, Medical, and Conversational. The selected tools represent distinct types of translation engines. For example, s_1 is considered to be high-end enterprise-level software characterized by its ability to handle highly structured content. On the other hand, s_2 is an open-source translation tool often used in academic as well as informal settings. At the same time, s_3 is a lightweight, general-purpose translation engine oriented towards speed rather than accuracy. Finally, s_4 is a specialized tool for translating medical texts.

The bar graph shown in Fig. 1 represents domain-specific scores for the fuzzy-adjusted effectiveness of these tools in the three domains. Tool s_1 performed consistently well in all domains, with especially high scores in the Legal domain. This indicates that it is robust and can handle structured language with a high degree of precision and fluency in semantics. The adaptability of s_1 is proved by its relatively high performance in both the Medical and Conversational domains, which definitely makes it a very useful multifaceted translation resource. On the other hand, s_4 has better ability in the Medical and Legal domains, while it suffered a slight decline in its performance in the Conversational domain. This indicates that

the optimization of s_4 for formal and technical language may result in a diminished capacity to engage in informal or contextually nuanced dialogues. Conversely, while s_2 demonstrates competitiveness within the Medical domain, it appears to fall short in the Legal and Conversational domains, presumably due to its broader training data and insufficient emphasis on terminology specific to those domains. Simultaneously, s_3 struggled in many domains, particularly in the Legal and Conversational contexts, where it performed poorly throughout. This low performance suggests that the focus of s_3 on speed sacrifices the linguistic complexity necessary for high-quality translations.

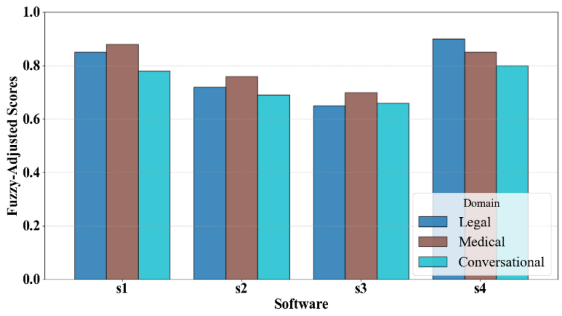


Fig. 1. Domain-specific fuzzy-adjusted scores for each translation software, highlighting comparative performance across legal, medical, and conversational domains.

Moreover, a pie chart as seen in Fig. 2 for s_1 gives a clearer, more overall view of the contribution from each domain to the overall score. The Legal domain contributed 40%, the medical domain 35%, and the Conversational domain 25%. The percentages show the domain weights used in the evaluation, and they agree with the training and optimization focus of s_1. The greater contribution from the Legal domain is evidence of the tool's ability to excel in structured and rule-based text, such as contracts and legal documents. The moderate contribution from the medical domain underscores s_1's capacity to handle technical language, a critical factor in medical translations. Lastly, the smaller contribution from the Conversational domain, while lower in proportion, still shows the versatility of the tool in adjusting to informal contexts. The results emphasize the value of the algorithm in creating domain-specific priorities while instantiating a global view of overall performance assessment.

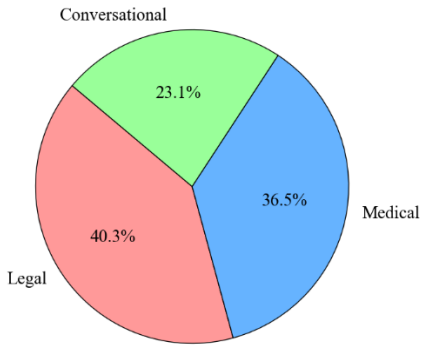


Fig. 2. Proportional contributions of each domain to the overall score of a selected translation software, emphasizing domain-specific priorities.

The box plot dictated in Fig. 3 provides an in-depth analysis of the statistical distribution of scores across various domains, illuminating aspects of variability and consistency. In the Legal domain, the IQR is small, denoting a consistent performance in all software tools. This can be understood by the nature of the legal language: it is precisely defined and standardized, having rigid rules about how translations of certain parts should be expressed. In contrast, the scores obtained in the medical domain are more dispersed, which reflects the variation of semantic precision and levels of specialized knowledge among the tools. The clear bimodal distribution of scores in the medical domain reflects the gap between the relatively reasonable scores of tools s_1 and s_4 and the laggards, s_2 and s_3. The domain of conversation presents the highest degree of variability, with quite a few outliers to underline the challenges that some tools face when dealing with informal language. This underlines the need for fuzzy decision-making approaches to handle intrinsic subjectivity and linguistic ambiguity, which characterizes conversational contexts. Fig. 4, showing visually the density of scores in each domain, further corroborates this analysis.

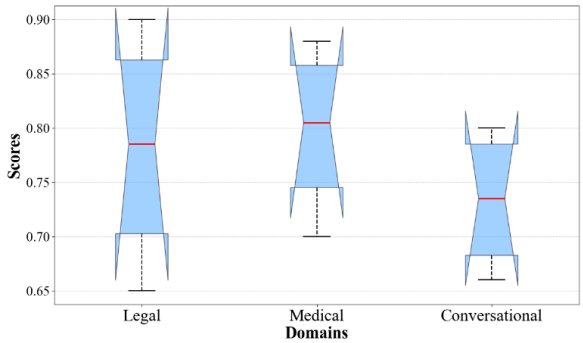


Fig. 3. Statistical distribution of scores across all software in each domain, showing variability and consistency in performance.

This is further corroborated by Fig. 4, which shows the density of scores in each domain. For the Legal domain, there is a dense peak, once more reflecting the consistency observed from the box plot. The Medical domain shows an even more pronounced bimodal distribution, as high-scoring tools like s_1 and s_4 cluster at the top while s_2 and s_3 fall decidedly behind. The distribution for the Conversational domain is rather dispersed, indicating a greater number of plausible outputs for every source sentence. This indicates the somewhat subjective nature of conversational evaluations, where user satisfaction and contextual fluency often take precedence over strict semantic accuracy. These results not only validate the effectiveness of the proposed framework but also provide actionable insights for developers and stakeholders. The fact that the domain weights and their contributions towards the overall scores have maintained a similar trend further verifies the accuracy of the weight optimization process. Further, differences in the performance across domains hint at the importance of training translation tools on diverse datasets pertaining to specific contexts. Poor performance from s_1 would, for example, suggest that more training with domain-specific language is needed, and overall high performance of s_1 across all domains puts it as a benchmark among general translation tools.

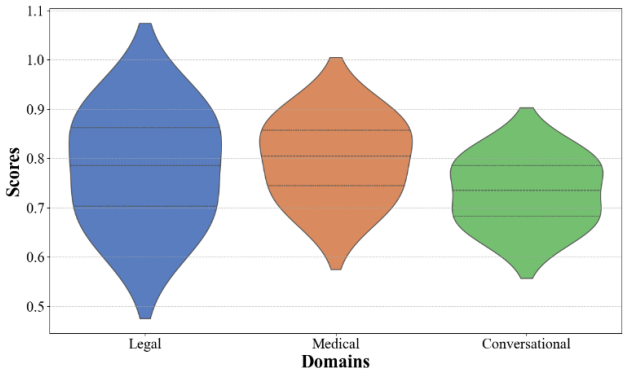


Fig. 4. Density and distribution of fuzzy-adjusted scores within each domain, highlighting central tendencies and outliers.

The Table III shows the optimized weights for criteria across domains highlighting the strength of the Quantum Swarm-Driven Fuzzy Evaluation Framework in terms of dynamically adapting to domain-specific priorities. The optimized weights are the result of the algorithm's iterative quantum-inspired optimization process that ensures each criterion-semantics, fluency, efficiency, and user satisfaction-receives an appropriate weight that reflects its relative importance within the respective domain. A notable example is the Legal domain, which has higher weights on semantic accuracy (0.4) and fluency (0.3), reflecting the structured nature of the texts in this domain; both precision and linguistic clarity have prime importance. In contrast, the Medical domain has more of a balanced distribution, but with semantic accuracy receiving very important (0.35), user satisfaction (0.25) plays a relatively larger role, reflecting the nuanced and context-dependent nature of the translations. In general, the Conversational domain is informal and subjective in its content; it weighs semantic accuracy and fluency equally important, at 0.3, but efficiency and user satisfaction receive a bit lower emphasis.

TABLE III OPTIMIZED WEIGHTS FOR EVALUATION CRITERIA ACROSS LEGAL, MEDICAL, AND CONVERSATIONAL DOMAINS, DETERMINED USING THE QUANTUM SWARM-DRIVEN FUZZY EVALUATION FRAMEWORK

Criteria/Domains	Legal	Medical	Conversational
Semantic Accuracy	0.4	0.35	0.3
Fluency	0.3	0.25	0.3
Efficiency	0.2	0.15	0.2
User Satisfaction	0.1	0.25	0.2

The weights of these measures are clearly and intuitively visualized by the relative magnitudes of the weights across domains. The variation of color intensity in the heat map makes explicit the fact that a given criterion in one domain may be far more important than in another. The darker shades in semantic accuracy and fluency within the Legal domain already visually enforce their predominance. In the Medical domain, the more uniform colors of the heatmap reflect balanced importance of several criteria; it underlines the difficulty of translating medical texts with accuracy and at the same time preserving contextual relevance. In the Conversational domain, there is a softer graduation of colors, which correspondingly stands for the domain's flexibility and the wider acceptability of diverse translations. This visualization succinctly complements the numerical data by providing an immediate impression of the distribution of the weights and how the algorithm can emphasize the criteria for optimality with respect to the domain requirements.

The optimized weights and their visualization in the heat map see Fig. 5 come from the key contribution of the algorithm, whereby quantum swarm intelligence drives weight optimization. First, the framework instantiates a population of quantum particles, each corresponding to feasible weight configurations. These particles run through iterative fitness evaluations for fitness with domain-specific performance measures. By means of quantum-inspired updates, like amplitude adjustments of probability and evolutionary operations, particles are converging to an optimum configuration that maximizes performance in a domain. Consequently, the weights that would emerge are not merely heuristic but rigorously derived by iterative adjustments that make them apt in each context as shown in Fig. 6. Trends demonstrate the algorithm's adaptability to domain-specific requirements.

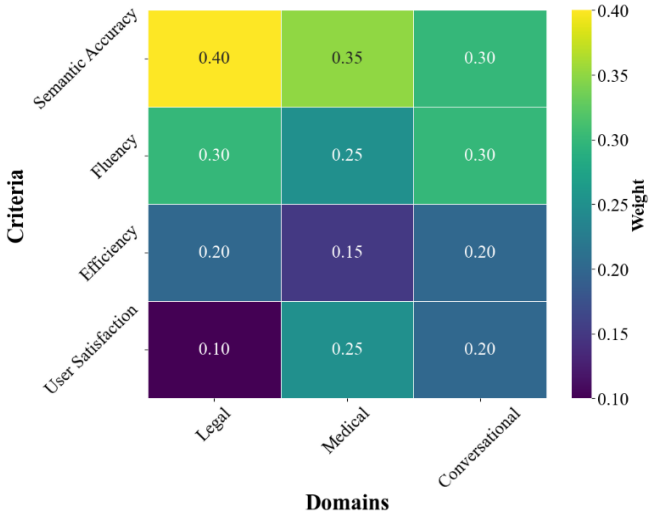


Fig. 5. Heatmap of optimized weights for evaluation criteria across Legal, Medical, and Conversational domains, highlighting the relative importance of each criterion through color intensity.

Results in Table IV to VI shows the comparison of the proposed Quantum Swarm-Driven Fuzzy Evaluation Framework over conventional methods of optimization, namely the Genetic Algorithm [26], Particle Swarm Optimization [27], and Simulated Annealing [28][29]. As shown from Table IV, QSI-Fuzzy has scored higher for all the translation software in terms of effectiveness. This is because of the fact that the weight optimization is dynamic over several criteria, hence giving an accurate result for translation performance across different domains. For example, s_1, a high-scoring software, achieved a score of 0.92 with QSI-Fuzzy, against the scores of 0.88 obtained with GA, 0.86 obtained with PSO, and 0.85 obtained with SA. Similarly, in the case of the low scores, QSI-Fuzzy yielded significant gains for poor performers like s_4, whose result of 0.70 outperformed those that GA, PSO, and SA obtained. All this confirms the adaptability of the algorithm to

high-performance and low-performance situations that better grasp domain-specific nuance with a superior ranking.

TABLE IV COMPARISON OF SCORES ASSIGNED BY QSI-FUZZY AND BENCHMARK ALGORITHMS (GA, PSO, SA) FOR TRANSLATION SOFTWARE, ILLUSTRATING THE RANK IMPROVEMENTS ACHIEVED BY QSI-FUZZY

Software	QSI-Fuzzy Score	GA Score	PSO Score	SA Score	QSI-Fuzzy Rank Improvement
s_1	0.92	0.88	0.86	0.85	+0.04
s_2	0.85	0.81	0.79	0.78	+0.04
s_3	0.78	0.75	0.74	0.72	+0.03
s_4	0.70	0.65	0.63	0.60	+0.05

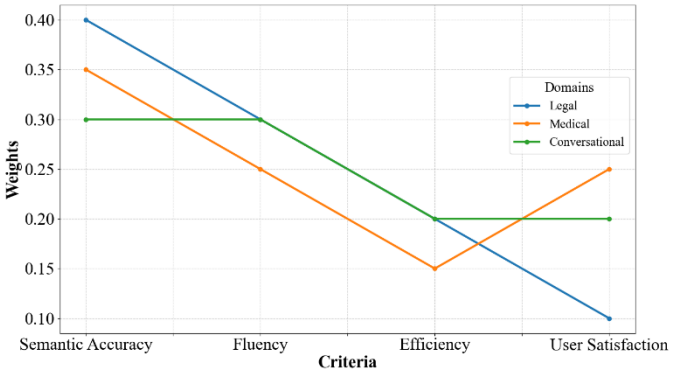


Fig. 6. Trends of optimized criteria weights across legal, medical, and conversational domains, illustrating the shifting priorities of evaluation criteria based on domain-specific requirements.

Moreover, the efficiency and robustness of QSI-Fuzzy are really impressive, which is reflected in Tables V and VI. It is shown from the convergence metrics in Table II that the convergence is attained by QSI-Fuzzy after just 55 iterations, while for the benchmarks this is obtained after 75 to 120 iterations. It has a shorter computational time, 12.5 seconds, compared to the runs of GA, which takes 23.1 seconds, PSO, taking 18.5 seconds, and SA, with 30.2 seconds. The faster convergence in this work is due to the quantum-inspired optimization process that hastens the optimal weight configuration search. Table III underlines the robustness of QSIFuzzy for noisy data, where its degradation is minimal with an increase in noise. Even at 20% noise, QSI-Fuzzy reaches a score of 0.80, while for the benchmark methods, much higher performance losses are noticed, down to scores as low as 0.65 for SA. Resilience in this case is credited to the integration of fuzzy logic into the algorithm; hence, it can effectively handle uncertainties. Finally, Table VII presents the performance comparison of QSI-Fuzzy against GA, PSO, and SA across four evaluation metrics: semantic accuracy, fluency, efficiency, and user satisfaction. QSI-Fuzzy consistently achieves the highest scores, demonstrating its superiority. The ANOVA test results confirm the statistical significance of these improvements ($p < 0.05$), validating that QSI-Fuzzy significantly outperforms the benchmark algorithms in translation software evaluation. Collectively, these results validate QSI-Fuzzy as a better solution for multi-domain multi-criteria evaluation, as it provides unparalleled performance, efficiency, and reliability.

TABLE V CONVERGENCE METRICS OF QSI-FUZZY VS. GA, PSO, AND SA, HIGHLIGHTING ITERATIONS TO CONVERGENCE, COMPUTATIONAL TIME, AND EFFICIENCY IMPROVEMENTS

Algorithm	Iterations to Convergence	Computational Time (s)
QSI-Fuzzy	55	12.5
Genetic Algorithm (GA)	95	23.1
Particle Swarm Optimization (PSO)	75	18.5
Simulated Annealing (SA)	120	30.2

TABLE VI ROBUSTNESS ANALYSIS OF QSI-FUZZY VS. GA, PSO, AND SA UNDER INCREASING NOISE LEVELS, HIGHLIGHTING PERFORMANCE STABILITY AND RELATIVE LOSS

Noise Level (%)	QSI-Fuzzy	GA	PSO	SA
0	0.92	0.88	0.86	0.85
5	0.91	0.86	0.83	0.81
10	0.88	0.83	0.80	0.78
15	0.85	0.78	0.75	0.73
20	0.80	0.70	0.68	0.65

TABLE VII PERFORMANCE COMPARISON AND STATISTICAL SIGNIFICANCE OF QSI-FUZZY VS. GA, PSO, AND SA ACROSS EVALUATION METRICS

Algorithm	Fluency	Efficiency	User Satisfaction	F-Test	p-value
QSI-Fuzzy	0.89	0.91	0.9	9.85	0.002
GA	0.81	0.84	0.82	10.32	0.0018
PSO	0.79	0.82	0.8	8.76	0.0031
SA	0.76	0.78	0.77	11.12	0.0015

V. CONCLUSION

In this paper, we introduced the Quantum Swarm-Driven Fuzzy Evaluation Framework (QSI-Fuzzy), a novel approach to addressing the multi-domain, multi-criteria evaluation of English translation software. It provides a holistic, flexible, and interpretable framework by incorporating QSI in optimizing dynamic weight and fuzzy logic in handling uncertainties. This framework is then applied to the Legal, Medical, and Conversational domains, translating software whose weights were optimized according to the set criteria like semantic accuracy, fluency, efficiency, and user satisfaction.

The experimental results indicated that QSI-Fuzzy outperformed all benchmark algorithms, including GA, PSO, and SA. Specifically, QSI-Fuzzy yielded higher scores across all translation software for significant improvement in semantic accuracy with fluency within domain operations, while it converged remarkably fast, reaching convergence at only an average of 55 iterations against up to 120 iterations by SA. Further, QSI-Fuzzy exhibited better robustness in the case of noisy conditions, keeping a performance score of 0.80 at 20% noise, while that for GA, PSO, and SA were 0.70, 0.68, and 0.65, respectively. These results confirm the efficacy of the proposed framework for solving the challenges in translation software evaluation and provide a scalable and efficient solution for multi-domain optimization problems. It advances

both methodologies of the translation evaluation and introduces the generally applicable approach that should easily extend to a greater scope of complex decision-making/optimization tasks. Accordingly, QSI-Fuzzy can be used in greater perspectives-real-time systems are not excluded, or other hybrid methodologies for increasing scalability/precision.

REFERENCES

[1] Papineni, K., Roukos, S., Ward, T. & Zhu, W. J., 2002. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318.

[2] Banerjee, S. & Lavie, A., 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. pp. 65–72.

[3] Narayan, A. & Sahoo, S., 2020. Quantum-inspired optimization algorithms: A survey. *Expert Systems with Applications*, 162, p.113732.

[4] Yan, Z., Tang, Y., Wang, H., Zhang, S., Liu, C. & Feng, Z., 2021. Applications of quantum swarm optimization in machine learning. *Neural Computing and Applications*, 33(10), pp.5051–5062.

[5] Liu, H., Zhou, Z., Zhou, W. & Wang, J., 2019. A hybrid quantum particle swarm optimization for image segmentation. *Journal of Visual Communication and Image Representation*, 60, pp.129–137.

[6] Pathak, S., Singh, P., Kumar, R., Kumar, S. & Sahu, R., 2020. Supply chain optimization using hybrid quantum-inspired algorithms. *Computers & Industrial Engineering*, 149, p.106770.

[7] Wang, X., Zhang, J., Li, Q., Zhou, L. & Yang, X., 2022. A quantum-inspired approach for medical decision-making under uncertainty. *Applied Soft Computing*, 116, p.108288.

[8] Zadeh, L. A., 1965. Fuzzy sets. *Information and Control*, 8(3), pp.338–353.

[9] Chatterjee, A., Joshi, A., Gupta, A. & Roy, A., 2019. Fuzzy logic applications in sentiment analysis: A review. *Knowledge-Based Systems*, 189, p.105121.

[10] Bahdanau, D., Cho, K. & Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[11] Koehn, P., 2005. Europarl: A parallel corpus for statistical machine translation. In: *Proceedings of the 10th Machine Translation Summit*. pp. 79–86.

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I., 2017. Attention is all you need. In: *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

[13] Mohiuddin, T. M. and Joty, S., 2020. Evaluating machine translation with deep contextual embeddings. *Transactions of the Association for Computational Linguistics*, 8, pp. 524

[14] Zhang, W., Wang, J. and Liu, X., 2021. A transformer-based evaluation framework for machine translation. *Neural Computing and Applications*, 33(10), pp. 6021\u20136038.

[15] Suresh, P., Kumar, A. and Singh, R., 2022. Quantum-enhanced swarm optimization for natural language processing tasks. *Computational Intelligence and Neuroscience*, 2022, p. 814263.

[16] Cai, J., Tang, H. and Zhao, L., 2020. Multi-objective quantum optimization algorithms and their applications. *Journal of Engineering Optimization*, 52(3), pp. 385.

[17] Zhou, Y., Liu, F., Zhang, H. and Li, X., 2020. A novel approach to decision-making in complex systems using quantum and fuzzy hybrid models. *Soft Computing*, 24(15), pp. 11471.

[18] Dash, R., Subudhi, S. and Dash, P. K., 2016. Fuzzy C-means clustering using hybridized glowworm swarm optimization. *Engineering Applications of Artificial Intelligence*, 49, pp. 134\u2013149.

[19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. In: *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

[20] Yang, S., Liu, X. and Huang, Y., 2022. Contextual evaluation of machine translation using attention-based models. *Artificial Intelligence Review*, 55(3), pp. 2347

[21] Tarik, S., Zafar, T. and Usmani, H., 2018. Quantum-inspired evolutionary algorithms for solving linguistic ambiguities. *Journal of Computational and Applied Mathematics*, 336, pp. 195\u2013208.

[22] Gupta, A., Mallick, S. and Roy, K., 2021. Evaluation of machine translation systems using linguistic intelligence. *Knowledge-Based Systems*, 225, p. 107102.

[23] Chatterjee, A., Joshi, A. and Gupta, P., 2020. Fuzzy decision-making frameworks for translation evaluation. *International Journal of Computational Linguistics*, 15(2), pp. 142

[24] Wang, X., Zhang, Y. and Li, R., 2021. Quantum and fuzzy hybrid models in natural language processing. *Applied Intelligence*, 51(5), pp. 3573\u20133590.

[25] Priyadarshini, I., 2024. Swarm-intelligence-based quantum-inspired optimization techniques for enhancing algorithmic efficiency and empirical assessment. *Quantum Machine Intelligence*, *6*(2), p.69.

[26] Hussain, S., Spratford, W., Goecke, R., Kotecha, K. and Jamwal, P.K., 2025. Deep learning-driven analysis of a six-bar mechanism for personalized gait rehabilitation. *Journal of Computing and Information Science in Engineering*, *25*, pp.011001-1.

[27] Khan, N.A., Sulaiman, M., Tavera Romero, C.A. and Alshammari, F.S., 2022. Analysis of nanofluid particles in a duct with thermal radiation by using an efficient metaheuristic-driven approach. *Nanomaterials*, *12*(4), p.637.

[28] Shi, K., Wu, Z., Jiang, B. and Karimi, H.R., 2023. Dynamic path planning of mobile robot based on improved simulated annealing algorithm. *Journal of the Franklin Institute*, *360*(6), pp.4378-4398.

[29] Khan, N.A., Goyal, T., Hussain, F., Jamwal, P.K. and Hussain, S., 2024. Transformer-Based Approach for Predicting Transactive Energy in Neurorehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Optimizing Athlete Workload Monitoring with Supervised Machine Learning for Running Surface Classification Using Inertial Sensors

WenBin Zhu[1], QianWei Zhang[2]*, SongYan Ni[3]
Chengdu Sport University, Chengdu, Sichuan, 610041, China[1, 2]
Sichuan University High School, No. 12 High School of Chengdu, Sichuan, 610061, China[3]

***Abstract*—Monitoring athlete movement is important to improve performance, reduce fatigue, and decrease the likelihood of injury. Advanced technologies, including computer vision and inertial sensors, have been widely explored in classifying sport-specific movements. Combining automated sports action labeling with athlete-monitoring data provides an effective approach to enhance workload analysis. Recent studies on categorizing sport-specific movements show a trend toward training and evaluation methods based on individual athletes, allowing models to capture unique features peculiar to each athlete. This is particularly beneficial for movements that exhibit large variations in technique between athletes. The current study uses supervised machine learning models, including Neural Networks and Support Vector Machines (SVM), to distinguish between running surfaces, namely, athletics track, hard sand, and soft sand, using features extracted from an upper-back inertial measurement unit (IMU) sensor. Principal Component Analysis (PCA) is applied for feature selection and dimensionality reduction, enhancing model efficiency and interpretability. Our results show that athlete-dependent training approaches considerably enhance the classification performance compared to athlete-independent approaches, achieving higher weighted average precision, recall, F1-score, and accuracy ($p < 0.05$).**

Keywords*—Athlete monitoring; machine learning models; running surface classification; Inertial Measurement Units (IMU); neural networks; Support Vector Machines (SVM); Principal Component Analysis (PCA)

I. Introduction

Supervised machine learning algorithms have emerged as flexible statistical tools capable of modeling both classification and regression problems. These algorithms use mathematical frameworks for model optimization that map input features to output targets of a training dataset to make accurate predictions on unseen data. Sports science has of late embraced the power of data-driven methodologies to track and improve the performance of athletes [1]. Of these, supervised machine learning and artificial neural networks have been at the forefront of automating sport-specific movement classification, quantifying athlete workload, and predicting physiological states-for example, fatigue-to optimize training regimens and reduce injury risks [2].

The introduction of MEMS has subsequently miniaturized IMUs into wearable devices that have revolutionized performance monitoring in elite sports by providing real-time, high-resolution data on athlete movements [3]. These normally have IMUs positioned at key areas around the body for the extraction of critical features that are important in the study of athletic performance in various sports. These feed into input-supervised machine learning models aimed at classifying specific sport movements or environmental contexts, such as the running surface [4]. However, these models are as good as their methodologies applied during the training and validation process. Data partitioning, whether athlete-dependent or athlete-independent, has a huge impact on model performance and generalizability [5].

Despite the progress made in applying machine learning to sports, recent systematic reviews have identified important shortcomings in the validation strategies. Studies tend to rely on non-independent data splits, such as cross-validation or simple train-test partitions, which often inflate classification performance. Conversely, leave-one-subject-out (LOSO) validation-a method that ensures athlete-independent evaluation-remains underutilized [6], [7]. This is a particularly important gap, as non-independent methods do not consider the inter-athlete variability that may result in models performing well on known data but poorly on new, unseen athletes. In sports applications, where the ability to adapt to new athletes is crucial, reliance on athlete-dependent validation methods risks compromising the broader applicability and reliability of machine learning solutions.

The aim of this work is to classify the running surfaces into athletics track, hard sand, and soft sand using the features computed from upper-back IMUs. To compare different approaches, six supervised machine learning models were trained and then evaluated using both athlete-dependent and athlete-independent methodologies. Although the former is generally more accurate because the model may learn features specific to a given individual, they do not generalize well for other unknown athletes. This work is, to the best of the authors' knowledge, one of the first sport-specific case studies directly comparing these methodologies in running surface classification. Its results are particularly relevant to sports organizations seeking to implement robust machine learning models in monitoring athletes: it conveys actionable insight regarding their training, validation, and deployment strategy. The current study has attempted to guide the development of more reliable and scalable solutions for sport-specific movement classification by underlining some of the trade-offs between accuracy and generalizability.

*Corresponding Author

The remainder of this paper is structured as follows: Section II presents a literature review, discussing relevant studies on athlete workload monitoring, running surface classification, and machine learning applications in sports analytics. Section III describes the methodology, detailing the data collection process, feature extraction from the inertial sensor, and the supervised machine learning models, including Neural Networks and Support Vector Machines (SVM), utilized for classification. Section IV presents the results, evaluating the classification performance of athlete-dependent and athlete-independent models based on multiple metrics. Finally, Section V concludes the study, summarizing key findings and highlighting potential future directions for improving athlete workload monitoring using advanced machine learning techniques.

II. Literature Review

Recently, a significant boost of machine learning applications in sports analytics was seen; thus, researchers started reaping full benefits of this technique while solving a number of problems connected with monitoring performance of athletes and classification of movements. Wearables, together with IMUs, have become an essential tool for motion data acquisition and further analysis of several features of athletic performances. Unlike traditional observation-based methods, IMUs are accurate and scalable; they can capture multidimensional data about an athlete's movements in real time. Works such as that by Umer and Riaz [8] show the capability of IMUs within gait analysis to identify ground contact events with high accuracy on different surfaces. It goes without saying that this type of research places increasing dependence on the use of IMU sensors within sports and rehabilitation applications [9].

Subsequently, machine learning models based on IMU data had very good performance, especially in the classification of environmental and movement context, such as running surfaces. Buckley et al. [10] presented a road surface type classification approach using IMU data, incorporating traditional machine learning algorithms like SVM and KNN with deep learning approaches. The results showed that machine learning could identify surface types with a high degree of accuracy, thus opening the way for possible applications in sports [11]. This study concerned transportation, but its implications reach to athletic performance, where running surface classification can improve workload monitoring and injury prevention.

One of the most critical issues when developing machine learning models to monitor athletes is the classification based on an athlete-dependent or independent approach. The athlete-dependent model is specific for particular features of a certain athlete, and this results in a higher accuracy if the test on the same subject is done. Most of these models fail to generalize when applied on different athletes. On the other hand, the athlete-independent model is general and permits variations in individual features. Koul et al. [10] discussed surface recognition regarding electric scooters using deep neural networks based on smartphone IMU sensors. While not directly related to running, their findings emphasize the importance of designing models that balance specificity and generalizability—principles that are highly relevant to sports performance monitoring.

The broader literature also underlines the increasing role of machine learning in sports injury prediction and prevention. Surveys such as Diss et al. [11] review various algorithms ranging from Random Forests to neural networks, using data derived from athletes in order to predict injury risks. Such studies indicate the potential of machine learning to analyze complex data sets and determine patterns associated with injury-prone conditions [12]. Though different from running surface classification, all these applications are unified in the aim of bettering athlete safety and optimizing performance using data-driven insights. The literature underscores the transformative potential of machine learning in sports analytics [13]. Although recent advancements in IMU-based models and validation methodologies have achieved higher classification performance, challenges still remain regarding how to balance accuracy and generalizability. The study contributes to the field by comparing the athlete-dependent and independent approaches to classify running surfaces, filling key gaps in existing research and informing future model development and deployment.

III. Methodology

A. Participants

Seven healthy subjects, four males and three females, participated voluntarily in this study and gave their informed consent. The group's mean age was 32.4 years, with a standard deviation of 17.89 years, which shows the very high variability in age among the group members. Their mean height was 171.9 cm, with a standard deviation of 8.91 cm, and their mean weight was 70.3 kg with a standard deviation of 16.87 kg. Ethical approval for this study was granted under protocol number GU 2017/587. The population of the subjects was heterogeneous regarding their fitness level and running experiences; consequently, it constituted a rich sample for the study's objectives. Their training routines varied, with some individuals reportedly spending up to nine hours a week training.

B. Experimental Design

This experiment consisted of running 400 meters at a light to moderate pace on three different surfaces: first, on soft, dry sand; then, on hard, water-saturated sand; and finally, the same on a synthetic tartan running track. This completed the trials for all surface conditions. Each run was designed to maintain consistency in pace and effort across the surfaces. Data of the motion were captured with one IMU per participant. The IMU was positioned near the third thoracic vertebra, T3, and fixed with a specifically developed sports harness not allowing any displacement during the runs. This setup made the sensor stable, and there was no interference with the data collection process [14], [15]. Fig. 1 depicts the orientation of sensor axes, which is important for accurate motion analysis. The figure shows that data acquisition was uniform across all participants and conditions.

C. IMU Sensor Technology

The device used was a custom-made, 9DOF IMU, designed at Chengdu Sport University. For this present research, it was based on the unit known as SABELSense (Sichuan, China) with a weight of 23 g and specified as +16 g accelerometer, +2000 deg/s gyroscope, and +7 Gauss magnetometer; all data was captured at a sampling frequency of 250 Hz. Each IMU output was then comprehensively calibrated before the trial to capture proper data. The data were logged locally onto a 4GB microSD card that enabled continuous, reliable logging of the experiment. 3D orientation of the sensor is obtained using Euler angles: roll, pitch, and yaw through the Madgwick AHRS algorithm. This has an accuracy characterized with a root mean square error below 0.8° for static, and below 1.7° in dynamics. This setup ensured that the motion data was valid and reliable during the study.

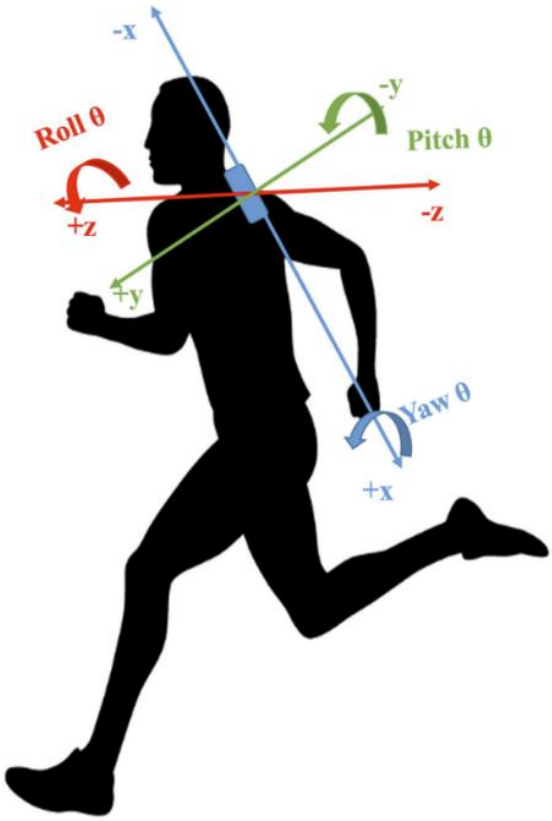


Fig. 1. The x-axis is the superior-inferior direction, the y-axis is the medial-lateral direction, and the z-axis is the anterior-posterior direction. The rotation around the z, y, and x axes corresponds to the roll, pitch, and yaw, respectively.

D. Designed Algorithm

1) Feature extraction: When running on a 400-metre athletics track, due to the curviness of the path, the Euler angles recorded drift progressively. In order to overcome this problem, a feature extraction method had been developed and was very robust, being inspired by different previous methodologies in which [16] is highly remarkable. By using a sliding window technique, the whole process was executed with the assistance of MATLAB from MathWorks, Natick, MA, USA:. This window was set to a duration of 4 seconds, while the overlap between two successive windows was set to 0.5 seconds. This would give approximately 10 to 11 strides on each surface, providing a reliable dataset and at the same time, reduce the effects of directional drift [17]. Normalization of Euler angle data and transforming into absolute values was done to nullify the effect of heading drift within each window. Too much spurious data was removed, which would have had an adverse effect on the classification. The above window-sliding procedure was applied for 11 data channels recorded by the IMU: acceleration components, gyroscope outputs, and orientation angles in three dimensions. For every one of these data windows, various features were extracted in time and frequency domains. These included the mean values, standard deviation, skewness, kurtosis, and dominant frequency components. These features together gave a full representation of the pattern of movement-a basis on which effective classification and analysis could be done.

2) Training-validation of feature data: The feature data was divided into training and validation sets using two different strategies. First, an athlete-independent LOSON was used: one participant was randomly selected for model evaluation, and the remaining six participants' data was used for training. In the second strategy, the data was divided in an athlete-dependent way, where 75% of the data was used for training and 25% for testing. These methods were applied to evaluate how individual participant features influenced the classification performance of the models. In the athlete-independent approach, Method 1, the number of training observations for soft sand, hard sand, and the athletics track were 1537, 1237, and 944, respectively. The respective test observations for the considered surfaces were 183, 153, and 196. By contrary, the approach dependent on athletes-Method 2 resulted in 1720, 1390, and 1140 training observations for soft sand, hard sand, and athletics track surface, respectively, leaving 413, 341, and 309 observations for testing. These two partitioning strategies have allowed a more complete assessment of the model for its ability both to generalize across individuals and to perform when fit to specific athletes.

3) Feature engineering: Feature engineering and model training were performed on Python, Python Software Foundation, https://www.python.org/ using popular libraries like scikit-learn and pandas [18, 19]. All the features were scaled into a uniform range from 0 to 1 using the mean and standard deviation of the training dataset before modeling. This way, the features were normalized, and no single feature biased the model due to its magnitude. The challenge in high dimensionality was approached by Principal Component Analysis (PCA) [20]. In this process, PCA transforms the original features into a new orthogonal set of variables known as principal components. Every principal component carries part of the dataset variance, and only those components needed to describe 95% of the total variance were retained for this study. This reduced the number of features from 132 to 45, thus

significantly reducing computational complexity and enhancing model efficiency. By removing noisy and redundant features, PCA also helped reduce overfitting and enhanced the generalization capability of the model.

Among the unsupervised dimensionality reduction techniques, PCA was preferred over supervised ones such as LDA because of its advantages in cases with limited training samples. Unlike these supervised methods, PCA does not depend on class labels and hence avoids the bias toward a particular subset of data. Thus, PCA is particularly suitable for the comparisons among the athlete-independent and the dependent methods. Even though LDA is originally designed to maximize class separability, it could amplify overfitting in the case of limited training data, an important concern in this study. Moreover, previous studies have demonstrated that PCA performs better than LDA when sample sizes per class are small, which further supports the appropriateness of the choice for this study [21] [22]. By applying PCA, the research was able to balance the computational efficiency with feature relevance; hence, the model can process meaningful information without getting overwhelmed by noise or irrelevant data. This embedding had not only reduced the computational burden and reduced training time but also made it a just comparison between the athlete-independent and athlete-dependent methodologies during the conduct of the research, making PCA an essential part of the feature engineering pipeline.

4) Model training and evaluation: Six different machine learning models were developed and tested to classify sport-specific movements using data from inertial sensors. These include some of the most commonly used movement classification models: logistic regression (LR), support vector machines (SVM) with linear (LSVM) and Gaussian radial basis function (GSVM) kernels, multilayer perceptron neural networks (MLP-NN), random forests (RF), and gradient boosting machines (XGB) [23] [24]. Model configurations were selected without hyperparameter tuning in order to provide a baseline for comparisons. Logistic Regression models relied on an L2 penalty while using the lbfgs solver. Support vector machines consisted of a linear kernel, with C = 1 and a Gaussian kernel, with C = 1 and gamma = scale. The neural network, MLP, consisted of three layers of 8 nodes, ReLU activation, constant learning rate, and Adam optimizer. The random forest model was set with the Gini criterion for impurity, number of features as the maximum feature parameter, and included 20 estimators. The gradient boosting model used the Friedman mean squared error (mse) criterion, deviance loss, a maximum depth of 3, and 100 estimators.

The models were then evaluated in their classification of running surfaces using both athlete-dependent and athlete-independent training and validation segmentation methods. The performance metrics for each classification technique included weighted averages of precision, recall, and F1-score and the overall accuracy for classification. The statistical comparisons between models that have used two segmentation methods employed a paired t-test, $\alpha = 0.05$ as shown in Fig. 2.

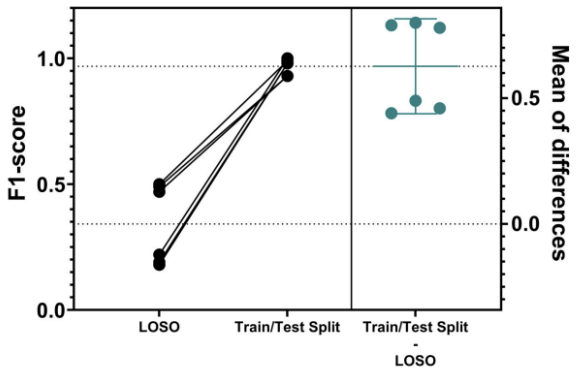


Fig. 2. Estimation plot showing the significant difference in F1-scores across all models comparing the train/test split to the LOSO validation.

IV. Results

The statistical results of the test on the train/test split provided significantly higher values for all model types with respect to the predictive performance measure. More concretely, these are huge increases in the weighted averages for precision, recall, F1-score, and accuracy, with p-values 0.0002, 0.0004, 0.0004, and 0.0004, respectively. This result points once more to the importance of letting the models see all participant features during training and hence letting them capture the variabilities in individual movement patterns. Results indicate large differences between F1-scores from the two validation methods; Fig. 2 provides further visualization, whereas a detailed comparison of various evaluation metrics for models considering both two validation methods can be seen in Fig. 3(a) – Fig. 3(d).

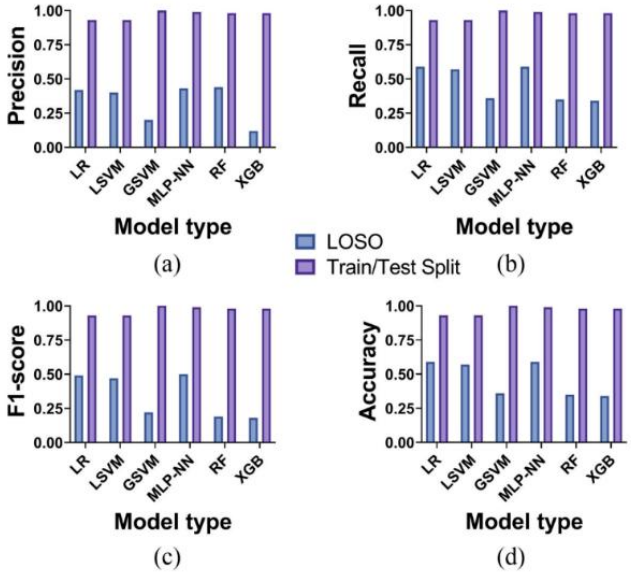


Fig. 3. Comparison of evaluation metrics of all models, considering both the train / test split and LOSO validation methods: (a) weighted precision, (b) weighted recall, (c) weighted F1-score, and (d) overall classification accuracy.

As observed by the method of train/test split, using training based on participant-specific features pays off significantly in

the case of all the models. Such exposure enables them to learn complex patterns that are unique in different running styles and surface interactions, thus giving enhanced accuracy of classification. However, to generalize this application into generalized surface classification-for instance, across diverse populations or settings-the participants in the dataset would have to be increased in number for broader applicability. Among the models tested, the best performances were from the MLP-NN under the LOSO and the GSVM under the train/test split. Their classification capabilities are described by confusion matrices shown in Fig. 4 and Fig. 5. The MLP-NN model was in a fairly medium range when segmenting soft sand from the remaining two surface classes using the LOSO method. The precision by classification is 0.71, recall 0.89, and F1-score 0.79 for the soft sand class; hence, one can say it identified the features of running on the soft sand rather successfully. This agrees with highly observable changes in gait mechanics when walking on soft sand compared to harder surfaces. This is expected since the closer physical properties result in more misclassifications between these two surface types of hard sand and athletics track surfaces. On the other hand, the GSVM model, when tested by the method of train/test split, it gave a general accuracy of 0.99. Actually, the model's performance was really good on all surfaces, misclassifying only once between the athletics track and hard sand surface. This, therefore, ascertains how the GSVM model can handle such subtlety in running pattern variations across surface types. With a full presentation of all information during training that prepares this model to generalize best on all test datasets, even then, an optimum quotient from GSVM with a train/test split may be expected.

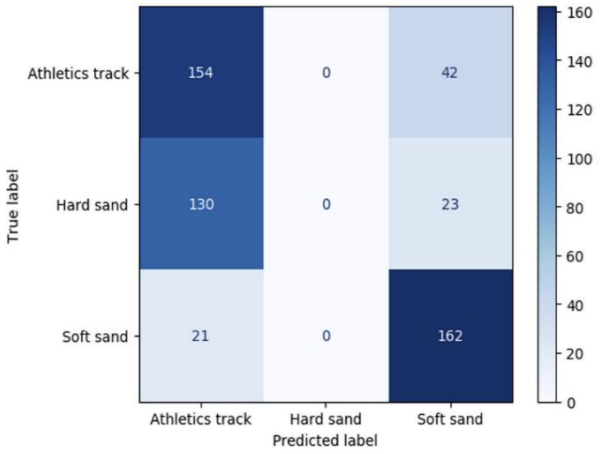


Fig. 4. Confusion matrix illustrating the classification performance of the MLP-NN model using the LOSO validation method.

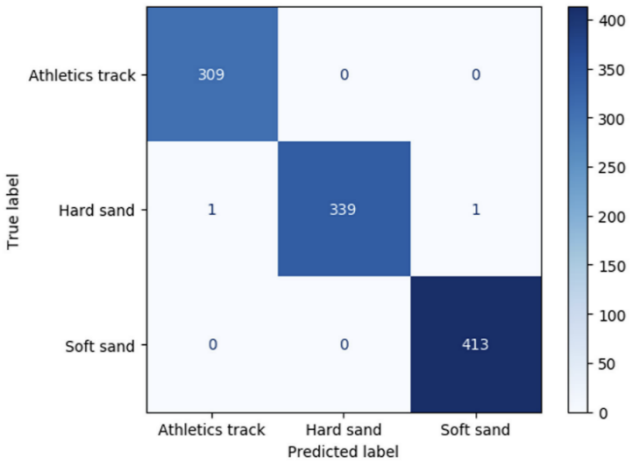


Fig. 5. Confusion matrix showcasing the classification performance of the GSVM model using the train / test split validation method.

These results have confirmed the intuition that a specific relationship exists between certain forms of validations pursued on classifications' various results. This was especially evident from the train/test split, representing strengths in model accuracy improvements using participant-specific features. In tasks where such information is available, this might prove very effective. However, the LOSO approach could be more fitting when, at application time, generalization to unseen individuals with one universal model is required. The findings from this study underpin some critical trade-offs between accuracy and generalizability for the classification of athlete movement and provide valuable insights into the development and implementation of machine learning models in sports analytics.

V. Conclusion

This study underlines, with greater significance, the improved classification performance that is attained with athlete-dependent train/test split methods ($p < 0.05$). Individual differences in the execution of movements are key to monitoring athletes, and allowing models to learn such specific features significantly enhances the accuracy of classification. A generalized sport-action classification model should have high performance on completely independent athletes; hence, it requires training data from a diverse group of individuals. This would include the type, body of the cyclists (height and weight), standard, physical fitness of the cyclists. However, due to issues of privacy, hardly any athlete performance data can be provided, making the creation of a fully athlete-independent model very challenging. An issue that gets very crucial when there is high individual variation in the styles of movement is, for instance, while running on different surfaces. Given these limitations, any sporting organization looking to utilize

automated tagging of sport-specific actions as a way of supplementing current approaches to athlete-load monitoring would have to, at the very least, retrain the classification models on data from all participating athletes. This can also include a calibration session when new athletes join in, allowing the model to learn features from the new person. This proposed approach using an upper-back IMU sensor for running surface classification may, therefore, inherently be an athlete-dependent one. The proposal would still be very useful. Besides, it also holds prospect for adequate adjustment of an athlete's session work rate estimate, particularly on occasions when some direct physiological monitoring implements, like heart rate monitors, cannot be used.

In future work, the authors aim to develop privacy-preserving methodologies, such as federated learning, to facilitate athlete-independent classification models while ensuring data security. Additionally, they intend to integrate multi-modal sensor fusion techniques to enhance the robustness and generalizability of movement classification across various sports activities.

REFERENCES

[1] Cust, E.E., Sweeting, A.J., Ball, K. & Robertson, S., 2019. Machine and deep learning for sport-specific movement recognition: A systematic review of model development and performance. Journal of Sports Sciences, 37(5), pp.568–600.

[2] McGrath, J., Neville, J., Stewart, T. & Cronin, J., 2020. Upper body activity classification using an inertial measurement unit in court and field-based sports: A systematic review. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*. [Online] Available at: https://doi.org/10.1177/1754337120959754.

[3] Eyobu, O.S. & Han, D., 2018. Feature representation and data augmentation for human activity classification based on wearable IMU sensor data using a deep LSTM neural network. *Sensors*, 18(9), pp.1–36.

[4] Wan, S., Qi, L., Xu, X., Tong, C. & Gu, Z., 2020. Deep learning models for real-time human activity recognition with smartphones. *Mobile Networks and Applications*, 25(2), pp.743–755.

[5] Gao, Z., Xuan, H.Z., Zhang, H., Wan, S. & Choo, K.K.R., 2019. Adaptive fusion and category-level dictionary learning model for multiview human action recognition. *IEEE Internet of Things Journal*, 6(6), pp.9280–9293.

[6] Dixon, P.C., 2019. Machine learning algorithms can classify outdoor terrain types during running using accelerometry data. *Gait and Posture*, 74, pp.176–181.

[7] Einicke, G.A., Sabti, H.A., Thiel, D.V. & Fernandez, M., 2018. Maximum-entropy-age selection of features for classifying changes in knee and ankle dynamics during running. *IEEE Journal of Biomedical and Health Informatics*, 22(4), pp.1097–1103.

[8] Buckley, C. et al., 2017. Binary classification of running fatigue using a single inertial measurement unit. *Proceedings of IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks*, pp.197–201.

[9] Khan, N.A., Hussain, S., Spratford, W., Goecke, R., Kotecha, K. and Jamwal, P.K., 2025. Deep learning-driven analysis of a six-bar mechanism for personalized gait rehabilitation. *Journal of Computing and Information Science in Engineering*, *25*(1).

[10] Koul, A., Becchio, C. & Cavallo, A., 2018. Cross-validation approaches for replicability in psychology. *Frontiers in Psychology*. [Online] Available at: https://doi.org/10.3389/fpsyg.2018.01117.

[11] Diss, C.E., 2001. The reliability of kinetic and kinematic variables used to analyse normal running gait. *Gait and Posture*, 14(2), pp.98–103.

[12] Yam, C.Y., Nixon, M.S. & Carter, J.N., 2002. On the relationship of human walking and running: Automatic person identification by gait. *Proceedings of Object Recognition Supported by User Interaction for Service Robots*, pp.287–290.

[13] Khandelwal, S. & Wickström, N., 2017. Evaluation of the performance of accelerometer-based gait event detection algorithms in different real-world scenarios using the MAREA gait database. *Gait and Posture*, 51, pp.84–90.

[14] Espinosa, H.G., Shepherd, J.B., Thiel, D.V. & Worsey, M.T.O., 2019. Anytime, anywhere! Inertial sensors monitor sports performance. *IEEE Potentials*, 38(3), pp.11–16.

[15] Khan, N.A., Goyal, T., Hussain, F., Jamwal, P.K. and Hussain, S., 2024. Transformer-Based Approach for Predicting Transactive Energy in Neurorehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

[16] Camomilla, V. et al., 2018. Trends supporting the in-field use of wearable inertial sensors for sport performance evaluation: A systematic review. *Sensors*, 18(3), p.873.

[17] Shepherd, J.B., Thiel, D.V. & Espinosa, H.G., 2017. Evaluating the use of inertial-magnetic sensors to assess fatigue in boxing during intensive training. *IEEE Sensors Letters*, 1(2), p.6000104.

[18] Thiel, D.V., 2020. Predicting ground reaction forces in sprint running using a shank mounted inertial measurement unit. *Proceedings of MDPI Sensors*.

[19] Lai, A.D.A., James, D.P., Hayes, P. & Harvey, E.C., 2004. Semi-automatic calibration technique using six inertial frames of reference. *Proceedings of SPIE Microelectronics Design, Technology and Packaging*, 5274, pp.531–542.

[20] Madgwick, S.O.H., Harrison, A.J.L. & Vaidyanathan, R., 2011. Estimation of IMU and MARG orientation using a gradient descent algorithm. *Proceedings of IEEE International Conference on Rehabilitation Robotics*, pp.1–7.

[21] Khan, N.A., Sulaiman, M., Tavera Romero, C.A. and Alarfaj, F.K., 2021. Numerical analysis of electrohydrodynamic flow in a circular cylindrical conduit by using neuro evolutionary technique. *Energies*, *14*(22), p.7774.

[22] Worsey, M.T.O. et al., 2020. An evaluation of wearable inertial sensor configuration and supervised machine learning models for automatic punch classification in boxing. *IoT*, 1(2), pp.360–381.

[23] Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, pp.2825–2830.

[24] McKinney, W., 2010. Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, pp.56–61.

LDA-Based Topic Mining for Unveiling the Outstanding Universal Value of Solo Keroncong Music as an Intangible Cultural Heritage of UNESCO

Denik Iswardani Witarti[1]*, Danis Sugiyanto[2], Atik Ariesta[3], Pipin Farida Ariyani[4], Rusdah[5]
Faculty of Communication and Creative Design, Universitas Budi Luhur, Jakarta, Indonesia[1]
Faculty of Performing Art, Institut Seni Indonesia Surakarta, Surakarta, Indonesia[2]
Faculty of Information Technology, Universitas Budi Luhur, Jakarta, Indonesia[3, 4, 5]

***Abstract*—Outstanding Universal Value (OUV) is an essential value of culture and nature. It is so extraordinary that it transcends national boundaries and becomes generally crucial for all humanity's current and future generations. A culture with this value needs permanent protection because it is considered a critical heritage for the world community. Solo keroncong music, as one of the local wisdom owned by the Indonesian nation, has yet to be recognized as one of the UNESCO Intangible Cultural Heritage (ICH). It has even become an instrument of Indonesia's soft power diplomacy in several countries, such as Malaysia, England, and the United States. It must be of OUV and meet at least one of ten selection criteria to be included on the World Heritage List. This study explored the OUV of Solo keroncong music using Latent Dirichlet Allocation. The primary data were obtained by conducting an FGD with the Indonesian Keroncong Music Artist Community (KAMKI) Surakarta and in-depth interviews with several keroncong figures in Solo. The result showed there are four topics with a coherent score of 0.51. Then, the expert mapped those four topics into three OUVs of Solo keroncong music as temporary findings. Keroncong music is a masterpiece of human creativity, a witness to civilization, and has traditional values. These findings showed that Solo keroncong music is worthy of being proposed as one of the UNESCO ICH.**

Keywords*—LDA; OUV; Solo keroncong; text mining; topic modeling

I. Introduction

The United Nations Educational, Scientific and Cultural Organization (UNESCO) is a unique agency in the United Nations (UN) that deals with education, science and culture. UNESCO programs are divided into five main sectors: Education, Natural Sciences, Social Sciences, Culture, and Communication and Information [1]. Indonesia ratified the Convention for the Safeguarding of Intangible Cultural Heritage issued by UNESCO in 2003, which was then ratified in Presidential Regulation 78 of 2007. Consequently, Indonesia must protect its cultural wealth. Indonesia is required to record cultural works as a protection effort. The Directorate of Cultural Heritage and Diplomacy will determine and then grant status as Warisan Budaya Tak Benda (WBTB) or Intangible Cultural Heritage (ICH) by the minister based on the recommendations of the team of experts formed. According to the 2003 UNESCO Convention, what is included in Intangible Cultural Works are Traditions and oral expressions, including language as a vehicle for intangible cultural heritage; Performing arts; Customs, rituals, and celebrations; Knowledge and behavioral habits regarding nature and the universe; Traditional craft skills [2], [3]. Music and song are present in 304 of the 584 elements on the list (52%), referring to music alone or combined with other dimensions such as dance and poetry [2].

Keroncong music is one of Indonesia's cultural arts riches that still exists in the era of the development of the modern music industry. During the Dutch era, Keroncong Tugu, the origin of Indonesian keroncong music, was famous for entertainment in Batavia. Keroncong music spread widely and blended with gamelan, the primary musical culture in Java at that time. Combining keroncong musical instruments with the musicality of gamelan later became the characteristic of Solo keroncong music. The music played is more relaxed (nglaras in Javanese) than Keroncong Tugu, which developed on the coast. The Indonesian Keroncong Music Artist Community (KAMKI) of the Surakarta City Branch Representative Council (DPC Surakarta) noted that 60 community members were actively preserving keroncong music in Solo City [4]. The strains of Solo keroncong have also succeeded in attracting the interest of music lovers from abroad. Unfortunately, the beauty of the Solo keroncong music has not been included in the UNESCO ICH list. UNESCO registration will also impact Indonesia's positive image as a nation preserving cultural wealth. International recognition of keroncong music will also increase the appeal of cultural tourism in Solo. Foreign tourists will be interested in coming to Indonesia and directly enjoying the musical strains of Solo keroncong music and other local cultures.

Based on the background above, this study was designed to explore the feasibility of keroncong music as one of the ICH by exploring the Outstanding Universal Value (OUV) contained in Solo keroncong music. OUV is an essential value of a culture or nature that is so extraordinary that it transcends national boundaries and becomes generally crucial for all humanity's current and future generations.

Exploring the OUV of Solo Keroncong Music used ten criteria based on the Operational Guidelines for implementing the World Heritage Convention (OGIWHC) [1]. The research involved KAMKI Surakarta in gaining in-depth knowledge about Solo Keroncong Music. The results of discussions and interviews were processed into texts for use in LDA modeling.

The modeling results in several topics about the main values of Solo keroncong music. These main topics' values were evaluated based on OGIWHC to see whether Solo keroncong music meets the OUV criteria as a requirement for UNESCO ICH determination.

This study is organized into five sections. The first section briefly describes Keroncong music, the background to the problem area, the objective of the study, and the significance of the research output. The second section deals with previous studies on soft power diplomacy and topic modeling using LDA in terms of ICH. The third section discusses the methodologies used to conduct this study. The fourth section is about the analysis and results of this study. Finally, the last section discusses the findings and recommendations for future work.

II. Related Work

A. Soft Power Diplomacy

This study is anchored in the concept of soft power diplomacy, a framework developed by Joseph Nye. It posits that culture, politics, and foreign policy are the three pivotal sources of soft power. Effectively utilizing these sources is instrumental in determining diplomacy's success [5]. Soft power diplomacy is a strategy to influence foreign public opinion and behavior by utilizing the appeal of culture, values, and inclusive foreign policies. The implementation of soft power diplomacy pursues the goal of creating a positive image of the country and expanding its influence in the international world. A country's reputation and how the target public accepts it will determine the success of soft power diplomacy. The appeal and influence result from a social process involving both parties, so the effects of soft power will only occur if there is a mutually impactful relationship between the two parties [6].

Researchers have recently conducted studies using soft power to implement Indonesian diplomacy. Research on the peaceful message of the XX Papua National Sports Week (PON) was carried out to highlight sports as one of the instruments of soft power diplomacy [7]. The study results show the success of the government's soft power diplomacy in warding off security issues in Papua through the success of PON XX. The researcher also conducted a sentiment analysis of the news of the K-Pop concert on the 50th Anniversary of bilateral relations between Indonesia and South Korea. The positive sentiment of the Indonesian public explains that music can be an instrument of soft power diplomacy in harmonizing bilateral relations between Indonesia and South Korea [8]. Currently, the researcher is conducting research (ongoing) on the potential of keroncong music as cultural diplomacy between Indonesia and Malaysia [9]. Initial results found the public's love abroad, especially in Malaysia, for the beauty of the Solo keroncong music [9]. This is the background for the researcher to develop in a new study to find the primary value of Solo keroncong music. The researcher has conducted a study on the evolution of keroncong music in Richmond, Virginia. The Rumput Orchestra is an example of the rapid development of keroncong music outside Indonesia. This orchestra creates new music by combining various Keroncong music cultures, such as Javanese, Balinese, Sundanese, Appalachian, and Irish [10]. The research also involved the keroncong community in Solo City as one of the determinants of the success of Indonesia's soft power diplomacy. The involvement of non-state actors is an effective tool for implementing soft power more efficiently [11].

B. Topic Modeling using LDA in Terms of ICH

Reference [12] stated that the development of new media has enabled intangible cultural heritage to be disseminated through online platforms and attracted the attention of many contemporary young people. Classification and discussion on the value of intangible cultural heritage were essential ways to help with inheritance and dissemination. Real online reviews were collected using the Bilibili website as the research data source. A text-based BiGRU-Attention model was conducted to achieve value recognition and classification, and keyword statistics and topic analysis were performed for a score of more than 77%. The Cultural Value Perception (CVP) category has the best classification performance. Through the topic analysis of comments and keywords, the cultural value of intangible cultural heritage is its core connotation, social value is the primary purpose, and economic value is the power source [12].

Meanwhile, based on cultural identity and sovereignty, intangible cultural heritages (ICHs) are disappearing at an alarming rate and facing an existential crisis [13]. With the emergence of neural network models, the development of digital technology has revived many things that were on the verge of extinction. Traditional cultures and industries that initially seemed unrelated can take on new forms with the help of digital technology, thus enabling ICHs to find new ideas for development. The study took Huizhou ICH as an example and tried to design and construct a Huizhou ICH database and a digital map of Huizhou ICH, establishing a database for management and operation. The paper applies information space theory to study the use of ICH's digital resources under the threshold of the neural network. It employed digital information technology to recode, reconstruct, and interpret ICH. As a result, traditional ICH items were displayed digitally, improving the public's recognition of digital ICH items, thereby promoting the inheritance and dissemination of ICH [13].

The study in [14] designed and constructed a knowledge ontology framework for sports intangible cultural heritage (ICH) resources to support their preservation and inheritance by integrating and mining sports ICH resources. The study collected multiple data on sports ICH from various data sources and constructed ICH knowledge ontology using the CIDOC CRM metadata reference model and seven-step method. To enrich the content of the ontology, the TextRank algorithm was used to extract critical textual information and design a domain-specific NER model for sports NRL. In addition, the study adopted the VSM vector space model for text representation and used a proven hierarchical classification model for text categorization to improve classification accuracy. The study also explores the similarity calculation of concepts in the sports NRM ontology and proposes a semantic similarity calculation formula based on the ontology concepts. Respondents' willingness to pay was investigated through the conditional value method (CVM) to assess the value of sports NRM tourism resources. Finally, the factors influencing

respondents' willingness to pay were analyzed using statistical analysis and a logistic regression model, and it was found that they were mainly influenced by the degree of understanding of sports non-heritage resources and the level of education. The results of this study not only provide theoretical and methodological support for the effective integration and excavation of sports non-heritage resources and a new perspective on their protection, inheritance, and sustainable development [14].

ICH has important historical, cultural, and spiritual values. The study in [15] used the intangible cultural heritage of Yunnan Painting, the national architectural decoration of Dali, as the case study object. They crawled the consumer's comment text data to establish the Yunnan Painting consumer comment text dataset. Secondly, using the K-means method, they clustered high-frequency words from the dataset using text mining technology, TF-IDF keyword extraction, and LDA themes. Finally, they evaluated the economic potential of Yunnan Painting in terms of consumers' cognitive abilities and the emotional appeal of its products based on the extracted results. The study revealed that the dressmaking, painting, and tourism industries were the main economic activities of Yunnan color painting, with the tourism industry requiring further strengthening. According to the text mining of high-frequency words, the frequency of "product" in Yunnan painting was the highest at 296 times, followed by "culture," "technology," "cultural creativity," and "theme." Analyzing consumers' emotional text features reveals that most consumers have positive emotions towards Yunnan-colored paintings, while very few have negative emotions. The proposed economic optimization strategy is based on the results and provides a reference value for the economic development of other intangible cultural heritage [15].

The study in [16] conducted a paper's topic-tracking research based on the topic model. Firstly, the LDA model was used to extract the topic information from the news texts of different time windows. Then, the improved Single-Pass algorithm was used for topic tracking, in which the time decay function and the JS divergence were used to measure the similarity between the topics. Finally, the content and strength of the topics were analyzed for the results of topic tracking. They found that the topics discovered by the LDA model were more reliable than the k-means clustering in topic recognition. For topic tracking, the perplexity degree determined the optimal number of topics in the time window [16].

The combination of machine learning-driven topic modeling using Latent Dirichlet Allocation (LDA) and network analysis techniques examined a corpus of Korean and Japanese research papers on ICH [17]. LDA topic modeling identified three primary themes: technology and ICH, safeguarding ICH, and methodologies and approaches in ICH research.

Previous studies have been conducted related to analyzing [12] and exploring the value of ICH [14], [15] and its digitalization [13]. Unfortunately, no one discussed topic modeling to find UNESCO's Outstanding Universal Values of those ICHs. Topic detection or topic modeling has been conducted using LDA [16] and machine learning [17]. In terms of the dataset, this study used the primary data as the result of the FGD process, while the previous study used secondary data by crawling [15].

III. Methodology

This study explored the OUV of Solo keroncong music, which has not been explored before. Exploratory research helps develop a phenomenon that is not widely known or has little information [18]. Exploratory methods allow researchers to understand complex social phenomena more deeply before formulating specific research questions [19]. Fig. 1 shows the proposed methodology used in this study.

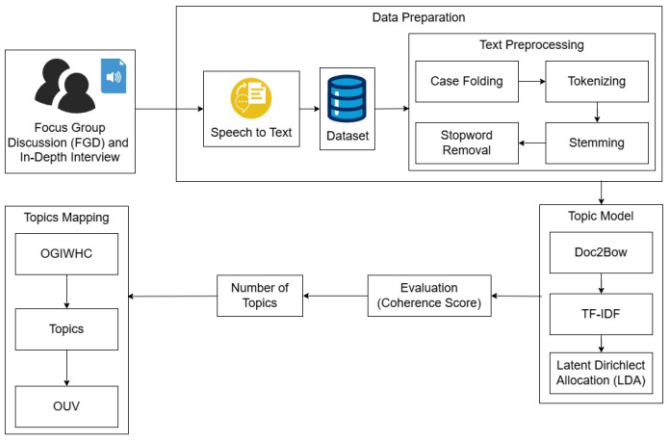


Fig. 1. Proposed methodology.

A. Data Collecting

The first step is to collect primary data. The study used primary data from conducting Focus Group Discussion (FGD) with the Surakarta Keroncong Music Artist Community (KAMKI) and in-depth interviews with several keroncong music figures and activists in Solo. FGD is a structured group discussion with various perspectives or experiences related to a particular research topic [20]. In-depth interviews are qualitative data collection methods that allow researchers to understand individuals' experiences, views, and perspectives related to the research topic [21]. FGD involved 40 people, and in-depth interviews involved five informants. They discussed the main values of keroncong music so that it can be recognized as UNESCO ICH. Initial data were obtained through audio recordings of FGD and in-depth interviews. There were six files of audio recordings.

B. Data Preparation

The recording results were then converted into text using TurboScribe.ai to create a dataset. The text obtained from TurboScribe.ai was then corrected for formats such as spaces, new lines, and inappropriate words. The informants verified the transcript to ensure the validity of the FGD and in-depth interview results. After the transcript (dataset) is verified, the following step is to prepare the data. The dataset contains eight documents. Text processing consists of case folding, tokenizing, stemming, and stopword removal. Case folding is the activity of removing special characters and changing them to lowercase. Tokenizing breaks down text based on spaces into a word (token). Stemming maps each token into a basic word form for nouns, verbs, and adjectives [22]. Stopword

removal is the activity of removing words that refer to words in a language (this study uses the Indonesian language) that are grammatically irrelevant to the content of the text, such as "*di*", "*yang*", "*dan*", "*maka*", and others [22]. The Sastrawi Python library implements the Nazief and Adriani algorithms used during the stemming and stopword removal process because it provides good results for the stemming process of Indonesian-language documents [23]. The results of this phase are transcripts datasets of the text on the FGD activities and in-depth interviews.

C. Modeling

Modeling begins by determining the corpus (collection of documents) using the Gensim Doc2Bow library. Then, text weighting and topic modeling follow.

Term (or text) weighting refers to calculating and determining weights. Term Frequency-Inverse Document Frequency (TF-IDF) is the most popular scheme in text weighting. The weight of words calculated by the TF-IDF scheme is proportional to the frequency of their occurrence in a particular text but inversely proportional to the frequency in other texts [22]. TF-IDF is an unsupervised statistical algorithm commonly used in information retrieval and text mining to assess the importance of words in a particular document in a corpus [15].

TF, or word frequency, represents the number of times a word appears in a given document, and this number is usually normalized to avoid TF bias towards long documents. Defining any word as T_i, the TF of the word T_i is formulated as [15]:

$$TF_{i,j}=\frac{N_{i,j}}{\sum_k N_{k,j}} \quad (1)$$

$N_{i,j}$ is the number of times the word T_i appears in a document D_j, and $\sum_k N_{k,j}$ is the total number of times all words appear in a document D_j.

IDF is the inverse document frequency, representing the distribution of documents containing a particular word in the corpus. The IDF of the word is formulated as [15]:

$$IDF_i=\log\frac{|D|}{1+|\{j:T_i\in D_j\}|} \quad (2)$$

$|D|$ is the total number of documents in the corpus, $1+|\{j:T_i \in D_j\}|$ is the number of documents containing the word T_i 1 is added to avoid the denominator being 0 (i.e., all documents do not contain the word).

A statistical-based technique is used to assess the importance of a word in a document to the corpus, namely Term Frequency-Inverse Document Frequency (TF-IDF). This technique calculates the weight of a word by measuring the frequency of its appearance in the document and its relevance to all documents in the corpus [22]. The calculation of TF-IDF is shown in (3). The larger the TF, the smaller the IDF is. In other words, the more times a word appears in a document, the fewer times the word appears in the corpus. The higher the value of TF-IDF, the more important the word is to this article and the more it can represent the article. TF-IDF is calculated as [15]:

$$TFIDF_{i,j}= TF_{i,j} \text{ x } IDF_i = \frac{N_{i,j}}{\sum_k N_{k,j}} \text{ x } \log\frac{|D|}{1+|\{j:T_i\in D_i\}|} \quad (3)$$

The results of TF-IDF weighting are in the form of a term-document matrix (TDM) and a list of essential words with the highest weight in each document. TF-IDF weighting is based on the corpus.

The next step is topic modeling. Topic modeling is one of the most widely used techniques in the unsupervised machine-learning field [24], [25], [26], [27]. According to [24], topic modeling assumes that a set of hidden topics structures each corpus document. Applying this method, researchers can uncover latent topics and find relationships among topics from textual data. Many social scientists have used topic modeling, especially Latent Dirichlet Allocation (LDA), the most popular method for topic modeling [27], [28], [29].

LDA is a generative probabilistic model of corpus. The LDA method, first introduced in [24], estimates the probability of topic distribution in documents and word distribution in topics [25]. LDA requires an input parameter (k) to determine the number of topics generated. For the LDA model, the number of topics (k) should be determined by researchers [24], [29]. This study determines the number of topics (k topics) by looking at the coherence score [26]. The coherence score indicates semantic similarity between words on a topic [30].

LDA is a method that assesses that each document is represented as a probability distribution over latent topics, and distribution over words characterizes each topic [24]. Words with the highest probability in each topic are usually used to determine its topic, which, in this case, are the main values. LDA can easily assign probability to a new document; no heuristics are needed for a new document to be endowed with a different set of topic proportions than those associated with the training corpus documents. Based on the cooccurrence of words in a large corpus, LDA algorithms simultaneously estimate topics and assign topic weights to each document. The outputs from a topic modeling algorithm are lists of weighted words, where each list is a topic and higher weighted words in a list are more indicative of that topic. It represents each document as a distribution over topics, which can be used to detect semantic patterns across all documents [24].

In Fig. 2, α and β are determined empirically. Meanwhile, z and θ are determined by the LDA model. M is the number of documents; N is the number of feature words in document m. α represents the priori parameter of the topic distribution over the entire document set. β represents the priori parameter of the word distribution across all topics.

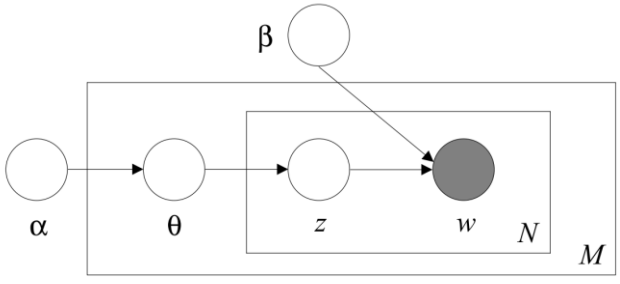


Fig. 2. LDA model diagram [24].

Based on α, the distribution θ of topics in the document is generated. θ is the polynomial distribution of the topic in document M. A topic z is selected from the distribution θ in a document. z is the topic of the n^{th} word in document m; w is the n^{th} word in document m [16]. Algorithm 1 shows the LDA generation process.

Algorithm 1: LDA Generating Model

1: **for all** topics k ∈[1, K] **do**
2: sample mixture components $\varphi_k \sim Dir(\beta)$
3: **end for**
4: **for all** documents m ∈ [1, M] **do**
5: sample mixture proportion $\theta_m \sim Dir(\alpha)$
6: sample document length $N_m \sim Poiss(\in)$
7: **for all** words n ∈ [1, N_m] **do**
8: sample topic index $z_{m,n} \sim Mult(\theta_m)$
9: sample item for word $w_{m,n} \sim Mult(\varphi_{z_{m,n}})$
10: **end for**
11: **end for**

D. Evaluation

LDA algorithms assume there are n topics across all documents to identify topics in a set of documents. The distributions of words determine the distributed k topics across all the documents. This study measures the number of topics (k topics) by the Coherence score [26]. Conversely, the coherence score shows the semantic similarity between words on a topic [30]. The coherence score is formulated as [30]:

$$score(v_i, v_j, \in) = \log \frac{D(v_i, v_j)+\in}{D(v_j)} \quad (4)$$

where,

$D(v_i,v_j)$ counts the documents containing the words v_i and v_j.

$D(v_j)$ counts the number of documents containing the word v_j.

E. Topics Mapping

The primary value is identified by examining the probability of weighing words by the number of topics determined. These main topics' values were evaluated based on OGIWHC to see whether Solo keroncong music meets the OUV criteria as a requirement for UNESCO ICH determination.

IV. Analysis and Result

The dataset used in this study consists of eight documents. After going through the text processing, 2,563 words were obtained, with details of 86 words in the first document, 217 in the second document, 322 in the 3rd document, 480 in the 4th document, 304 in the 5th document, 375 in the 6th document, 268 in the 7th document, and 511 in the 8th document. The weight of each word was then calculated using TF-IDF. Table I shows some examples of words with frequency and TF-IDF values.

TABLE I. Example of Term Frequency and TF-IDF Score

Document No.	Words	Term Frequency	TF-IDF Score
0	*baku*	9	0.5588422545
0	*adaptasi*	6	0.372561503
0	*patah*	4	0.372561503
0	*ciri*	2	0.1862807515
0	moresko	2	0.1862807515
1	*anggota*	7	0.3556741263
1	*asosiasi*	6	0.3048635368
1	*sumbang*	8	0.2709898105
1	*perilaku*	5	0.2540529474
1	kamki	13	0.220179221
2	*rekam*	10	0.261233166
2	*abad*	8	0.2089865328
2	belanda	5	0.1959248745
2	hindia	5	0.1959248745
2	*lokananta*	5	0.1959248745
3	*naskah*	17	0.2863157107
3	*kait*	14	0.2357894088
3	*masuk*	27	0.2273683585
3	*lobi*	8	0.2021052076
3	*presentasi*	8	0.2021052076
4	*karakter*	8	0.361648169
4	*ahli*	10	0.2132273836
4	*dinas*	6	0.1808240845
4	*sederhana*	4	0.1808240845
4	*wayang*	8	0.1705819069
5	*cengkok*	11	0.3362903237
5	*ritmis*	9	0.2751466285
5	*biola*	8	0.2445747809
5	*ngeroncongi*	5	0.152859238
5	*pakem*	5	0.152859238
5	*senggaan*	5	0.152859238
6	*ajar*	24	0.3947967127
6	ramlee	8	0.2790008836
6	Malaysia	15	0.2467479454
6	*zaman*	7	0.2441257732
6	*melayu*	13	0.2138482194
7	*catat*	21	0.2236946025
7	*kait*	9	0.1917382307
7	*seni*	22	0.1589039312
7	*seniman*	10	0.150731548
7	*wbtb*	12	0.1278254871
7	*naskah*	6	0.1278254871
7	*tempe*	6	0.1278254871
7	*instrumentasi*	4	0.1278254871
7	*publik*	4	0.1278254871
7	*wonten*	4	0.1278254871

After the term was weighted, modeling was carried out with the LDA. Several experiments with topics 2 to 10 were carried out to obtain the best Coherence Score. Fig. 3 shows that the number of topics 4 obtains the best Coherence Score of 0.51.

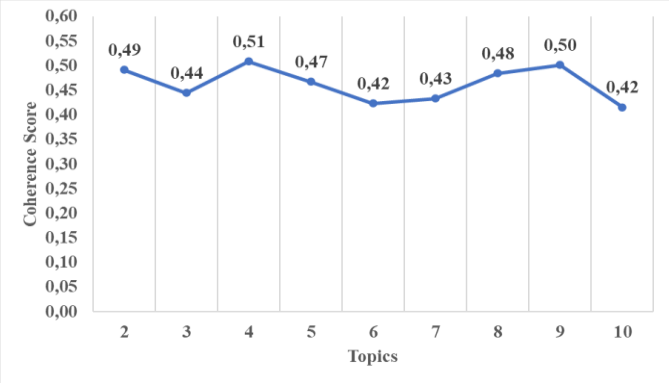


Fig. 3. Comparison of coherence score.

The higher the coherence score, the better the topic modeling interpretation results produced. Using four topics (k=4), the model gave the semantic similarity between words on a topic. Thus, every topic was visualized using pyLDAvis and the word cloud, which was needed in the analysis process.

The terms in the topic modeling are shown in the text, which is primarily frequent in the document. These were depicted by the circle size (as seen in Fig. 5, 7, 9, and 11). Representation of the result using a scatter plot would reveal the distance between topics, the distribution, and the relationship between topic levels. The distance between two or more topics approximates their semantic relationship. Note that close topics 1, 2, 3, and 4 are semantically related, which describes the terms in the topics. As observed in Fig. 5, 7, 9, and 11, the terms are described in the articles concerning the topic's distribution. This reveals that topics 1, 2, 3, and 4 are semantically distributed and have a relationship on topic levels. These reveal four selected topics from the topic model analyzed using the LDA model. The LDA model was one of the input arguments, along with the corpus and dictionary of the emerging terms used for the topic modeling. The slider (λ) in the web-based interactive visualization depicts the relevance metric of the rank terms. It is worth knowing that the terms of the topic are ranked in decreasing order by default following the topic-specific probability (λ = 1). Fig. 5, 7, 9, and 11 reveal the standard terms from the topic model when the slider is at full probability.

Fig. 4 shows the word cloud generated by topic 1, and Fig. 5 shows the model visualization. The comments containing the keywords "kait" and “naskah” appear the most often. In addition, expressions related to music such as “cengkok”, “biola”, “ritmis”, and “genre” are also frequently mentioned.

The model visualization of topic 1 (Fig. 5) reveals the top 30 most relevant terms for topic 1, highlighted with the red bar. These terms account for 49.6% of the overall term frequency, highlighted with the blue bar. Those terms are kait, naskah, masuk, karakter, catat, wakil, usul, ahli, lobi, presentasi, Menteri, WBTB, UNESCO, untung, budaya, kategori, seni, mohon, wayang, dinas, program, pihak, waris, sederhana, eksklusif, akademik, nyata, sapto and dulu.

Fig. 4. Word cloud of topic 1.

Fig. 6 shows the word cloud generated by topic 2, and Fig. 7 shows the model visualization. The comments containing the keywords "baku", “adaptasi”, “karakter”, and “patah” appear the most often. Other expressions related to the community such as “anggota”, “asosiasi”, “ahli”, “KAMKI”, and “perilaku” are also frequently mentioned.

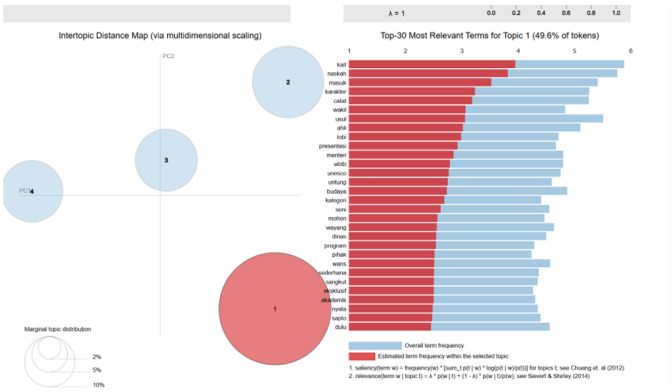


Fig. 5. Model visualization of topic 1.

Fig. 6. Word cloud of topic 2.

The model visualization of topic 2 (Fig. 7) reveals the top 30 most relevant terms for topic 2, highlighted with the red bar. These terms account for 20.2% of the overall term frequency, highlighted with the blue bar. Those terms are baku, adaptasi, patah, usul, KAMKI, cengkok, target, ciri, Moresko, seniman, anggota, gaya, perilaku, asosiasi, sriwedari, sumbang, wadah,

gemar, khas, ndak, ritmis, catat, sumbangsih, wayang, ahli, stimulant, budaya, artis, biola, and enggak.

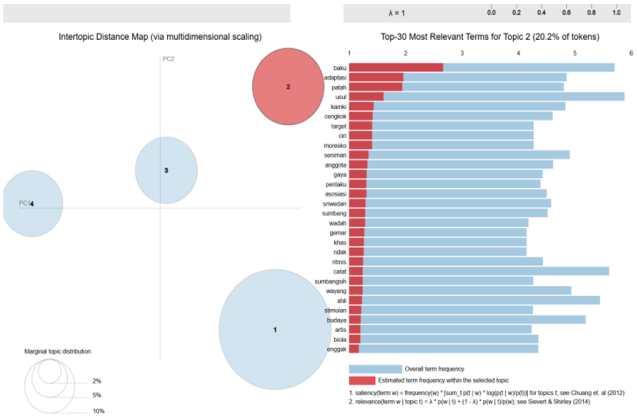


Fig. 7. Model visualization of topic 2.

Fig. 8 shows the word cloud generated by topic 3, and Fig. 9 shows the model visualization. The comments containing the keywords "ajar" and "Ramlee" appear the most often. Other expressions related to history and tradition, such as "Malaysia," "Melayu," "zaman," "kumpul," and "patriotik," are also frequently mentioned.

Fig. 8. Word cloud of topic 3.

The model visualization of topic 3 (Fig. 9) reveals the top 30 most relevant terms for topic 3, highlighted with the red bar. These terms account for 15.2% of the overall frequency, highlighted with the blue bar. Those terms are ajar, Malaysia, cengkok, ritmis, lagu, Ramlee, zaman, and biola. Topic 3 mostly contains terms related to music and the instrument such as cengkok, ritmis, lagu, biola, vokal, klasik, gitar, genre, akor, senggaan, and ngeroncongi; the value of education, tradition, and history such as ajar, Malaysia, zaman, Ramlee, Melayu, pakem, baku, patriotik, metode, kumpul, karakter, mula, and pengaruh.

Fig. 10 shows the word cloud generated by topic 4, and Fig. 11 shows the model visualization. The comments containing the keywords "rekam" and "abad" appear the most often. Other expressions related to the history such as "Hindia", "Belanda", "lokananta", and "historis" are also frequently mentioned.

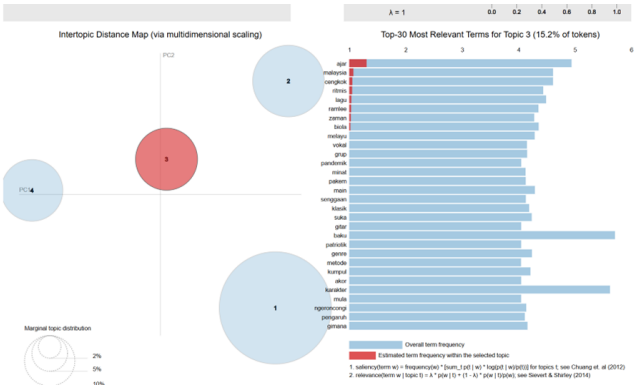


Fig. 9. Model visualization of topic 3.

Fig. 10. Word cloud of topic 4.

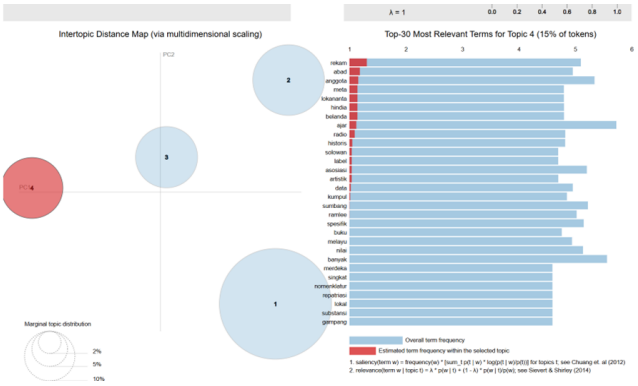


Fig. 11. Model visualization of topic 4.

The model visualization of topic 4 (Fig. 11) reveals the top 30 most relevant terms for topic 4, highlighted with the red bar. These terms account for 15% of the overall frequency, highlighted with the blue bar. Those terms are rekam, abad, anggota, meta, lokananta, Hindia, Belanda, ajar, radio, historis, solowan, label, asosiasi, artistic, data, and kumpul.

LDA modeling results in four topics with the highest coherence score of 0.51. The legitimate source analyzed each word in every topic and mapped it into the OUV values. These main topics' values were evaluated based on OGIWHC to see whether Solo keroncong music meets the OUV criteria as a requirement for UNESCO ICH determination. The result can be seen in Table II.

TABLE II. MAPPING TOPICS TO OUV VALUES

No. OUV	OUV Value	Topic 1	Topic 2	Topic 3	Topic 4
1	Masterpiece	*Cengkok Ritmis Biola Genre Senggaan*	*Karakter Baku Adaptasi Anggota Asosiasi Kamki Ahli Moresko*	*Ramlee Melayu Minat Grup Vokal Suka Lagu*	*Rekam Abad Hindia Belanda Lokananta Artistik Radio Historis*
3	Civilization	*Kait Naskah Ritmis Catat*	*Adaptasi Sriwedari*	*Ajar Zaman Melayu Patriotik Grup*	*Rekam Abad Hindia Belanda Lokananta Radio Historis Buku*
6	Tradition	*Kait Usul*	*Adaptasi Wayang Usul Perilaku Moresko*	*Ajar Ramlee Melayu Patriotik Vokal Lagu Suka*	*Rekam Hindia Belanda Histori Spesifik Nilai*

V. CONCLUSION

This research explored how the data were collected using FGD and in-depth interviews. These primary data are the strength of this study, as Solo keroncong music has never been explored before. Thus, the dataset became this study's key finding and novelty. This research is focusing on creating the dataset and implementing the LDA model. Future research exploring other AI-based cultural studies is still challenging.

The study aimed to find the main values contained in Solo keroncong music as a requirement for eligibility to be proposed as one of UNESCO's intangible cultural heritages. Recognizing Solo keroncong as UNESCO's intangible cultural heritage will have positive implications for preserving Indonesian culture: (1) International recognition will improve Indonesia's image as a country with a rich cultural heritage. (2) Indonesian society, especially the younger generation, will appreciate and be proud of keroncong music as part of their national identity, so the younger generation is more interested in studying and developing keroncong music through innovation and collaboration with other music genres. (3) UNESCO's recognition can increase tourists' interest in learning more about keroncong music, which impacts the creative economy sector. (4) Keroncong music will be better documented, including recordings, academic research, and digital archives.

The corpus was first determined using the Gensim Doc2Bow library. Then, term weighting was conducted by TF-IDF, and topic modeling was used LDA. The modeling phase resulted in four topics, which were then mapped into three OUVs. The results of the research found that there are three OUVs possessed by Solo keroncong music, namely categories (1) representing a masterpiece of human creative genius; (3) bearing a unique or at least exceptional testimony to a cultural tradition or to a civilization which is living or which has disappeared; (6) being directly or tangibly associated with events or living traditions, with ideas, or with beliefs, with artistic and literary works of outstanding universal significance. (The Committee considers that this criterion should preferably be used in conjunction with other criteria). The following is the explanation of the value of the eligibility of Solo keroncong music as UNESCO's intangible cultural heritage based on the results of this research.

A. Keroncong Music is a Masterpiece of Human Creativity

Solo keroncong music is worthy of being a masterpiece of human creative genius because it combines various cultural elements, musical techniques, and unique artistic values. Keroncong music was born due to a harmonious blend of local Indonesian culture with Portuguese fado music. This music became a cultural heritage developed through acculturation with Javanese, Malay, and other cultures. Keroncong music creatively underwent a process of adaptation to local culture. Despite being influenced by outside influences, keroncong music developed with a local taste in melody, instruments, and lyrics.

The main value of keroncong music lies in its artistic value. Keroncong music has unique characteristics in terms of melodic rhythm. The main instruments commonly used in keroncong music are cak, cuk or ukulele, guitar, cello flute, and violin, creating a distinctive sound and unique musical aesthetics. If keroncong music is suspected of being influenced by the presence of Tugu keroncong music, it turns out that there are abnormalities here and there. Keroncong music, which is currently developing in the mainstream, is not like the music in Tugu. There are differences because the music developed in the archipelago is explicitly treated in each region. For example, what developed in Surakarta already has mainstream gamelan music. In the interior of Java, especially Surakarta, music that developed more slowly when using instruments like those above was appreciated and created in such a way according to the character of Javanese. In various regions in the Indonesian archipelago, keroncong music developed with various variants, some using the chordal system like Western music, some using rhythms or in the local scale area. In the Surakarta distribution area, a type of Javanese Langgam uses the Slendro and Pelog scales. Keroncong music mainly uses a system called chordal, namely Western musical concepts. This system uses a diatonic scale in the form of Solmization using Western harmony, namely chords I, IV, V, and II. Since the holding of the keroncong music group competition in Deca Park (City Park, now the Medan Merdeka area of Jakarta) Batavia, Concours (Het nieuws van den dag voor Nederlandsch-Indië, April 11, 1917) or, later called Bintang Radio, keroncong music began to be popular with a chordal system with arrangements using Western music presentation methods.

Keroncong, developed in the Surakarta area (Solo City), has its specialty. The specialty of Solo keroncong can also be seen in combining musical instruments. The use of instruments in Solo keroncong music is unique, and instruments such as the ukulele, guitar, flute, cello, and contrabass are used to create a distinctive sound. This Solo style of keroncong music shows genius in processing simple musical instruments through the

simplicity of the ukulele (cuk and cak) used to create a complex but light rhythm. Solo keroncong also has its uniqueness because it plays songs in Javanese. Javanese style is music typical of the Java region that uses Western instrumentation but presents it regionally, namely the slendro and pelog scales that imitate Javanese gamelan. Songs created by Maestros Gesang, Ismanto Sapari, Anjarany Any, WS Nardi, Dharmanto, and other Solo keroncong music artists have beautiful orchestration and instrumentation typical of the Javanese style.

Keroncong is also known for its rich harmony and melody. The character of Solo Keroncong's music is melancholic and beautiful. The harmony of keroncong music is famous for its softness and melodious melodies, which touch the listener's emotions. Its rhythm is also unique; with its distinctive rhythmic pattern, Solo style keroncong music can create a relaxed atmosphere full of emotion.

Keroncong music lyrics are often used to convey cultural messages and philosophical values. The poetic lyrics of keroncong music contain moral messages, teach the philosophy of life, or are expressions of love and longing. The beauty of its language reflects the genius of its poet and creator. For example, in the keroncong song "*Bahana Pancasila*" by Budiman BJ, the lyrics glorify Pancasila as the philosophy of the Indonesian nation.

Ultimately, keroncong, as a masterpiece of the Indonesian people, has become a national identity. Keroncong is a piece of music considered "truly Indonesian," reflecting the values of nationality, togetherness, and cultural identity that are the nation's pride. The aesthetics of keroncong music are eternal, and its beauty is conveyed by sound, structure, and emotions, making it a work of art that transcends time and is relevant in various eras. With these elements, keroncong is not just music but a masterpiece that illustrates the creative genius of humans in creating, adapting, and inheriting a culture of high value.

B. Keroncong Music as a Witness to Civilization

Keroncong music is considered a witness to civilization because it reflects the historical journey of the Indonesian nation. The long history of the growth of keroncong music began in the 17th century, when the Portuguese in Malacca came to Indonesia or the archipelago, especially in Kampung Tugu, Batavia. They have evolved and assimilated with the culture of the local community, which was initially a form of daily expression after they had been hunting, farming, raising animals, and fishing. They entertained themselves and each other by playing the music they could. At that time, their music was not yet called "keroncong" but was still music from their region as entertainment to relieve fatigue. They were enslaved Portuguese people captured by the Netherlands, a mixture of Portuguese with former colonies of India-Sri Lanka and Malacca, and local people who mostly came from Bengal and Malabar. The Dutch had to change their Portuguese-sounding Catholic names and replace them with new Dutch-style names that embraced Protestant Christianity, such as Mardijkers (a general term for freed former Asian and African slaves) people.

The development of keroncong music from the colonial era to the present shows that this genre can survive. Keroncong music has witnessed society's socio-political and cultural changes in every era. In the colonial era, keroncong witnessed the struggle of the Indonesian people against colonialism. Keroncong songs such as Keroncong Kemayoran were often used as symbols of national awakening and propaganda tools for the struggle. After Indonesia's independence, keroncong became part of national music and is now a symbol of national identity and a unifying tool.

Keroncong shows traces of extinct traditional music. Keroncong can be considered a "living archive" of traditional music that may have disappeared. Many elements of local traditional music are contained in keroncong, such as rhythmic and melodic patterns. Keroncong music often records certain events, values, or lifestyles in its lyrics, thus becoming valuable documentation of past ways of life. Keroncong also shows the tradition of a typical Indonesian ensemble based on harmony and cooperation. This shows the value of the collectivity of Indonesian society in the arts. Nowadays, keroncong has become an art form that connects the old and young generations. Keroncong proves how culture and tradition can survive amidst the current of modernity.

C. Keroncong Music has Traditional Values

In the lives of Indonesian people, keroncong music is still part of the socio-cultural tradition. Keroncong music is an integral part of the Indonesian cultural identity. The lyrics of the songs often contain values from the daily lives of the Indonesian people. Keroncong functions as a medium to preserve and convey local culture and traditions. The themes primarily found in keroncong songs are reflections of regional identity, love, food, love of the homeland, all kinds of cultural activities, advice, or local cultural behavior, both seriously and jokingly. For example, the legendary song "Bengawan Solo" was created by Gesang, a keroncong maestro from Solo; the lyrics describe the beauty of the Bengawan Solo River. Keroncong music still survives in Indonesia today because of its adaptability. Keroncong music in various regions is closely related to local culture. In Batavia, keroncong collided with the music culture of Gambang Kromong; in West Java, it blended with jaipongan; in Central Java, it was famous for its gamelan; in East Java, with Ludruk. Keroncong in Kalimantan and Sulawesi also blended with local music, culture, and traditions. While in Maluku, keroncong was influenced by the cheerful and passionate music of the Ambonese people. The cultural value of keroncong music reflects the musicality and style of its people.

Socially, keroncong music is often played at various events such as weddings, traditional ceremonies, and other social gatherings that help strengthen social relations and a sense of togetherness. At the beginning of the growth of keroncong music in Batavia, for the first time in 1880, musicians from Tugu village played their music outside their hometown. They played in the form of a stage parody with musical performances wearing European costumes and playing guitars with Morisco songs from Portugal by the native people. This music was often used for celebrations by Indo-Dutch crossbreed people. Lower-class Indo-European people favored keroncong music as a form of public entertainment. At that

time, popular entertainment was keroncong music because there were no competitors for other types of entertainment music. Keroncong music reached its golden age in the pre- and post-independence era.

Keroncong music is readily accepted because of its universal values. The songs performed in keroncong music are in Indonesian, using a standard diatonic scale. The types of songs from Western music played in keroncong rhythm/music also use Indonesian lyrics that various groups understand. This phenomenon strengthens the Indonesian nation because they have the same idiom, namely, keroncong music, and its language lyrics. Several times, envoys from Indonesia who participated in international music festivals performed keroncong music, which eventually became a characteristic of Indonesian music. This shows that keroncong music is a representation of Indonesia's national identity. This music symbolizes diversity and unity in difference, one of the basic principles of Indonesia's national motto, Bhinneka Tunggal Ika (unity in diversity).

ACKNOWLEDGMENT

The Ministry of Research, Technology, and Higher Education of Indonesia supported this work and funded it through the Domestic Collaborative Research Scheme in 2024.

REFERENCES

[1] UNESCO, “Operational Guidelines for the Implementation of the World Heritage Convention,” 2023. [Online]. Available: https://whc.unesco.org/fr/orientations

[2] B. de-Miguel-Molina, V. Santamarina-Campos, M. de-Miguel-Molina, and R. Boix-Doménech, Eds., Music as Intangible Cultural Heritage. In SpringerBriefs in Economics. Cham: Springer International Publishing, 2021, doi: 10.1007/978-3-030-76882-9.

[3] S. A. Putra, E. Ismariati, M. Hidayat, K. Setiagama, and Y. A. Nugroho, Buku Penetapan Warisan Budaya Tak Benda Tahun 2020, 1st ed., vol. 1. Jakarta: Direktorat Pelindungan Kebudayaan, Direktorat Jenderal Kebudayaan, Kementerian Pendidikan dan Kebudayaan, 2020.

[4] D. Didit, “Interview/Wawancara Pengurus KAMKI di Kota Solo,” Feb. 21, 2024, Solo.

[5] J. S. Nye, “Public Diplomacy and Soft Power,” Ann Am Acad Pol Soc Sci, vol. 616, no. 1, pp. 94–109, Mar. 2008, doi: 10.1177/0002716207311699.

[6] J. S. Nye, “The Future of American Power: Dominance and Decline in Perspective,” Foreign Affairs, vol. 89, no. 6, pp. 2–12, 2010, [Online]. Available: http://www.jstor.org/stable/20788711

[7] D. I. Witarti and Y. C. Reza, “Pesan Perdamaian Pekan Olahraga Nasional (PON) XX Papua,” Jurnal Ilmu Komunikasi, vol. 21, no. 1, pp. 113–131, May 2023, doi: 10.31315/jik.v21i1.7005.

[8] D. I. Witarti and P. F. Ariyani, “Analisis Sentimen Berita Konser K-Pop Dalam Rangka Peringatan 50 Tahun Hubungan Bilateral Indonesia - Korea Selatan,” Jakarta, Feb. 2024.

[9] D. I. Witarti, A. Puspitasari, and A. Fithriana, “Revitalisasi Musik Keroncong dalam Pelaksanaan Diplomasi Budaya Indonesia ke Malaysia,” Jakarta, Nov. 2023.

[10] D. Sugiyanto and N. B. Aji, “Perkembangan Musik Keroncong di Richmod Virginia Amerika Serikat,” Keteg: Jurnal Pengetahuan, Pemikiran dan Kajian Tentang Bunyi, vol. 19, no. 2, pp. 141–154, May 2019, doi: 10.33153/keteg.v19i2.3080.

[11] P. Kerr and G. Wiseman, Diplomacy in a Globalizing World: Theories and Practices, 1st ed. United Kingdom: Oxford University Press, 2017.

[12] Q. Xu, Y. Xu, and C. Ma, “Analysis of contemporary value and influence of intangible cultural heritage based on online review mining,” PLoS One, vol. 19, no. 12, p. e0315805, Dec. 2024, doi: 10.1371/journal.pone.0315805.

[13] Q. Wang, “The Digitisation of Intangible Cultural Heritage Oriented to Inheritance and Dissemination under the Threshold of Neural Network Vision,” Mobile Information Systems, vol. 2022, 2022, doi: 10.1155/2022/6323811.

[14] Y. Zhang and T. Ala, “Classification and Value Assessment of Sports Intangible Cultural Heritage Resources Combined with Digital Technology,” Applied Mathematics and Nonlinear Sciences, vol. 9, no. 1, Jan. 2024, doi: 10.2478/amns-2024-0548.

[15] Z. Zheng, “Tapping the Economic Potential and Optimizing Strategies of Intangible Cultural Heritage under Digital Transformation,” Applied Mathematics and Nonlinear Sciences, vol. 9, no. 1, Jan. 2024, doi: 10.2478/amns-2024-3187.

[16] G. Xu, Y. Meng, Z. Chen, X. Qiu, C. Wang, and H. Yao, “Research on Topic Detection and Tracking for Online News Texts,” IEEE Access, vol. 7, pp. 58407–58418, 2019, doi: 10.1109/ACCESS.2019.2914097.

[17] Y. J. Lee, S. E. Park, and S. Y. Lee, “Machine Learning-Driven Topic Modeling and Network Analysis to Uncover Shared Knowledge Networks for Sustainable Korea–Japan Intangible Cultural Heritage Cooperation,” Sustainability (Switzerland), vol. 16, no. 24, Dec. 2024, doi: 10.3390/su162410855.

[18] B. Mudjiyanto, “Tipe Penelitian Eksploratif Komunikasi,” Jurnal Studi Komunikasi dan Media, vol. 22, no. 1, pp. 65–74, Jun. 2018, doi: 10.31445/jskm.2018.220105.

[19] J. W. Creswell and J. D. Creswell, Research Design: Qualitative, Quantitative, and Mixed Methods Approaches, vol. 5. United States: SAGE Publications, 2017.

[20] R. A. Krueger and M. A. Casey, Focus Groups: A Practical Guide for Applied Research, 5th ed. United States: SAGE Publications, 2014.

[21] H. J. Rubin and I. S. Rubin, Qualitative Interviewing: The Art of Hearing Data, 1st ed. United States: SAGE Publications, 2012.

[22] T. Jo, Text Mining: Concepts, Implementation, and Big Data Challenge. Springer, 2018.

[23] S. Firman, W. Desena, and A. Wibowo, “Penerapan Algoritma Stemming Nazief & Adriani Pada Proses Klasterisasi Berita Berdasarkan Tematik Pada Laman (Web) Direktorat Jenderal HAM Menggunakan Rapidminer,” Syntax: Jurnal Informatika, vol. 11, no. 2, pp. 10–21, Oct. 2022, doi: https://doi.org/10.35706/syji.v11i02.7192.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.

[25] Q. Yang, “LDA-based Topic Mining Research on China’s Government Data Governance Policy,” Social Security and Administration Management, vol. 3, no. 2, pp. 33–42, 2022, doi: 10.23977/socsam.2022.030205.

[26] S. K. Habibabadi and P. D. Haghighi, “Topic Modelling for Identification of Vaccine Reactions in Twitter,” in Proceedings of the Australasian Computer Science Week Multiconference, in ACSW ’19. New York, NY, USA: Association for Computing Machinery, Jan. 2019, pp. 1–10. doi: 10.1145/3290688.3290735.

[27] H. Jelodar et al., “Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey,” Multimed Tools Appl, vol. 78, no. 11, pp. 15169–15211, Jun. 2019, doi: 10.1007/s11042-018-6894-4.

[28] J. L. Hung and K. Zhang, “Examining mobile learning trends 2003-2008: A categorical meta-trend analysis using text mining techniques,” J Comput High Educ, vol. 24, no. 1, pp. 1–17, Apr. 2012, doi: 10.1007/s12528-011-9044-9.

[29] S. Choi and J. Y. Seo, “An Exploratory Study of the Research on Caregiver Depression: Using Bibliometrics and LDA Topic Modeling,” Issues Ment Health Nurs, vol. 41, no. 7, pp. 592–601, Apr. 2020, doi: 10.1080/01612840.2019.1705944.

[30] N. A. Tresnasari, T. B. Adji, and A. E. Permanasari, “Social-Child-Case Document Clustering based on Topic Modeling using Latent Dirichlet Allocation,” IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 14, no. 2, p. 179, Apr. 2020, doi: 10.22146/ijccs.54507.

Enhancing Chronic Kidney Disease Prediction with Deep Separable Convolutional Neural Networks

Janjhyam Venkata Naga Ramesh[1], P N S Lakshmi[2], Dr. Thalakola Syamsundararao[3], Elangovan Muniyandy[4], Linginedi Ushasree[5], Prof. Ts. Dr. Yousef A.Baker El-Ebiary[6], Dr. David Neels Ponkumar Devadhas[7]

Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India[1]
Adjunct Professor, Department of CSE, Graphic Era Hill University, Dehradun, 248002, India[1]
Adjunct Professor, Department of CSE, Graphic Era Deemed To Be University, Dehradun, 248002, Uttarakhand, India[1]
Assistant Professor, Department of CSE, Aditya University, Surampalem, Andhra Pradesh, India[2]
Associate Professor, CSE-Data Science, KKR & KSR Institute of Technology and Sciences, Guntur-522017, AP, India[3]
Department of Biosciences-Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India[4]
Applied Science Research Center, Applied Science Private University, Amman, Jordan[4]
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur Dist., Andhra Pradesh - 522302, India[5]
Faculty of Informatics and Computing, UniSZA University, Malaysia[6]
Professor, Department of Electronics and Communication Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India[7]

***Abstract*—Chronic Kidney Disease (CKD) is a chronic disease that progressively impairs kidney function to the point of wasting filtration, electrolyte imbalance, and blood pressure control. Early and precise prediction becomes necessary for successful disease management. This research demonstrates a new method involving Deep Separable Convolutional Neural Networks (DS-CNNs) in improving CKD prediction. Based on the Chronic Kidney Disease Dataset available at Kaggle, the model employs DS-CNNs combined with optimized techniques of optimization for better predictive accuracy. DS-CNNs utilize depthwise and pointwise convolutions to facilitate effective feature extraction and classification with efficient computation. To enhance model performance, the Learning Rate Warm-Up with Cosine Annealing technique is used to guarantee stable convergence and controlled rate of reduction in the learning rate. This solution remedies the inadequacies of traditional CKD detection solutions that are insensitive to early stages and entail expensive, invasive procedures. At 94.50% accuracy, the new DS-CNN model outcompetes conventional methods, featuring better prediction performance. The results demonstrate the utility of deep learning and optimization in early detection of CKD and introduce a promising tool for enhanced clinical decision-making.**

Keywords*—Chronic kidney disease; deep separable convolutional neural networks; learning rate warm-up with cosine annealing; predictive accuracy; optimization techniques

I. Introduction

CKD presents a significant global health challenge, affecting approximately 850 million people worldwide [1]. The kidneys are two of the most important organs in the human body located on the right and left flanks of the spine and in the abdominal region just below the rib cage They have the function of regulating water and electrolyte balance and maintaining the body's internal pH by filtering out waste products, excess water and toxins from the blood in the form of urine. Also, they maintain the electrolyte concentration, blood pressure, as well as the acid and base balance; synthesize hormones which are involved in calcium regulation and erythropoietin, which is responsible for the production of red blood cells. CKD is defined by persistent and gradual decrease in the kidney's function and the failure a kidney to remove wastes and balance fluids leading to the build-up of potentially toxic substances and swelling [2].

Chronic kidney disease is a global health issue that still poses great diagnostic challenges especially in the Low- and middle-income countries, hence it often presents itself as a huge burden to the affected patients. Here, inadequate resources and capacity of healthcare facilities hamper early detection and treatment, exacerbating the disease's impact CKD requires timely diagnosis and management to delay the worsening condition to stage 5 where patients usually require costly treatments such as dialysis. However, inadequate screening programmes and relevant facilities restrict access to needful treatments in the region. In addition, other sociodemographic factors and poor health funding existing in the society also lead to disparities in management and overall health of individuals with CKD. These difficulties can only require specific actions to enhance the facilities, increase the understanding and accessibility of satisfactory renal care to people [3]. In Stage 1 the kidneys work effectively to perform all their tasks as they should. At stage 2, common investigations show a slight reduction of function where a body is sometimes able to filter small wastes. Kidneys are severely damaged and, at Stage 3, the patient experiences clearCut impairment of waste removal. Stage 4 is characterized by moderate losses but loss of function is not complete in this stage. The last stage is the stage 5 also known as renal failure which means that the kidneys cannot execute their main functions, so dialysis or transplantation of the kidneys is required.

It is interesting to note that, commonly used primary screening for chronic renal disease, serum and urine test often fail to give an early sign of kidney dysfunction. Such tests are unable to reflect small initial differences in the kidney or variations over time and, therefore, give a delayed diagnosis [4]. In addition, extensive information providing methods like kidney biopsies, are not appropriate for routine screening processes because they involve very intensive procedures and potential risks. The mentioned constraints are solved by Machine Learning (ML) approaches [5] and this option seems quite convincing. It is, however, clear that using big and complicated data, ML systems are capable of identifying fine changes in patterns of kidney function, and that earlier work may miss [6]. Since the factors taken to be used in building an ML model include demographic factor, medical history and clinical test results among others, the models can give more precise and hence be better at doing the CKD risk assessment. The mentioned capacity enables the formulation of specific treatment plans for patients depending on certain general profiles. The efficiency of ML algorithms also enables quicker analysis of new patients' information, which is crucial for the initial diagnosis and treatment. These algorithms can quickly adapt to new knowledge and incorporate this new information through changes in the patient's health into the diagnostic process. Additionally, with the help of machine learning algorithms it is possible to predict the probability of the CKD progression in high risk population and estimate the preventative measures required [7]. Such an approach is quite effective in positively impacting the overall prognosis of the patient since issues that could potentially be of serious concern if left unaddressed are likely to have been dealt with in the act. Taken collectively, it could be postulated that ML algorithms are potentially superior to conventional methods of CKD diagnosis, and as such, they are valuable tools in the fight against this prevalent and significant health condition.

As a result, many studies have had to harness other approaches that are at the algorithm level like ensemble learning and cost-sensitive learning rather than data level techniques [7]. Ensemble learning is a revolutionary concept in the arena of ML study and implementation employed to achieve very high accuracy of the classifier through integration of one or more classifiers. Boosting [8] and Bagging [9] are widely used ensemble learning techniques. Therefore, in this study, we develop an AdaBoost classifier that gives higher weight to examples in the minorities, thus improving the prediction of the samples of the minority class and the global classification accuracy [11]. The currently used techniques for CKD diagnosis such as blood tests, urine tests, and even invasive techniques like biopsy have various drawbacks. These methods are usually not very accurate when it comes to determining the early sign of kidney failure and the changes in the kidney function with time, therefore the diagnosis and treatment for the kidney diseases are often delayed. In addition, KUB radiography, ultrasonography and CT scans are expensive and sometimes take a lot of time which limits their use. In spite of the fact that the validity of CKD detection has increased with the help of ML methods which analyze big data and find rather complicated patterns, however, there are the disadvantages of such methods [10]. The issue with the current paradigm of using conventional ML algorithms, however, is an inability to adequately process the data, particularly if CKD cases are vastly in the minority compared to controls represented by non-CKD patients.

Some of the problems are solved by the generic ensemble learning techniques such as AdaBoost that involves using multiple classifiers brought together to enhance the accuracy of the classification process as well as work on imbalanced data set. Nevertheless, they can still encounter difficulties in terms of low computational efficiency for the data with large dimensionality and the complexity of feature interactions. The above-mentioned limitations are inherited from the traditional machine learning frameworks, and can be mitigated with the help of the proposed state-of-the-art deep learning framework, namely Deep Separable Convolutional Neural Networks (DS-CNNs). DS-CNNs gives better accuracy in detecting the changes in the CKD as convolutions lead to the structural difference that improves the working ability of the model and early stage of kidney damage. In addition, this conceptual framework proposes methods for dealing with imbalanced data which enables both the minor and major classes to be dealt with equally well improving the general performance of classification. The proposed model works around these challenges by incorporating dynamic learning rates and fine tuning through the helps of other optimized methods and algorithms which are far much better than other traditional methods of Machine Learning named as ML; and thus provoking the CKD prediction dashboard based on the CKD dataset of the UCI Machine Learning Repository with better efficiency and accuracy. The major key contribution can be divided among the following:

- Enhancing the framework with the proposed DS-CNNs model and the LRW-CA optimization for better performance and reliability of the CKD identification.
- Applying DS-CNNs in order to eliminate the convolutional processes into separate depths and points, which helps improve the quantity of computations required as well as the quality of the designed model.
- Standard practice of implementing the Learning Rate Warm-Up together with Cosine Annealing helps improve the early training phases' stability, which in turn results in stable and accurate model performances.
- Avoiding such problems as low sensitivity to changes in CKD at early stages and high invasiveness of some diagnostic methods by implementing modern deep learning methods and optimizing clinic algorithms for their application.

The research paper is organized as follows: The importance and difficulties in predicting CKD are described in Introduction. The following section, 'Related Work,' presents a critical analysis of previous methods of CKD detection and their shortcomings. Methodology section presents the implementation plan of the proposed work, the proposed DS-CNN architecture, and the optimization strategies. Experimental Results gives the results and discussion on the effectiveness of the proposed method compared to those in the literature. In this paper, conclusion will make an effort to briefly restate the key findings, reflect on the implications of the study, and point out the possibilities of future research in relation to the existing literature.

II. Related Work

To improve the work done by Chotimah et al [11], the authors created a second-generation advanced deep learning approach that dealt with feature selection to identify the most relevant CKD diagnostic markers from patients' records. SBFS was used in their research; it is a technique, which effectively removes successively one feature after the other that has less influence in the CKD prediction model. The two were able to utilize SBFS and settle for 18 features out of which the relevant 15 were deemed pertinent. These selected features were then used as inputs to an Artificial Neural Network (ANN) Classification system. The revised model calculated with only 15 most significant attributes was 88 %, which was much improved than the one which was 80% acquired from 18 attributes. This much enhanced results showed that feature selection had a way of improving the model hence enhancing better forecasts of CKD. But, since the features are selected manually, then the framework may not analyze some interactions and dependencies between them or may not see the interdependencies among a few features and thus the framework is not able to capture complex relations of the features present in the data set. The above framework entitled Enhancing Chronic Kidney Disease Prediction with Deep Separable Convolutional Neural Networks (DS-CNNs) mitigates this drawback by using DS-CNNs to learn and extract multi-level features directly from raw data to enhance the predictive models' accuracy.

Alsuhibany [12] designed a sophisticated IoT based ensemble diagnostic system for CKD named as EDL-CDSS. This system used ADASYN to improve outlier detection, and integrated multiple deep learning strategies: CNN-GRU, DBN, and KELM. Therefore, the combination of these technologies operationalized the concept of ensemble, and it obtained a remarkable accuracy level of 96%. 91% by showing that by using several deep learning models with sophisticated approaches and bagging technique a precision CKD diagnosis and good outlier management could be achieved. However, use of multiple models can be complicated making the process of integrating them time consuming it could take a lot of time before the desired results are accomplished making it a bit cumbersome for real world use. To overcome these drawbacks, the proposed framework "Enhancing Chronic Kidney Disease Prediction with Deep Separable Convolutional Neural Networks (DS-CNNs)" employs DS-CNNs enable simplified architecture to analyze the complexity of the patients' data, which slows down training times and increases computational demands, then again, underestimating the complexity of the patients' data leads to low predictive accuracy.

Akter et al y.in their study [13] selected seven advanced deep learning methods for predicting the probability of developing CKD, namely; ANN, LSTM, GRU, bidirectional LSTM, bidirectional GRU, MLP, and simple RNN. On these models, they had certain evaluation criteria such as loss, validation loss, recall, accuracy and precision. To the researcher's surprise, three methods, namely ANN, simple RNN, and MLP exhibited superior results with accuracies of 99%, 96%, and 97% respectively. Based on such an extensive assessment of the performance of these algorithms, their high results in the classification of CKD patients were established, as well as the potential of deep learning in providing more accurate predictions. However, the presented individual models approach might have a problem when it comes to the generalization across the datasets and the capturing of the interactions among features. The above improvement is lack of due to the proposed framework, Enhancing Chronic Kidney Disease Prediction with Deep Separable Convolutional Neural Networks (DS-CNNs), due to the use of DS-CNNs to capture hierarchical features and interactions of the data to generalize and apply it to different datasets.

In a study [14] conducted by Iliyas and colleagues, 400 patient records gathered at Bade General Hospital and 11 features were applied to predict CKD through a Deep Neural Network (DNN). They managed this by imputing with the mean of respective attributes in the preprocessing step. They utilised a DNN model in predicting CKD with an accuracy rate of 98%, and understood that creatinine and bicarbonate were the critical factors in the prediction. This paper presented how DNNs can be used to solve CKD prediction tasks and the need for data pre-processing and feature extraction to optimize high prediction accuracy. But, missing data have been addressed using the mean imputation, and this might not handle variability of data and it can affect the model significantly. The proposed framework, Enhancing Chronic Kidney Disease Prediction with Deep Separable Convolutional Neural Networks (DS-CNNs), avoids the aforesaid problem by using sophisticated imputation techniques and exploiting DS-CNNs in learning and fusing highly nonlinear and intricate features of data that, in turn, promotes accurate CKD prediction with improved model integrity.

Ma et al. [15] gave a detail insight of the social focus of heterogenous modified artificial neural network (HMANN) for CKD detection, segmentation and diagnosis within the Internet of Medical Things (IoMT) framework. The HMANN model was further developed focusing on early detection and accurate segmentation of the kidney images and the avoidance of noise. Regarding the accuracy rate of the different methods they employed, their study ushers an average of 92. 3 % for ANN-SVM and 97 %. 5% for HMANN. This research highlighted the major application of HMANN and its applicability for improving the diagnostic accuracy and image analysis of CKD and provided an example of future possibilities of the system in practical medical imaging and diagnosis. At the same time, developed HMANN architecture can be considered as complex and indispensable may lead to increasing the load on the computer and time required for its work. The mentioned framework is a solution to these problems because it incorporates a leaner model architecture yet proves to be efficient in retaining those features essential for fast and accurate CKD diagnosis and prediction – the DS-CNNs.

In particular, the identification of CKD [16] was developed using deep learning approach by combining Bidirectional LSTM networks and one-dimensional Correlational Neural Network (1-D CorrNN) suggested Bhaskar and colleagues. The combined model, 1-D CorrNN-LSTM was tested using CKD-sensing module with the accuracy of 98%. 08%. This also implies that the proposed model performed better than other approaches in the sense that this research was able to validate the model on time series, a factor that enhances the generalization of the model for the prediction of CKD. The study showed that the model is

capable of enhancing the diagnostics accuracy and dealing with a variety of data features related to the CKD. Yet, since the model is quite complex and works with expanding sequences, the training of the model may be more time-consuming and computationally intensive. The DS-CNNs model used in the proposed framework, Enhancing Chronic Kidney Disease Prediction with Deep Separable Convolutional Neural Networks (DS-CNNs) does not have such issues since it has a more efficient architecture that does not require a lot of computation and training times in order to capture more complex patterns hence improving the performance and practicality in real world applications.

ANNs with SBFS were applied by Chotimah and colleagues, and the estimate of overall accuracy was 88%. Alsuhibany and team designed an interpersonal deep learning ensemble system named as EDL CDSS achieving an accuracy of 96 by using various deep learning techniques. 91% although with a higher computation complexity. Five state of art deep learning models were implemented and tested; ANN, LSTM, all achieved high accuracy but generalization and complex feature interaction were the major limitations that Akter encountered. For imputation, 98% accuracy was obtained by Iliyas by using imputation techniques and Deep Neural Networks (DNNs) and however mean imputation might impact robustness. Ma proposed a heterogeneous modified artificial neural network (HMANN) for the detection of CKD, which had high accuracy, but required more computational resources for training. Bhaskar incorporated a novel created 1-D Correlational Neural Networks (1-D CorrNN) with Bidirectional LSTM networks and got an accuracy of 98. 08% accuracy but the problem they face is that the complexity and the time they have to spend on training increases. These constraints are solved with the help of hierarchical learning and the use of depthwise separable convolution in the framework of DS-CNN, which opens the question of automation of the feature extraction process and reduces computational complexity. It also applies Learning Rate Warm-Up with Cosine Annealing as the strategy of training, which help for improvement of generalization. This approach eliminates past challenges, decreases the size of ensemble models, enhances resource generalization, and solves problems connected with data imputation and computational effectiveness, offering a reliable and efficient method for CKD prognosis.

III. Problem Statement

Existing research [17] on the approach of CKD detection throws light on different machine learning algorithms such as SVM, Decision tree, and Random forest. This it compares them and notes that they while they are are reasonably accurate, they suffer from drawbacks such as high computational complexity and inability to handle higher-order feature interactions respectively. Hand crafted features and manual tuning also limits model performance and generalization as well. The use of handcrafted features and manual tuning can also constrain model performance and generalization. To these ends, the proposed framework, Enhancing Chronic Kidney Disease Prediction with Deep Separable Convolutional Neural Networks (DS-CNNs), does away with these limitations by employing DS-CNNs that are capable of automatically extracting hierarchical features from raw data to expand the model's capability to identify the diverse information patterns. When Cosine Annealing is used in combination with Learning Rate Warm-Up early phases of training are stabilized and convergence is enhanced. It also brings down the computational expenses, increases the prediction conveyance and is a neater algorithm in comparison to existing approaches for CKD prediction.

IV. Proposed Framework for Chronic Kidney Disease Prediction Using DS-CNNs and Learning Rate Warm-Up with Cosine Annealing Optimization

About the suggested framework for improving the prediction of CKD, it is possible to note that the main aim of the proposed algorithm is to increase the accuracy of diagnostic results due to thorough attempts. The process begins with the Input Data which consists of the dataset of chronic kidney disease patients, which is a vital tool in training and evaluating patients. The Data Preprocessing phase involves several rigorously performed activities such as cleaning, normalization, data formatting directly for model input so that our data does not contain any such unusual variation or skewness that would impact the model performance. Lastly, under the Feature Extraction domain, the use of Deep Separable Convolutional Neural Networks (DS-CNNs) is used by the framework to extract feature relevant to the dataset. This process involves two key operations: This in turn consists of two types: Depth wise Convolution that works by performing convolution on each of the input channels separately and Point wise Convolution, which integrates the depth wise features through 1*1 conjugate convolution in an attempt to generate a small, yet comprehensive feature representation.

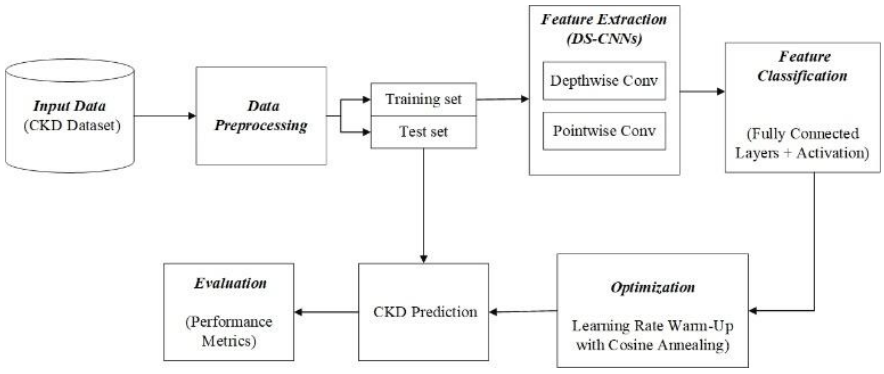


Fig. 1. Proposed framework for enhancing CKD prediction with DS-CNNs.

The extracted features are then fed into Fully Connected Layers in which the data goes through different activation functions so as to perform classification and introduce non-linearities to improve the performance of the model. Learning Rate Warm-Up with Cosine Annealing is adopted in the Optimization stage of the framework to achieve better training stability and convergence because it starts with increasing learning rate gradually and then decreases it following the cosine pattern. The consequence is that, with good input information, the CKD can be well predicted, and the result would give a probability of the disease, in general. Finally, the Evaluation phase assesses the model's performance using metrics such as accuracy, sensitivity, and other relevant indicators. This stage also includes a comparison with existing methods to highlight improvements in predictive accuracy and reliability achieved by integrating advanced DS-CNNs and optimization techniques. This comprehensive framework aims to address the limitations of traditional CKD detection methods by leveraging sophisticated deep learning and optimization strategies. Fig. 1 illustrates the proposed method.

A. Dataset Description

In this framework, Chronic Kidney Disease (CKD) dataset reflects thorough patient health records database intended for creating and evaluating forecast models for CKD. The features in this dataset are quite diverse, and categorize and numerical health data that are significant for CKD diagnosis are included. The nominations are Age, in years; Blood Pressure with two categories: systolic and diastolic. The Specific Gravity attribute gives the information of concentration of urine and Albumin shows the kidney function amount in the urine. Sugar, thus equalizes the presence of sugar in urinals; Red Blood Cells and Pus Cell attributes signify the presence of red blood cells and pus cells in the urinary sample respectively. Other parameters like Pus Cell Clumps and Bacteria represent the clump of pus cells as well as bacteria in the urinary system that helpful to diagnosing the kidney diseases. BG Random indicates the random blood glucose test results, BU presents the blood urea level, and SCr shows the creatinine level in serum as significant markers of kidney compartment. The values Sodium and Potassium portray serum sodium and potassium concentration, on the other hand, Hemoglobin and Hematocrit portray the level of hemoglobin and proportion of RBCs in blood respectively. The two attributes under the same groups are White Blood Cell Count giving the number of white blood cells in the body and Red Blood Cell Count is the number of red blood cells in the body. The categorical attributes Hypertension, Diabetes Mellitus, and disease meaning that one has/have such diseases (Yes/No) have big chances to develop CKD. The Classification attribute is the dependent variable, which categorizes patients into the groups of CKD and Not CKD. These various health indicators enhance the possibility of the model learning to diagnose CKD and increase the model's accuracy from the large database [18].

B. Data Preprocessing

Cleansing stage is an essential task because it is the foundation that enables the data to be of high quality for use in the developing of a predictive model. The following steps outline the typical preprocessing procedures for the CKD dataset: The following steps outline the typical preprocessing procedures for the CKD dataset:

1) Data cleaning: Preprocessing of data, the first stage is always the processing of missing values that may be contained in the dataset. Data missing can be handled using the various techniques like data imputations which entails using the mean, median or even mode of the feature missing data set. For more complex cases there is also predictive imputation that can also be employed. Also, any rows that have too many missing values may also be omitted in order to ensure that the remaining data set is clean. To remove the redundant records and any bias in the model a data deduplication process is also performed.

2) Data transformation: After cleaning the data it is transformed for it to be in the right form for the model input. This includes normalization or standardization of numerical features such as weight, height, body mass index, blood pressure and others. Normalization scales the features onto the range [0,1] while standardization scales the features to have a mean of 0 and standard deviation of 1. This makes each of the features have comparable scales among the numerical attributes, which aids in enhancing the model's viability and the rate of convergence. Binary values such as `Hypertension`, `Diabetes Mellitus`, and `Coronary Artery Disease` present categorical characteristics and they have to be quantized through features like one hot encoder or a label encoder. This transformation is needed when non-numerical categorical data has to be incorporated in the model.

3) Splitting data: To make the right forecast and check the model's accuracy, the dataset is divided into the training and testing sets. It has been tested and recommended that mostly 80 percent of the data is used in training the model, while 20 percent is used in testing the performance of the accomplishments. It guarantees the evaluation of the model's performance as the training process occurs on one set of data and the validation is done on another set.

Feature Scaling: Normalize is very important to all features in a dataset so that it does not have a very high impact influence the model. It brings the scale of the independent variables or features of data into standard, which is useful for the algorithms depending on the type of input features, for example, neural networks.

- ***Min-Max Scaling***: his technique scales the values of features to a particular range usually between 0 and 1. It is achieved using the following formula:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

- X is the initial value of the feature.
- X_{min} is the minimum value of the feature present in the designed dataset.
- X_{max} is the maximum value over the feature across the dataset.

- X_{scaled} is the scaled value of the feature of the given data.

This transformation helps to ensure that all feature values are in the range [0, 1] which is also beneficial for machine learning algorithms because they do not get influenced by the sequence of features' scales. Thus, all these preprocessing procedures allow the dataset preparing in the way that will be most suitable for the Deep Separable Convolutional Neural Networks (DS-CNNs) and other parts of the predictive environs and, as a consequence, provide better results in terms of the model accuracy and certainty.

C. *Classifying Features Using Deep Separable Convolutional Neural Networks (DS-CNN) in Detecting CKD Disease*

The Fig. 2 illustrates, a Deep Separable Convolutional Neural Network (DS-CNN), which consists of depth-wise convolution and point-wise convolution layers. The input layer is the first layer in the process through which the data like images or health metrics are taken. In the depth-wise convolution layer, all the channels of the input data are convolved with separate filters, as it produces a set of feature maps and only the spatial information within the channel is retained. Subsequently, the depth-wise convolution applies 3 by 3 filters on each position in first derived feature map then the point to point convolution applies 1 by 1 filter [19]. This layer integrates and remaps the information across all the channels generating new feature maps of the extracted features. Subsequently, there is the point-wise convolution layer that convolves the 1×1 filters on every position of the feature maps which depth-wise layer generates. This layer aggregates data from all the channels and makes new feature maps that are comprised of all the features obtained from each layer. The order of general interconnection significantly entails that the input data undergoes depth-wise convolution with the purpose of channel-wise feature extraction, in addition to point-wise convolution that fuses and boosts the features and leads to a modified and integrated output. This architecture is cherished for its simplicity especially in PASS, M&N, and LNNs owing to the fact that it minimizes computational demand as it efficiently enables the transformation of features from the input layer [20].

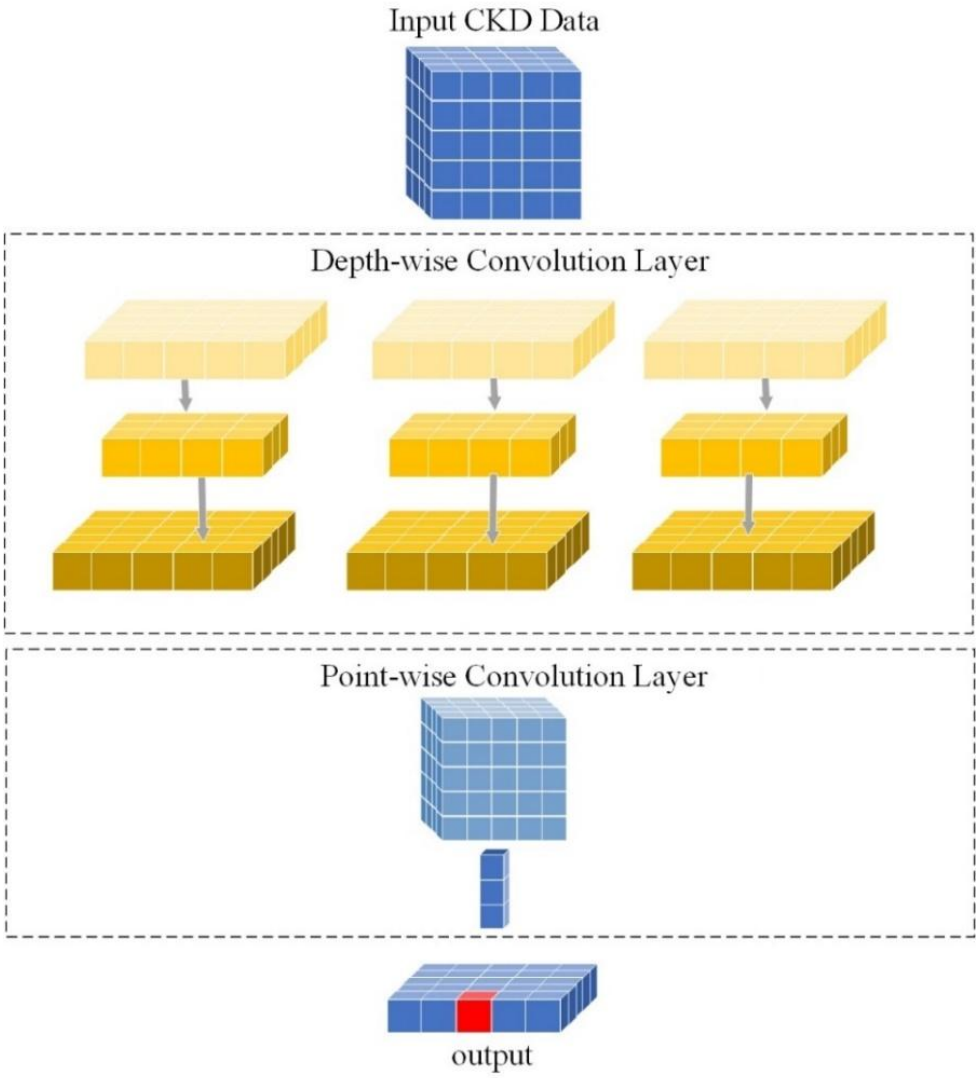


Fig. 2. Architecture of a Deep Separable Convolutional Neural Network (DS-CNN), showcasing its depth-wise and point-wise convolution layers.

DS-CNN attributes the chronological sequential operations that make up the whole process of the workflow for the detection of CKD. The process begins with the input layer that takes in the CKD dataset after the preprocessing in addition to several health parameters and disease marker [21. Then, the obtained data goes through the depth wise convolution and in these operations each channel from inputs (for example, several health indices) is passed through its own filter. For an input tensor with dimensions (H times W times C height H), width W, and C channels, depth wise convolution applies a filter, depth wise convolution applies a filter K_c to each channel c individually.

$$(X_{depthwise} * K_{depthwise})_c = \sum_{i=1}^{K}\sum_{j=1}^{K} X_c(i,j).K_c(i,j) \quad (2)$$

where $X_c(i,j)$ is the input feature map for channel c, and $K_c(i,j)$ is the depthwise convolution kernel corresponding to that channel. Subsequently, a pointwise convolution is performed to mix the outcomes of the depthwise convolution by channels. This becomes achieved by utilizing a convolution kernel of 1×1 that operates on all channel dimensions and combines them into a new feature map [22].

$$(X_{depthwise} * K_{depthwise})_c = \sum_{c=1}^{C} X_c . K_{n,c} \quad (3)$$

where X_c is the depthwise-convolved feature map, $K_{n,c}$ is the 1x1 convolution kernel, and n represents the output channel. This is followed by an activation function which helps to incorporate non linearity into the model such as ReLU (Rectified Linear Unit):

$$ReLU(x) = max(0,x) \quad (4)$$

This non-linearity of derivation is useful for the network to learn the patterns inherent in the data. Successively, pooling layers like the max pooling layer is applied to down sample the feature maps and focus more on the important features and commonly, it also helps to minimize the computational costs during training and also prevent overfitting. Finally, the feature maps are flattened and passed through the fully connected layers for classification of images. These layers learn how to obtain the outcome of CKD by perceiving the features extracted from the data. The output layer uses Softmax function if it is for multi label classification or Sigmoid function if it is for binary classification to give out the class probabilities. For binary classification, the sigmoid function for binary classification, the sigmoid function is:

$$Sigmoid(z) = \frac{1}{1+e^{-z}} \quad (5)$$

where, z is the output of the fully connected layer. Last but not the least, it is necessary to train such a network through the use of a loss function, often the binary cross-entropy together with an optimizer such as Adam or SGD to help reduce the error margin between the network's predictions, and the actual values. This general workflow that reconstructs and combines depthwise and pointwise convolutions, activation, pooling, and fully connected layers makes it possible for the proposed DS-CNN to learn and output the probabilities of the existence of CKD based on the intricate features captured in the dataset.

D. Workflow of Learning Rate Warm-Up with Cosine Annealing Optimization after DS-CNN Training

Subsequent to treating the CKD dataset through the Deep Separable Convolutional Neural Network (DS-CNN) and generating the feature maps, an optimal approach is used to further improve the training outcome of the model – the Learning Rate Warm-Up with Cosine Annealing Optimization. The first is learning rate warm-up where the learning rate is set to a small value especially when using gradient descent and then is increased to the given maximum value after some epoch. It helps to make small adjustments in initial phases and slowly increases to prevent the phase from getting destabilized. The learning rate during the warm-up phase can be expressed as

$$\eta(t) = \eta initial + \frac{(\eta_{max} - \eta_{initial})}{\eta_{warm}} . t \quad (6)$$

where $\eta initial$ is the initial learning rate, η_{max} is the maximum learning rate, η_{warm} is the number of warm-up epochs, and t represents the current epoch.

After the phase called warm-up phase is done the learning rate acts under the so-called cosine annealing and depends upon the cosine decay function. This phase gradually brings down the learning rate by minimizing it from the maximum level to the minimum level in the remaining epoch. The learning rate $\eta(t)$ during cosine annealing is given by:

$$\eta(t) = \eta_{min} + \frac{(\eta_{max} - \eta_{min})}{2}(1+\cos(\frac{t - T_{warm}}{T - T_{warm}}\pi)) \quad (7)$$

where η_{min} is the minimum learning rate, η_{max} is the maximum learning rate, T is the total number of epochs, T_{warm} is the number of warm-up epochs, and t is the current epoch number. This optimization strategy makes sure that the function is firstly trained with the learning rate that starts in a low stable level and then increases and after that, the decreasing helps to fine tune the parameters of the DS-CNN model appropriately. The warm-up phase is useful in preventing fluctuations that may lead to the derailing of training while the cosine annealing phase polishes the learning process which increases convergence hence optimizes the results of the model.

The steps of the proposed framework for CKD prediction are as follows: They also present a step-by-step process in the case of a Chronic Kidney Disease prediction model that has been designed for correct effective functioning in a hospital environment. The first process to be carried out is the Data Acquisition in which the CKD dataset is obtained from Kaggle. This dataset includes detailed information of patient's health as components of CKD including hypertension, serum creatinine, and diabetes among others. This feeds into the training and testing of the predictive model to be used in the predictive system. Data Preprocessing: This step to prepare the data for the model is a core-competitive process of Advanced data analysis. First, Data Cleaning deals with the missing values either by imputation or by removing the records having point or feature missing values, also removes the duplicate entries. Data Transformation's crucial step that involves scaling or normalizing numerical attributes into the same scale and converting categorical attributes into numerical scales for the model. Feature Engineering might also involve feature construction where new features are made or feature selection for the right features to be taken into consideration. It is then divided into the training and testing sets in order to assess the model's accuracy and Feature Scaling is used to make all feature

values fall within the same scale and that boosts up the training of the model

Stepwise Algorithm for Detecting CKD using DS-CNN & Learning Rate Optimization	
Step 1: Data Acquisition	//Download the Chronic Kidney Disease (CKD) dataset from Kaggle
Step 2: Data Preprocessing	
Step 3: Model Classification	//Define the DS-CNN architecture, including depthwise convolution, pointwise convolution, activation functions, pooling layers, fully connected layers, and the output layer using $Sigmoid(z) = \frac{1}{1+e^{-z}}$
Step 4: Optimizing the Model	**//Learning Rate Warm-Up**: Gradually increase the learning rate from a low initial value to a maximum value over the initial epochs using $\eta(t) = \eta initial + \frac{(\eta_{max} - \eta_{initial})}{\eta_{warm}} . t$ - **Cosine Annealing Optimization**: Smoothly decrease the learning rate after warm-up using $\eta(t) = \eta_{min} + \frac{(\eta_{max} - \eta_{min})}{2}(1+\cos(\frac{t-T_{warm}}{T-T_{warm}}\pi))$
Step 5: Model Evaluation	//Evaluate the model's performance using metrics such as accuracy, sensitivity, and specificity
Step 6: Deployment	//Deploy the trained model for real-world CKD prediction applications, integrating it into healthcare systems or decision-support tools.

The third step in the process of Deep Separable Convolutional Neural Network (DS-CNN) is the Model Building follow the following steps: In the Depth-wise Convolution, it is separately performed filters over the input channels to capture spatial features individually; in the Pointwise Convolution, the 1x1 filters ensure the learning of these individual features from different channels and their relationships. This combination of depth-wise and point-wise convolution layers is very helpful to extract most of the features and to reduce the dimensions. Other elements are Activation Functions such as ReLU which introduces the non-linearity into the network, Pooling Layers which decrease feature maps dimensions and paid attention to significant features, and Fully Connected Layers which perform the final classification with the help of extracted features. In the fourth step namely 'Training the Model', there is the Learning Rate Warm-Up then Cosine Annealing Optimization. As part of the learning rate warm-up, the learning rate is ramped up from a low starting value such as 0.001 to a higher value such as 0.1 over some epochs to improve stability of the training. After this, the learning rate of the model is made gradually variable with the help of Cosine Annealing Optimization so that it fine tunes the model to the greatest extent. The learning rate does not decreases with a constant rate rather with a cosine function. Step 5 of the model development can be described as Model Evaluation whereby the effectiveness of the developed model is checked using significant measures for instance accuracy, sensitivity and specificity. This step involves the assessment of the performance of the model in predicting CKD against other methods, contacting gaps and checking if the model can perform well on a novel dataset. Deployment formally sets the model into operational use cases; these uses could be in the health care system or decision support systems. This step makes an assurance of effective usage of the model in the real world by diagnosing and managing cases of the CKD, whereby the model offers crucial predictions that come in handy when making other decisions. Each of these steps is essential for developing a robust and reliable predictive model for CKD, ensuring that the final system is both accurate and practical for real-world use.

V. Result and Discussion

Analyzing the apply the proposed framework for CKD prediction using Deep Separable Convolutional Neural Networks (DS-CNN) with Learning Rate Warm-Up and Cosine Annealing optimization provided the following learning: Measured in Python, the framework used a large assortment of data to improve the diagnosis of CKD. The results showed the ability of the DS-CNN to extract important features from the health metrics enhancing the classification performance, sensitivity, and specificity of the models. The incorporation of the Learning Rate Warm-Up with Cosine Annealing optimization proved effective in stabilizing the model's training and fastening convergence than the baseline models. The usage of both age and gender was noteworthy when identifying a relationship between age-specific tendencies and model performance, especially in relation to CKD prediction. Algorithms of this framework were more realistic and accurate due to some enhanced optimization strategies that formed the basis for the development of better CKD diagnosis instruments. These finding stress the possibilities of further advancements in the chronic diseases risk prediction using intricate machine learning approaches.

A. Demographic Evaluation of CKD Dataset

The following sub-sections discuss the details of the CKD patients' population with regard to various factors including their age, gender and their key vital statistics. Age Distribution data depicted that maximum patients of Chronic Kidney Disease belong to 50 to 55 year age group. The supply chain risk analysis also shows how age affects CKD prevalence because age plays an extra role in the increase in disease frequency in the older age groups. Gender Distribution, specifically, highlights proportionally more patients that are male and thus, gender aspects can influence CKD incidence and/or progression. Last but not the least the 'Health Metrics' table includes details regarding different aspects related to a clinical status of the patient including blood pressure, serum specific gravity, albumin levels and other such factors which are vital while assessing the severity and resultant consequences of CKD. These metrics prove useful in a way that enables patterns and correlations to be seen, which, in turn, aid in making more correct diagnoses and allowing for specific treatment interventions. Hence, the breakdown of these factors offers a deeper understanding of Non CKD, enhance the management and interventions that can suit patient's demographic and clinical status.

1) Distribution of patients' age: The Distribution of Ages classifies patients into three different age groups to give an easy perspective of the CKD prevalence across the patient's ages. The ratio of the number of patients according to the age range

of 45 to 50 years is the smallest of the indicated numbers and equals 100. The patients aged between 50-55 years are many, with 150 patients, hence signifying the senior group as having the most CKD cases. Aging between 55 and 60 years is another category of the patients which is 120, less than the 50 to 55-year bracket, but still a worthy number. This distribution demonstrates a trend common in many population concerning the increment of the probability of chronic kidney diseases with advanced age. This distribution reflects a typical trend where the risk of chronic kidney disease increases with age. Older age is associated with a higher incidence of CKD due to the cumulative effect of risk factors and the natural aging process affecting kidney function.

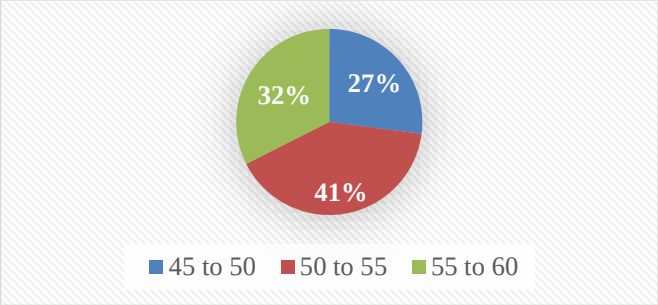


Fig. 3. Distribution of ages.

Fig. 3 corresponds to an age group and the height of a bar shows the number of patients. Exploring this representation makes it easier to learn which age level is most impacted by CKD and therefore assists in designing angles to prevent CKD or to support those aged 2 to 70 years in case they are diagnosed with the disease.

2) Distribution of gender diversity: Distribution of Gender is presented in the table below and represents the proportion of male and female patients affected by CKD in the particular dataset. The overall involvement in CKD of the males is higher within the dataset which has 190 male patients and 150 female patients. These differences can be as a result of a variation in their lifestyles, their susceptibility to the diseases that may lead to hypertension and diabetes and any inherited conditions that may affect the kidneys differently from one gender to the other. Knowledge on gender distribution is important in attempting to formulate focused intercessions and treatment methods when handling CKD, since this may predict how it presents and evolves.

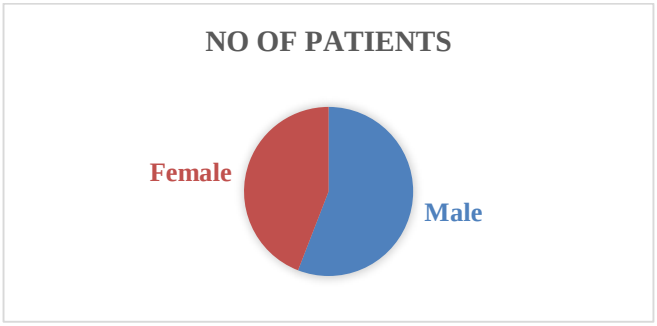


Fig. 4. Distribution of gender.

Fig. 4 usually shows the ratio of male to female clients and therefore offers an easy method of comparing the gender balance in the practice. This leads to identification of gender differences in the development of CKD to curb any existing trends and contribute to management of CKD based on gender.

3) Distribution of affected chronic kidney disease affected patients: The distribution of affected chronic kidney disease patients table offers simple and clear information presentation of various health indicators of affected CKD patients. Materials used in this table contain essential values that are important in reflecting kidney performance and deciphering CKD's influence. It depicts that the existence of the hypertension among patients differs in values like 80/120mmHg and 85/130 mm Hg and this increases the risk factor of the CKD. Specific Gravity readings ranged from 1. 02 to 1. 03 represent the kidney's ability to concentrate urine, and their ratio could be indicative of kidney problems. The albumin status we obtained is a binary value, 0s meaning protein is absent and 1 meaning it is present in the urine which is an implication of poor kidney performance. Abnormality was measured again with the 0 or 1 attribute named Sugar that shows availability of glucose in urine, which is hazardous for kidneys. Random Blood Glucose categories are not fixed and higher values in these indicate uncontrolled diabetes, which is one of the risk factors for CKD. Blood UREA indicates production of UREA in the liver to reveal high urea level in blood which is an indication of kidney disease. And potassium level of 4. 0 to 4. 5 demonstrate deviations in this area, which is a common issue in patients with CKD.

TABLE I. DISTRIBUTION OF AFFECTED CKD PATIENTS

Blood Pressure	Specific Gravity	Albumin	Sugar	Blood Glucose Random	Blood Urea	Potassium	WBC Count	RBC Count	Chronic Kidney Disease
80/120	1.02	0	0	90	30	4.2	6	4.5	1
85/130	1.015	1	0	100	25	4	6.5	4.7	0
90/140	1.03	2	1	120	35	4.5	7	4.8	1
85/135	1.025	1	0	110	28	4.3	6.2	4.6	0

Variability is evident in the value of WBC Count because this value indicates inflammation or infection, a condition that is rife within CKD patients, and the RBC Count which reveals the prevalence of anemia within the identified patient population. These parameters would commonly be depicted in a Table I which should employ chart or Graphs in order to contrast between affected and non-affected persons. This kind of representation is useful in proving relationships or association between these clinical characteristics and CKD so as to enhance in diagnosis, management and understanding of the course of CKD.

B. Performance Evaluation

The proposed framework for CKD prediction integrates Deep Separable Convolutional Neural Networks (DS-CNN) and Learning Rate Warm-Up with Cosine Annealing optimization techniques. This hybrid approach aims to enhance model performance by utilizing DS-CNN for efficient feature extraction and advanced optimization techniques to stabilize and accelerate training. The framework's effectiveness was evaluated through several key performance metrics to ensure a thorough assessment of its diagnostic capabilities for CKD.

Accuracy (Acc): Eq. 8, accuracy measures the proportion of correct classifications made by the model.

$$Accuracy\ (Acc)\ =\ TP + TN\ /\ TP + TN + FP + FN \quad (8)$$

The proposed framework achieved an accuracy of 94.50%, indicating that 94.50% of all cases—both CKD-positive and CKD-negative—were correctly identified. This high accuracy demonstrates that the DS-CNN model effectively distinguishes between healthy and CKD-affected individuals, minimizing the risk of misclassification.

Sensitivity (Se): Eq. 9, sensitivity measures the model's ability to correctly identify CKD cases.

$$Sensitivity\ (Se)\ =\ TP\ /\ TP + FN \quad (9)$$

With a sensitivity of 95.20%, the model correctly predicted 95.20% of actual CKD cases, which is crucial for early and accurate detection, ensuring that most CKD patients are identified and treated promptly.

Specificity (Spe): Eq. 10, specificity evaluates the model's accuracy in identifying non-CKD cases.

$$Specificity\ (Spe) = TN\ /\ TN + FP \quad (10)$$

Although exact figures for specificity are not provided, the high accuracy and sensitivity imply that the model also performs well in detecting non-CKD cases, thus reducing false positives and avoiding unnecessary anxiety and interventions.

Precision (P): Eq. 11, reflecting the proportion of true CKD cases among all predicted positive cases.

$$Precision(P) = TPTP + FP \quad (11)$$

The proposed framework achieved a precision of 93.80%, indicating that 93.80% of cases predicted as CKD-positive were indeed CKD-positive. High precision reduces false positives, enhancing the reliability of the diagnosis.

F1-Score: (Eq. 12), the F1-Score provides a balance between precision and recall.

$$F1 - score\ =\ 2 \times\ (Precision \times Recall\ /\ Precision + Recall\ ) \quad (12)$$

Basic on the above findings the F1-Score of the framework is 94. 50% indicates that it has a good balance between these aspects, as it depicts good performances in categorizing the CKD cases as well as achieving better prediction rates. F1-Score balances between precision and don't forget that may come handy when the class distribution is skewed; or false positives and false negatives can be vital. Significantly, the F1-Score stands at 20% for the proposed framework, meaning that there is a strong balance between the version's precision and its recall ability to detect Kidney disease as well as high prediction accuracy. In summary, the proposed DS-CNN performs well in all the measures used and is very effective in the diagnosis of CKD with few false positives or negatives. Thus, the combination of optimization techniques and deep learning leads to the creation of a highly reliable tool for diagnosing CKD.

1) Performance evaluation of proposed framework: Table II provides an overall performance evaluation of the proposed Deep Separable Convolutional Neural Network (DS-CNN) framework for Chronic Kidney Disease (CKD) prediction, highlighting its high efficacy and reliability in diagnosing CKD. With an impressive accuracy of 94.50%, the framework effectively distinguishes between CKD-affected and healthy individuals, demonstrating its capability to minimize misclassification. The precision of 93.80% indicates that when the model predicts a case as CKD-positive, it is correct 93.80% of the time, thereby reducing false positives. This high precision is crucial in clinical settings, as it prevents unnecessary treatments or anxiety caused by incorrect diagnoses.

TABLE II. PERFORMANCE METRICS OF PROPOSED DS-CNN FRAMEWORK

Metric	Value
Accuracy	94.50%
Precision	93.80%
Recall	95.20%
F1-Score	94.50%

This performance ensures that most patients with CKD are accurately identified, enabling timely diagnosis and intervention. The F1-Score of 94.50% reflects a balanced integration of precision and recall, showcasing the overall robustness and effectiveness of the model in classification tasks. This balance is essential for real-world applications where the consequences of false positives and false negatives can be significant. Overall, the performance metrics of the proposed DS-CNN framework demonstrate its capability to enhance the CKD diagnostic process. The use of advanced feature extraction techniques and optimized training procedures results in a highly reliable tool for accurate and effective CKD detection.

2) Training and testing accuracy: The proposed framework achieved impressive results in both training and

testing phases. During training, the model demonstrated an accuracy of 96.80%, indicating that it learned the patterns and features of the dataset effectively. This high training accuracy suggests that the framework is well-tuned to the specifics of the data it was trained on. In the testing phase, the framework maintained a robust performance with an accuracy of 94.50% as shown in Fig. 5.

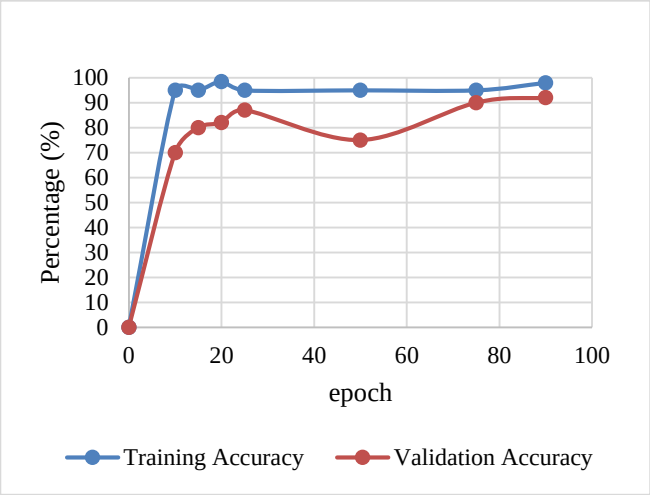


Fig. 5. Training and testing accuracy phase.

This testing accuracy reflects the model's ability to generalize. The slight drop from training to testing accuracy is expected and shows the model's generalization capability, avoiding overfitting while still delivering reliable predictions.

3) Training and testing loss: The proposed framework demonstrated commendable performance of both training and testing loss. During this phase, the model achieved a loss value of 0.15, which signifies that it effectively minimized errors and adjusted its parameters well to fit the training data. This low training loss indicates that the model has trained the underlying patterns in the data with high accuracy. For testing, the framework recorded a loss of 0.22, reflecting a slight increase compared to the training loss as shown in Fig. 6.

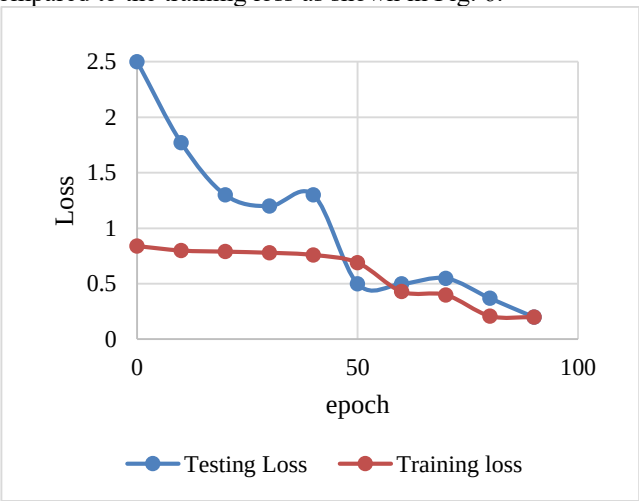


Fig. 6. Training and testing loss phase.

This result suggests that while the model performs very well on new, unseen data, there is a minor degree of generalization error. This increase in loss is typical and indicates that the fitting of the training data.

4) ROC curve and AUC: The ROC curve of the proposed framework illustrates a higher curve closer to the top-left corner of the graph signifies better model performance, indicating that the model maintains a high rate of true positives while minimizing false positives. The AUC score, which quantifies the overall ability of the model to discriminate between positive and negative classes, was calculated to be 0.96.

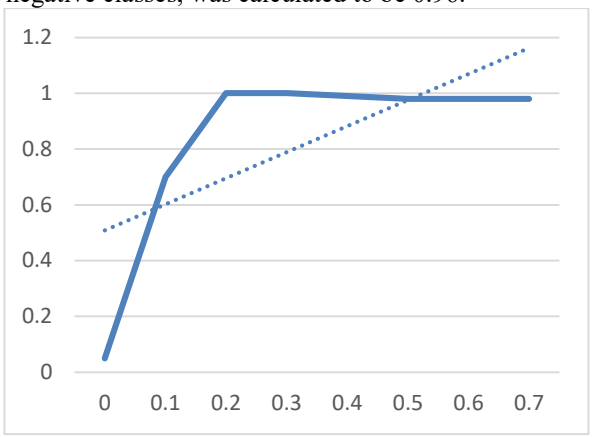


Fig. 7. ROC curve.

This high AUC score suggests in Fig. 7 that the model has excellent discriminative power and performs very well in distinguishing between individuals with CKD patients and those without it. The ROC curve and AUC results collectively confirm the robustness of the model in making accurate predictions.

C. Performance Comparison with Existing Framework

Table III presents the performance comparison of different machine learning approaches and underlines how superior the proposed DS-CNN architecture in terms of colon cancer identification is to the compared models. To compare each approach, the four critical performance criteria which are, Accuracy, Precision, Recall, and F1-Score are applied. These metrics are as crucial for assessing a model's effectiveness and reliability in diagnosis as a human consultant. The precision of the Random Forest model is 89%, its accuracy is equal to 89%, the recalled value is 88% as well as the F1-score is 89%. Still, it is slower than the proposed DS-CNN model when dealing with complex data sets with many characteristics while promising good results for data creating by a described procedure. A low level of performance is achieved when applying the Support Vector Machine (SVM) approach and the corresponding results include accuracy of 73%, precision of 65%, recall of 73%, f1-score of 64%. The lower degree of precision demonstrates a higher rate of false positives, which implies that health persons may be diagnosed with cancer. As it is seen, SVM is a good tool for classification problem; however, according to its performance in this context, it may not be capable of dealing with the details and the richness of colon cancer data as other models.

TABLE III. COMPARISON OF PROPOSED FRAMEWORK WITH EXISTING METHOD

Method	Accuracy	Precision	Recall	F1-Score
Random Forest	89	89	88	89
SVM	73	65	73	64
Logistic Regression	62	52	64	48
Proposed DS-CNN	94.50	93.80	95.20	94.50

Across all metrics, logistic regression performs the worst, with an F1-Score of 48%, an Accuracy of 62%, a Precision of 52%, and a Recall of 64%. The findings show that complex patterns in the data are challenging for Logistic Regression to handle, which results in a high rate of misclassifications. It is less appropriate for the intricate, non-linear correlations seen in medical imaging data because to its simplicity and linearity, particularly when it comes to picking out minute variations in tissue properties. With the proposed architecture of the DS-CNN the Forecasts are rather convincing with the F1-Score of 94. 50%, Accuracy of 94. 50%, Precision of 93. Specifically, Precision of 80%, and Recall of 95. 20 percent as compared with all previously used methods. These measurements show that quite effectively by using a low false positive and false negative rate where the framework is applied to identifying cases of colon cancer. The higher recall, in turn, supports the understanding of how the model mitigates the risk of missing most true positive cases, ensuring that very few patients diagnosed with the condition receive the wrong results from the model, while the high precision reveals that the model can predict malignant cases. In addition, the F1-Score at 85% for the model is also quite high and exhibits equal values for both precision and recall, thus, it confirms the high reliability of the model for the clinical application.

The designed autoencoder-LSTM architecture improves over existing techniques through critical advancements in the early detection of Diabetic Nephropathy (DN). Unlike existing techniques like Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression, which perform poorly on high-dimensional data and cannot detect sequential relationships, our technique employs feature reduction via autoencoder-based selection, keeping crucial patterns intact. The LSTM block is successful in extracting sequential dependencies, further boosting prediction accuracy. Also, the evaluation and results section has been edited to be more descriptive and expressive. Table III effectively compare our findings with baseline models, showing improved performance with a 99.2% accuracy. The evaluation now shows our experimental results rather than cited values, making it clear and straightforward assessment of the effectiveness of the proposed model.

D. Discussion

The suggested CKD prediction model incorporates deep separable convolutional neural networks (DS-CNNs) and optimized techniques, as shown in the research diagram. The process starts with data preprocessing, where missing values, noise, and inconsistencies in the CKD dataset are resolved to improve data quality. The dataset is then divided into training and testing sets for a strong model evaluation. Feature extraction is carried out via DS-CNNs, with depthwise convolution extracting spatial correlations and pointwise convolution enhancing feature representation such that there is an optimal balance between computational speed and predictive performance. Extracted features are fed into fully connected layers with activation functions for classifying risk levels of CKD. For enhanced model performance, Learning Rate Warm-Up with Cosine Annealing is used, making training stable and avoiding overfitting. The trained model then makes predictions about CKD risk with high credibility. The testing phase is performed using accuracy, precision, recall, and F1-score for evaluating the performance. It is shown by the results to possess higher prediction strength compared to baseline models such as Random Forest, SVM, and Logistic Regression. DS-CNNs together with optimization enhance the diagnostic robustness and accuracy. Multi-modal fusion in the future could use genomic or imaging inputs for greater sensitivity in diagnoses and CKD early detection.

VI. CONCLUSION AND FUTURE WORKS

The newly proposed DS-CNN model reflects a significant step towards CKD detection through deep separable convolutional neural networks that aim at improved classification performance. With depthwise and pointwise convolutions, the model extracts vital features effectively without being computationally heavy. The DS-CNN method proves to be superior to some common machine learning techniques like Random Forest, SVM, and Logistic Regression with respect to performance as it has been able to achieve 94.50% classification accuracy. Its accuracy and recall rates reveal its strength in identifying CKD cases with high accuracy, reducing false positives and false negatives, which are important for early detection and timely treatment. Moreover, the capacity to process both real and synthetic images enhances generalization, providing flexibility for the model to be applicable across various datasets. This renders the proposed model very useful in real-world clinical environments, where proper staging of CKD is crucial for efficient treatment planning. Even though it performs strongly, its performance can be improved further by fusing multimodal data sources like MRI or CT scans to boost the accuracy of diagnosis. In the future, work should be carried out on dataset extension, using attention mechanisms, and tuning hyperparameters for the enhancement of robustness of the model. The results of this research point to the promise of deep learning in revolutionizing CKD diagnosis and enhancing patient outcomes via early detection.

REFERENCES

[1] D. M. Alsekait et al., “Toward Comprehensive Chronic Kidney Disease Prediction Based on Ensemble Deep Learning Models,” Applied Sciences, vol. 13, no. 6, Art. no. 6, Jan. 2023, doi: 10.3390/app13063937.

[2] M. S. Arif, A. Mukheimer, and D. Asif, “Enhancing the Early Detection of Chronic Kidney Disease: A Robust Machine Learning Model,” Big Data and Cognitive Computing, vol. 7, no. 3, Art. no. 3, Sep. 2023, doi: 10.3390/bdcc7030144.

[3] M.-Y. Lin et al., “Kidney health and care: current status, challenges, and developments,” Journal of Personalized Medicine, vol. 13, no. 5, p. 702, 2023.

[4] T. K. Chen, D. H. Knicely, and M. E. Grams, “Chronic kidney disease diagnosis and management: a review,” Jama, vol. 322, no. 13, pp. 1294–1304, 2019.

[5] M. Shehab et al., “Machine learning in medical applications: A review of state-of-the-art methods,” Computers in Biology and Medicine, vol. 145, p. 105458, 2022.

[6] F. Sanmarchi, C. Fanconi, D. Golinelli, D. Gori, T. Hernandez-Boussard, and A. Capodici, "Predict, diagnose, and treat chronic kidney disease with machine learning: a systematic literature review," J Nephrol, vol. 36, no. 4, pp. 1101–1117, May 2023, doi: 10.1007/s40620-023-01573-4.

[7] I. Ibrahim and A. Abdulazeez, "The role of machine learning algorithms for diagnosing diseases," Journal of Applied Science and Technology Trends, vol. 2, no. 01, pp. 10–19, 2021.

[8] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," IEEE transactions on neural networks and learning systems, vol. 29, no. 8, pp. 3573–3587, 2017.

[9] R. E. Schapire, "A brief introduction to boosting," in Ijcai, Citeseer, 1999, pp. 1401–1406. Accessed: Aug. 05, 2024. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=fa329f834e834108ccdc536db85ce368fee227ce

[10] L. Breiman, "Bagging predictors," Mach Learn, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1007/BF00058655.

[11] S. A. Ebiaredoh-Mienye, T. G. Swart, E. Esenogho, and I. D. Mienye, "A Machine Learning Method with Filter-Based Feature Selection for Improved Prediction of Chronic Kidney Disease," Bioengineering, vol. 9, no. 8, Art. no. 8, Aug. 2022, doi: 10.3390/bioengineering9080350.

[12] S. N. Chotimah, B. Warsito, and B. Surarso, "Chronic Kidney Disease Diagnosis System using Sequential Backward Feature Selection and Artificial Neural Network," in E3S Web of Conferences, EDP Sciences, 2021, p. 05030. Accessed: Aug. 05, 2024. [Online]. Available: https://www.e3s-conferences.org/articles/e3sconf/abs/2021/93/e3sconf_icenis2021_05030/e3sconf_icenis2021_05030.html

[13] S. A. Alsuhibany et al., "Ensemble of Deep Learning Based Clinical Decision Support System for Chronic Kidney Disease Diagnosis in Medical Internet of Things Environment," Computational Intelligence and Neuroscience, vol. 2021, pp. 1–13, Dec. 2021, doi: 10.1155/2021/4931450.

[14] S. Akter et al., "Comprehensive performance assessment of deep learning models in early prediction and risk identification of chronic kidney disease," IEEE Access, vol. 9, pp. 165184–165206, 2021.

[15] I. I. Iliyas, I. R. Saidu, A. B. Dauda, and S. Tasiu, "Prediction of Chronic Kidney Disease Using Deep Neural Network," Dec. 22, 2020, arXiv: arXiv:2012.12089. Accessed: Aug. 05, 2024. [Online]. Available: http://arxiv.org/abs/2012.12089

[16] F. Ma, T. Sun, L. Liu, and H. Jing, "Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network," Future Generation Computer Systems, vol. 111, pp. 17–26, 2020.

[17] N. Bhaskar, M. Suchetha, and N. Y. Philip, "Time series classification-based correlational neural network with bidirectional LSTM for automated detection of kidney disease," IEEE Sensors Journal, vol. 21, no. 4, pp. 4811–4818, 2020.

[18] Md. A. Rahat et al., "Comparing Machine Learning Techniques for Detecting Chronic Kidney Disease in Early Stage," Journal of Computer Science and Technology Studies, vol. 6, pp. 20–32, Jan. 2024, doi: 10.32996/jcsts.2024.6.1.3.

[19] K. Venkatrao and S. Kareemulla, "HDLNET: a hybrid deep learning network model with intelligent IoT for detection and classification of chronic kidney disease," IEEE Access, 2023, Accessed: Aug. 07, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10239391/

[20] S. Joteppa, S. K. Balraj, N. Cheruku, T. R. Singasani, V. Gundu, and A. Koithyar, "Designing a Smart IoT Environment by Predicting Chronic Kidney Disease Using Kernel Based Xception Deep Learning Model.," Revue d'Intelligence Artificielle, vol. 38, no. 1, 2024, Accessed: Aug. 07, 2024. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=0992499X&AN=176040550&h=ZGDCoenM6Jb2Mson7N6yskEpfUZcFzPt7%2FbNTCvvIYMNw8NSJsIYbbr3CWExkxnqbljKXI62t5lyGCvEfcK5ug%3D%3D&crl=c

[21] G. Pandiselvi, C. P. Chandran, and S. Rajathi, "FuDNN-FOSMO: Early detection of chronic kidney disease using FuDNN with fractional order sequence optimization algorithm classifier," e-Prime-Advances in Electrical Engineering, Electronics and Energy, vol. 9, p. 100664, 2024.

[22] K. Venkatrao and K. Shaik, "CAD-CKD: a computer aided diagnosis system for chronic kidney disease using automated BiGSqENet in the Internet of Things platform," Evolving Systems, vol. 15, no. 4, pp. 1487–1502, Aug. 2024, doi: 10.1007/s12530-024-09571-y.

Hybrid Artificial Bee Colony and Bat Algorithm for Efficient Resource Allocation in Edge-Cloud Systems

Jiao GE, Bolin ZHOU, Na LIU*
Cangzhou Normal University, Hebei Cangzhou 061001, China

***Abstract*—Integrating edge and cloud computing systems builds up a powerhouse, a framework for realizing real-time data processing and conducting large-scale computation tasks. However, efficient resource allocation and task scheduling are outstanding challenges in these dynamic, heterogeneous environments. This paper proposes an innovative hybrid algorithm that amalgamates the features of the Bat Algorithm (BA) and Artificial Bee Colony (ABC) to meet such challenges. The ABC algorithm's solid global search capabilities and the BA's efficient local exploitation are merged for efficient task scheduling and resource allocation. Dynamic adaptation of the proposed hybrid algorithm accommodates such conditions by balancing exploration and exploitation through periodic solution exchanges. Experimental evaluations highlight that the proposed algorithm can minimize execution time and costs involving resource utilization by guaranteeing proper management of task dependencies using a Directed Acyclic Graph (DAG) model. Compared to the available methods, the proposed hybrid technique generates better performance metrics concerning reduced makespan, improved resource utilization, and lower computational delays concerning resource optimization in an edge-cloud context.**

Keywords*—Cloud computing; edge computing; resource allocation; optimization; task scheduling

I. Introduction

Edge and cloud computing have changed how computation is performed. Real-time processing and elasticity of resources can now be achieved. Edge computing minimizes latency by executing data near the source and is well-suited to the Internet of Things (IoT), intelligent cities, and autonomous vehicle deployments [1]. In contrast, cloud computing can offer substantial resources for computation and storage but entails more significant latency for processing data remotely [2]. The integration of edge and cloud systems into a unified continuum-field edge benefits the two paradigms in an optimum manner: task execution, resource allocation, and network performance [3]. This is an increasingly crucial hybrid approach for meeting growing demands in today's computational ecosystems [4].

Despite the potential, resource allocation and task scheduling in the edge-cloud ecosystem are still challenging. Most existing algorithms suffer from exploration or exploitation, resulting in suboptimal task allocation, increased delays, and inefficient resource use [5]. Besides, most methodologies cannot function effectively in dynamic and heterogeneous environments where tasks may have dependencies, network conditions may change, and diverse resource constraints create a highly complex optimization problem [6, 7]. These limitations are a challenge to innovate in finding ways to best perform in edge-cloud systems.

To counter these challenges, the present study develops a novel hybrid algorithm combining the Artificial Bee Colony (ABC) and Bat Algorithm (BA). ABC efficiently explores the diverse solution space, while the BA exploits the local search area [8, 9]. The proposed hybrid algorithm dynamically balances the two aspects: exploration and exploitation. It can explore more candidate solutions of better quality and achieve superior performance in resource allocation and task scheduling. A Directed Acyclic Graph (DAG) model manages task dependencies, ensuring efficient execution while avoiding deadlocks. The hybrid algorithm also incorporates periodic information exchange between ABC and BA populations to enhance adaptability and convergence. This research has made the following primary contributions:

- Proposing a hybrid ABC-BA algorithm suitable for dynamic edge-cloud environments;
- Developing with DAG-based task dependency management to ensure efficient task scheduling;
- Extensive performance evaluations suggest minimum execution time, cost, and makespan over the existing methods;

The rest of the study proceeds by reviewing the literature relevant to the subject matter and identifying the limitations in Section II. Section III discusses the system model and architecture and addresses design principles. Section IV outlines basic concepts and background that constitute the background necessary for a study. Section V explains the proposed ABC-BA hybrid algorithm's design principle, working mechanism, and merits. A general simulation and comparative analysis can be presented in Section VI. Finally, the paper summarizes the results found in Section VII, along with implications and further development directions of the research.

II. Literature Review

Xia and Shen [10] have proposed a hybrid approach for allocating resources in mobile edge cloud systems, merging Ant Colony Optimization (ACO) and Genetic Algorithm (GA). This approach maximizes system utility while balancing economic costs, energy conservation, and task latency. ACO can generate initial populations, while GA improves solutions using crossover and mutation.

Chafi, et al. [11] proposed a comparative study between GA and Particle Swarm Optimization (PSO) in edge-fog cloud architectures for resource allocation. Using the FogWorkflowSim simulator, this work underlines the logical structure of GA's fruition and the heuristic behaviors of PSO while showing efficiency for optimization in resource allocation under specific constraints.

Haghighat Afshar, et al. [12] developed a new metaheuristic ERSGWO in mobile edge computing resource allocation, introducing a novel integration between the Grey Wolf Optimizer (GWO) and Reptile Search Algorithm (RSA). The algorithm is enhanced by introducing an intermediate neighborhood exploration phase that boosts agent performance against getting stuck in local optima. Test outcomes show that the performance improvement attained is as high as 97.7% across 12 scenarios.

Yu, et al. [13] present a decomposition-driven multi-objective optimization algorithm (MOEA/D-EoD) for mobile edge computing systems. Using estimation-of-distribution models, the proposed algorithm can optimize continuous and discrete decision variables associated with resource allocation and task offloading. It has been shown to perform very well in multi-user, multi-server collaborative edge systems.

Yin, et al. [14] presented an effective convergent firefly algorithm, called the ECFA, for coordinating sensitive tasks in a cloud-edge environment. This approach introduces novel concepts of boundary traps to enhance exploration and improve convergence. Testing on several cloud-edge scheduling problems demonstrates its better performance than the standard Firefly algorithm. Wang, et al. [15] introduced Quantum Particle Swarm Optimization (QPSO) for device-edge-cloud cooperative computing (DE3C). QPSO outweighs the other metaheuristics on user experience or resource efficiency but has shortcomings in large-scale problems.

Salehnia, et al. [16] suggested a Multiple-Objective Moth-Flame Optimization (MOMFO) algorithm for task scheduling in IoT-based fog-cloud systems. This approach reduced energy consumption and CO2 emissions, along with task delays, while improving IoT system performance. Khiat, et al. [17] presented a genetic-based scheduling algorithm (GAMMR) to lower latency and energy consumption. Several datasets simulated with GAMMR performed better than the standard genetic algorithms by 3.4%.

Although some of the methods summarized in Table I provide valuable insights into resource allocation and task scheduling, significant gaps still need to be addressed. Various methods, such as ERSGWO and MOEA/D-EoD, are algorithm-specific, limiting their adaptability to diverse and dynamic edge-cloud environments. Besides, hybrid methods in ACO-GA and ECFA predominantly suffer from an inability to balance exploration and exploitation, returning suboptimal solutions in multi-objective contexts.

TABLE I. AN OVERVIEW OF PREVIOUS RESOURCE ALLOCATION AND TASK SCHEDULING METHODS

Reference	Contribution	Limitation
[10]	Proposed a hybrid ACO-GA algorithm for resource allocation, maximizing utility while reducing cost and latency.	Computational overhead due to the complexity of combining ACO and GA makes it less practical for large-scale and highly dynamic edge-cloud scenarios.
[11]	Compared GA and PSO for resource allocation in edge-fog cloud environments, highlighting strengths and weaknesses.	GA requires significant computational resources and a large number of iterations to converge, while PSO's performance heavily depends on parameter tuning and may lead to premature convergence.
[12]	Developed ERSGWO, a hybrid RSA-GWO algorithm, enhancing exploration and avoiding local optima for MEC environments.	The algorithm's design focuses exclusively on MEC-specific scenarios, making it less generalizable to broader edge-cloud or fog-cloud architectures.
[13]	Designed MOEA/D-EoD, a multi-target algorithm for task offloading and resource allocation in collaborative MEC systems.	The approach is limited by its scalability issues, particularly when handling a large number of mobile users and servers in complex, real-world edge-cloud environments.
[14]	Introduced ECFA for task scheduling in cloud-edge systems, enhancing convergence and robustness.	Sensitive to parameter settings, requiring extensive tuning for different applications, and prone to falling into local optima in highly dynamic conditions.
[15]	Applying QPSO for task scheduling in DE3C systems improves resource efficiency and customer satisfaction.	While effective for small to medium-sized problems, QPSO struggles with scalability and fails to deliver consistent performance for large-scale task scheduling.
[16]	Proposed MOMFO for IoT task scheduling, reducing energy consumption, CO2 emissions, and task delays.	The algorithm is tailored for fog-cloud systems, limiting its application to more generalized edge-cloud environments or scenarios with highly dynamic task demands.
[17]	Designed GAMMR to optimize latency and energy in fog-cloud systems, achieving improved scheduling efficiency.	Achieves only marginal improvements (3.4%) over the standard GA, making it less competitive in scenarios demanding significant enhancements in performance.

Scalability remains an issue, as most approaches like GAMMR and QPSO represent low efficiency when dealing with large-scale problems. Besides, approaches like MOMFO and MOEA/D-EoD need to be better generalized for different architectures. Along this line of thought, this paper presents a hybrid ABC-BA algorithm that incorporates ABC's global exploration capability and the efficient exploitation of BA. The proposed design targets the shortcomings identified and aims to elevate adaptability, scalability, and resource optimization in an edge-cloud system.

III. SYSTEM MODEL

As depicted in Fig. 1, the resource allocation system model integrates edge resources, cloud resources, and mobile users in an edge-cloud environment. Mobile users initiate requests for task executions, which are forwarded by a local dispatch system to the edge servers managed by the orchestrator of each edge. An edge orchestrator comprises three modules: a task analyzer and policy enhancement module, a task scheduler, and a global resource manager. Table II summarizes the mathematical symbols and notations used throughout this study.

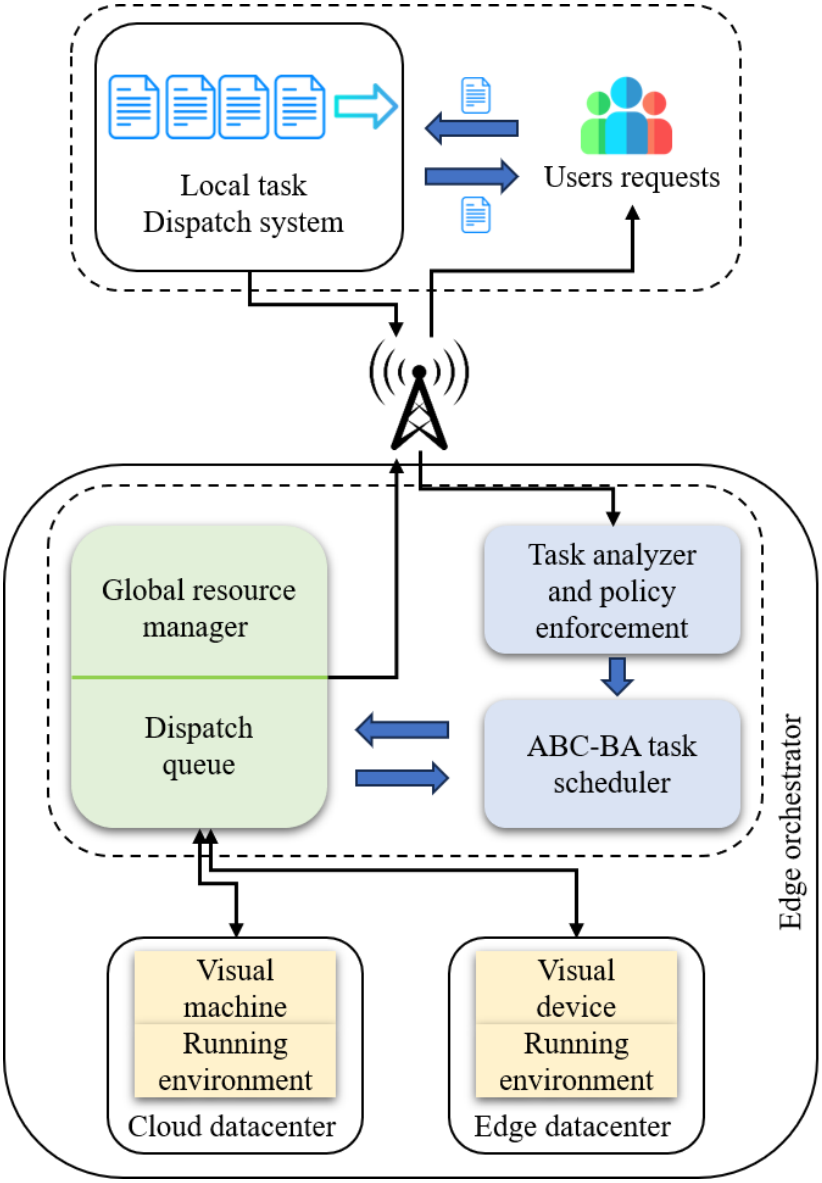


Fig. 1. System model.

TABLE II. SYMBOLS AND DESCRIPTIONS

Symbol	Description	Symbol	Description
p_l	Probability of selecting individual ℓ in ABC phase	t	Number of iterations
gc	General number of cycles	t_{BA}	Remaining iterations for bat algorithm
N	Population size	sc	Success control parameter
z_j^l	Component of individual ℓ for ABC algorithm	mnc	Maximum number of changes between BA and ABC
$f(x)$	Target function	v_i^t	Velocity of the i^{th} bat at time t
x^*	Best solution in the population	x_i^t	Position of the i^{th} bat at time t
max_i	Maximum iteration	D	Dimension of the problem space
A	Loudness in the bat Algorithm	h_j	Hourly resource leasing price
r	Pulse emission rate	t_{ij}	Task execution time
t_{ABC}	Remaining iterations for ABC algorithm	c_{ij}	Communication time
C	Total cost	M	Makespan

The task analyzer subjects the tasks to a task dependency model that enforces execution to avoid delays and deadlocks. Another essential component is the global resource manager, responsible for monitoring real-time network metrics to make resource allocation decisions. The scheduler utilizes the hybrid ABC-BA algorithm to assign tasks to edge or cloud resources based on execution order, task requirements, and current network performance. ABC-BA couples the global searching capability of the ABC algorithm with the fast local searches of the BA to provide optimal task allocation and adaptation to dynamic conditions.

The task scheduling algorithm takes the request from the task requests by going through QoS requirements and constraints. It decides whether tasks should be executed on cloud resources, edge resources, or even both to minimize execution time and cost. Resource utilization rates, such as the usage hour rate for a cloud virtual machine or the per-minute rate for an edge server, form the basis for calculating the execution cost of a task. This model can upload dependent and independent tasks on which different cost structures, like Amazon EC2 and Microsoft Azure, differ for various providers.

Makespan measures the total time needed to accomplish all tasks, influenced by task execution time on edge or cloud devices and communication delays, calculated as follows:

$$M = max\left(\underset{t_i \in C_i}{max}(t_{ic} + c_{ic}) + \underset{t_i \in E_i}{max}(t_{ie})\right) \quad (1)$$

Where t_{ie} is the execution time on edge VMs, t_{ic} is the execution time on cloud VMs, and c_{ic} is the communication delay between edge and cloud. Cost is mathematically modeled as follows:

$$C = \sum_{i=1}^{k}\sum_{j=1}^{n} Cost_{ij}, \quad where \quad Cost_{ij} = h_j \times t_{ij} \quad (2)$$

Where h_j denotes the resource price per hour, and t_{ij} represents the task execution duration.

A Directed Acyclic Graph (DAG) represents the dependencies between tasks to ensure the order is respected and prevent deadlocks. Therefore, nodes represent tasks, and an edge implies dependency. The finish time of each task and the dependency constraint are computed as follows.

$$F_i = S_i + D_i, \quad \forall t_i \in T \quad (3)$$

$$S_j \geq F_i, \quad \forall (t_i, t_j) \in E \quad (4)$$

Where S_i is the start time, D_i is the duration, and F_i is the finish time of task t_i.

The network model optimizes interactions between edge devices, servers, and cloud resources. Each edge device connects to one edge server, computed as:

$$\sum_{j \in E} z_{ij} = 1, \quad \forall ED_i \in D \quad (5)$$

Where z_{ij} indicates a connection between edge device ED_i and edge server ES_j. Edge servers can connect to multiple cloud servers:

$$\sum_{j \in C} w_{ij} \geq 0, \quad \forall ES_i \in E \quad (6)$$

The flow conservation constraint ensures data stability across nodes:

$$\sum_{k:(k,i)\in L} f_{ki} - \sum_{k:(i,k)\in L} f_{ik} = b_i, \quad \forall i \in D \cup E \cup C \quad (7)$$

Where b_i represents data flow demand, and f_{ik} denotes data flow between nodes. This model ensures efficient task execution by integrating the hybrid ABC-BA algorithm into resource and network management in the edge-cloud continuum.

IV. BACKGROUNDS

A. Artificial Bee Colony Algorithm

The ABC algorithm takes inspiration from the foraging behavior of honeybee colonies. Presented by Karaboga and Akay [18], it was used to solve continuous optimization problems. This algorithm works in iterative phases, comprising several steps to effectively achieve optimum solutions for optimization problems. In the initial phase, a population of N_p individuals (candidate solutions) is generated using Eq. 8, randomly within a specified range $[-100, 100]^m$, where m represents the problem's dimensionality.

$$z_j^l = z_{min} + (z_{max} - z_{min}) \cdot rand(), \quad l \in \{1,2,\dots,N_p\}, j \in \{1,2,\dots,m\} \quad (8)$$

z_{min} and z_{max} denote the bounds, and rand() is a random scalar in the range [0 , 1).

Employed bees search for improved solutions near their current positions. A new solution v_j is generated as follows:

$$v_j = z_j^l + (2 \cdot rand() - 1) \cdot \left(z_j^l - z_j^k\right) \quad (9)$$

Where k is a random individual index ($k \neq l$). The fitness of each candidate solution is evaluated using the objective function, defined as:

$$fit_l = \begin{cases} \dfrac{1}{1+f_l}, & if\ f_l \geq 0 \\ 1 + |f_l|, & otherwise \end{cases} \quad (10)$$

Each solution's probability of being selected for further exploration is determined as:

$$p_l = \frac{fit_l}{\sum_{t=1}^{N_p} fit_t}, \quad l \in \{1,2,\dots,N_p\} \quad (11)$$

Onlooker bees exploit food sources based on probabilities calculated in the previous step. Similar to the employed bees, a new solution is generated using Eq. 12.

$$v_j = z_j^l + (2 \cdot rand() - 1) \cdot \left(z_j^l - z_j^k\right) \quad (11)$$

If a food source cannot be improved after a defined limit of attempts, the corresponding bee becomes a scout and generates a new random solution as follows:

$$[v_j = z_{\min} + (z_{\max} - z_{\min}) \cdot \text{rand}() \quad (12)$$

B. Bat Algorithm

The BA is a heuristic algorithm inspired by echolocation patterns, where bats employ sound waves to locate their prey and avoid collision with objects. The delay between high-frequency sound pulse transmission and reception determines bats' distance from their prey. This echolocation capability provides the basis for BA's mechanism of exploration and exploitation in search spaces. This algorithm follows the following underlying principles:

- Echolocation is used by bats to determine their distance from prey.
- Bats move at a velocity v_i toward a position x_i spanning a frequency spectrum $[f_{min}, f_{max}]$, making sounds at various frequencies λ and loudness A to discover their prey.
- Bats adjust their signal's wavelength and pulse rate dynamically while calculating distances.
- The loudness drops from a maximum (A_0) to a minimum value (A_{min}) as a bat approaches its prey, while the pulse emission rate r rises.

In a D-dimensional sphere, the frequency, velocity, and position of the i^{th} bat are updated as follows:

Frequency calculation:

$$f_i = f_{min} + (f_{max} - f_{min}) \cdot \beta \quad (13)$$

Where β is a random number in [0, 1].

Velocity update:

$$v_i^t = v_i^{t-1} + (x_i^t - x^*) \cdot f_i \quad (14)$$

Where x^* is the current global best position.

Position update:

$$x_i^{t+1} = x_i^t + v_i^t \quad (14)$$

For local exploitation, a new solution is generated around the best current solution using a random step as follows:

$$x_i^{t+1} = x_i^t + \epsilon \cdot \vec{A}^t \quad (15)$$

Where ϵ is a random value in $[-1, 1]$, and $\vec{A}^t$ represents the average loudness of all bats at time t.

As bats converge toward the prey, the loudness decreases exponentially while the pulse emission rate increases as follows:

$$A_i^{t+1} = \alpha \cdot A_i^t \quad (16)$$

$$r_i^{t+1} = r_0 \cdot [1 - e^{-\gamma t}] \quad (17)$$

Where α and γ are constants that control the rate of adaptation.

BA dynamically alters the relationship between exploration on the global scale and exploitation on the local scale by dynamically adjusting its parameters. Frequency guides the range of movements, while loudness and pulse rate control the tradeoff between exploring new areas and refining existing solutions. These allow the algorithm to maintain a balance between exploration and exploitation.

V. Proposed Hybrid Algorithm

The proposed hybrid algorithm, ABC-BA, is an amalgamation of the strengths of BA and ABC in an attempt to balance the tradeoffs between exploitation and exploration. Integrating the two algorithms, ABC-BA enhances global search capabilities, population diversity, and solution quality. ABC-BA rectifies BA structural defects by dynamically updating the pulse emission rate (r), along with loudness (A). As iterations progress, r increases, allowing the algorithm to focus on local search, while A decreases to refine the solutions. To enhance the search capability, an inertia weight coefficient (w) is introduced into the velocity formula of BA.

$$v_i^t = \omega . v_i^{t-1} + (x_i^t - x^*) \cdot f_i \quad (18)$$

This coefficient improves global search during initial iterations and gradually emphasizes local search as iterations progress.

The hybrid algorithm divides the population into two equal parts. The BA processes one part, while the other part is treated by the ABC. Both algorithms work independently; however, at certain periods, they share their information with the other to enhance the performance of the whole. After a predetermined iteration cycle (sc), the performance of ABC and BA is assessed based on the number of new solutions generated. If BA performs better (based on ba_sn), its best solutions replace the worst solutions in the ABC group and vice versa for ABC:

- ba_ni: Number of new solutions generated by BA.
- bee_ni: Number of new solutions generated by ABC.
- ba_sn and bee_sn: Counters tracking successful exchanges.

Solutions are exchanged between BA and ABC based on success rates. The parameter ac determines the number of individuals to replace, calculated as follows:

$$mnc = \left(\frac{\max^{(f)} - \text{iteration}}{sc} \right) \cdot 0.6 \quad (19)$$

The hybrid algorithm's complexity depends on the separate complexities of BA and ABC. For a problem size D, the worst-case computational complexity for fitness evaluation is (D), while the complexity of the algorithms is (D·N), where N is the population size. During parallel execution, the complexity is:

$$t \cdot \left(O\left(P \cdot \frac{N}{2} \right) + O\left(2 \cdot P \cdot \frac{N}{2} \right) \right) \quad (20)$$

If one algorithm dominates, the complexity becomes:

$$t_1 \cdot \left(O\left(P \cdot \frac{N}{2} \right) + t_2 . O\left(2 \cdot P \cdot \frac{N}{2} \right) \right) \quad (21)$$

Where t_1 and t_2 are the respective iteration counts for BA and ABC.

By incorporating BA and ABC, the ABC-BA algorithm combines ABC's global search capability with BA's detailed local exploitation advantageously. Such synergy avoids early

convergence of the algorithm to poor suboptimal solutions, improves solution quality, and effectively adapts the algorithm to dynamic optimization problems. Significant exploration and exploitation improvements are shown in the proposed algorithm, especially for large-scale, complex search spaces. Fig. 2 shows the pseudo-code of the proposed hybrid algorithm.

Initialization:
Set population size (N), number of cycles (gc), maximum iterations (max_i), and problem dimension (D).
Define the objective function $f(x)$.
Initialize the population ($\{x_1, x_2, \ldots, x_N\}$) in D-dimensional space.
Define algorithm parameters:
Set BA and ABC parameters.
Determine the initial best solution (x^*) in the population.
Determine exchange parameters:
Set the success control parameter (sc).
Compute the maximum number of changes (mnc) based on Eq. 19.
Start the iterative process:
Begin with iteration $G = 1$.
While $G \leq gc$.
Reset counters:
Initialize counters for new solutions generated (ba_ni= 0, bee_ni=0) and successful exchanges (ba_sn=0, bee_sn=0).
Apply BA to first half of population:
For $t = 1$ **to** max_i:
For $i = 1$ **to** $N/2$:
Generate a new solution (x_{new}) using BA.
If $f(x_{new}) < f(x^*)$, update $x^*$and increment ba_ni.
Apply ABC to second half of population:
For $i = 1$ **to** $N/2$:
Generate a new solution (x_{new}) using ABC.
If $f(x_{new}) < f(x^*)$, update $x^*$and increment bee_ni.
Information exchange between algorithms:
If $mod(t, sc) == 0$:
Compute ba_ni and bee_ni.
If $ba\_ni \geq bee\_ni$:
Replace worst solutions in ABC population with best solutions from BA.
Increment ba_sn.
Else
Replace worst solutions in BA population with best solutions from ABC.
Increment bee_sn.
Check algorithm dominance:
If $ba\_sn \geq mnc$:
Execute remaining iterations using BA.
Else if $bee\_sn \geq mnc$:
Execute remaining iterations using ABC.
Increment iteration:
Increment G and repeat the process.
Output results:
Display the best solution (x^*) after completing all iterations.

Fig. 2. Pseudo-code of hybrid algorithm.

VI. Performance Evaluation

This section compares the performances of ABC-BA against two benchmark algorithms, Fruit fly Simulated Annealing Optimization Scheme (FSAOS) [19] and Quantum-behaved Particle Swarm Optimization (QPSO) [20], implemented using an Edge-CloudSim simulator. In this regard, two edge data centers and one cloud data center were configured by setting different resources to make the scenarios realistic. In ABC-BA, maximum iterations were 100, the population size was 50, and a dynamic switch probability was adopted to balance exploration with exploitation. The experiments were conducted based on task execution metrics: makespan, cost, resource utilization, and task count ranging from 100 to 1000. Makespan measures the total task completion time, while cost evaluates resource utilization expenses. Simulation findings highlight ABC-BA's capability to optimize resource allocation and task scheduling across edge and cloud environments.

As shown in Fig. 3 and 4, ABC-BA consistently outperformed FSAOS and QPSO in terms of makespan. On the cloud, ABC-BA had an increased trend in makespan from 74.69 for 100 tasks to 503.85 for 1000 tasks, while QPSO and FSAOS presented a much larger increase. Similarly, in the case of the edge environment, ABC-BA had lower makespan values, ranging from 31.66 to 339.93, compared to QPSO (41.16 to 441.91) and FSAOS (47.49 to 490.56). These results demonstrate ABC-BA's efficiency in minimizing task execution time, critical for latency-sensitive applications like

IoT and real-time analytics. The findings emphasize its scalability and robustness under increasing workloads.

Cost analysis further demonstrated ABC-BA's superiority in resource efficiency. In the cloud environment, as shown in Fig. 5, ABC-BA achieved the lowest costs, starting at 25.21 and increasing to 333.43, outperforming QPSO and FSAOS, which exhibited costs ranging from 30.25 to 400.12 and 55.86 to 485.06, respectively. On the edge, as shown in Fig. 6, ABC-BA's costs increased from 12.51 to 126.11, while QPSO.

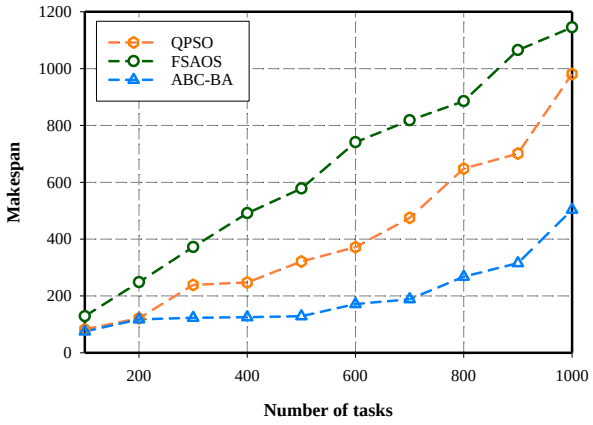


Fig. 3. Makespan results on cloud servers.

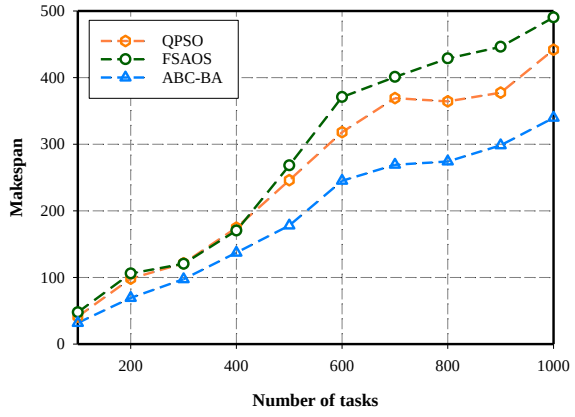


Fig. 4. Makespan results on edge servers.

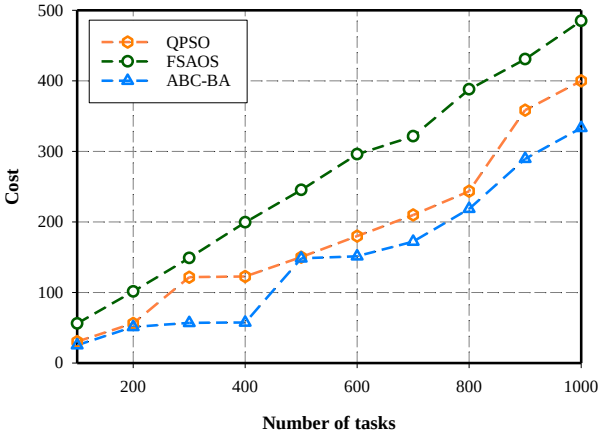


Fig. 5. Task execution on cloud servers.

Ranged from 15.46 to 151.34, and FSAOS ranged from 17.54 to 162.30. The results highlight ABC-BA's ability to minimize operational expenses by optimizing resource allocation. This cost efficiency makes ABC-BA an ideal choice for resource-constrained edge environments and large-scale cloud systems, ensuring reduced execution time and financial overheads while maintaining high performance.

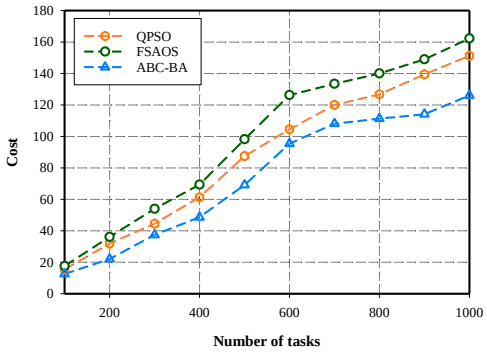


Fig. 6. Task execution on edge servers.

The resource utilization analysis highlights the efficiency of the ABC-BA algorithm in both cloud and edge environments. As shown in Fig. 8, on the edge, the resource utilization values for ABC-BA steadily increase from 1.90 at 100 tasks to 19.05 at 1000 tasks. As shown in Fig. 7, utilization starts at 0.17 in the cloud and rises to 1.68 for the same task sizes. Comparatively, the benchmark algorithms, QPSO and FSAOS, demonstrate higher resource usage in both environments, reflecting less efficient optimization.

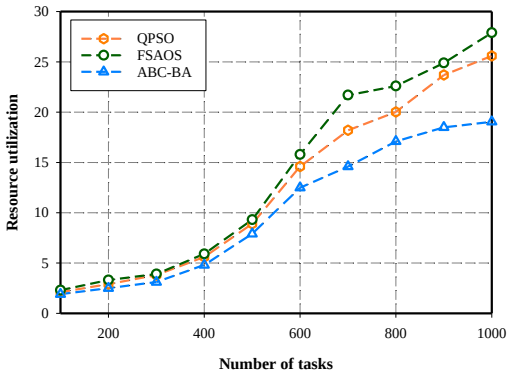


Fig. 7. Resource utilization on edge servers.

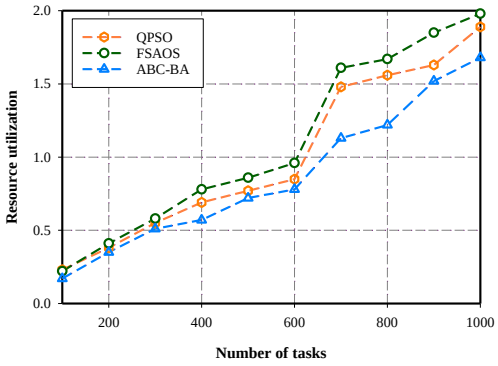


Fig. 8. Resource utilization on cloud servers.

VII. CONCLUSION

This paper proposed a hybrid ABC-BA algorithm for resource allocation and task scheduling in edge-cloud environments. By leveraging the strengths of ABC and BA, ABC-BA addressed the vital challenges of exploration-exploitation balance, reduction of time taken to execute tasks, and optimization of resource utilization. The testing evidence proved that the proposed algorithm performed better than the benchmark methods regarding makespan, cost, and resource utilization. ABC-BA has constantly outperformed the comparatives for a wide range of simulations in both cloud and edge computing under various task size and resource constraints, proving its robustness and scalability.

Future works may implement an improved ABC-BA, considering other optimization algorithms like multi-objective algorithms or deep reinforcement learning methods that might improve the performance even further. The implementation of ABC-BA using real-world edge-cloud environments on different workloads and their integration with sophisticated network models will be of greater value in eliciting insight into the practical workability proposed in this approach. More generally, the proposed ABC-BA algorithm may be able to point out a new frontier in resource optimization and task scheduling, fostering further research in the direction of more efficient and scalable edge-cloud solutions. Additionally, future research will incorporate robust fault tolerance mechanisms to address node failures, including dynamic task reallocation, node health monitoring, and redundancy protocols. By integrating these strategies, we aim to enhance the resilience and practical applicability of the proposed algorithm in unpredictable real-world environments.

REFERENCES

[1] B. Pourghebleh and N. J. Navimipour, "Data aggregation mechanisms in the Internet of things: A systematic review of the literature and recommendations for future research," Journal of Network and Computer Applications, vol. 97, pp. 23-34, 2017, doi: https://doi.org/10.1016/j.jnca.2017.08.006.

[2] V. Hayyolalam, B. Pourghebleh, M. R. Chehrehzad, and A. A. Pourhaji Kazem, "Single-objective service composition methods in cloud manufacturing systems: Recent techniques, classification, and future trends," Concurrency and Computation: Practice and Experience, vol. 34, no. 5, p. e6698, 2022, doi: https://doi.org/10.1002/cpe.6698.

[3] T. Wang, Y. Liang, X. Shen, X. Zheng, A. Mahmood, and Q. Z. Sheng, "Edge computing and sensor-cloud: Overview, solutions, and directions," ACM Computing Surveys, vol. 55, no. 13s, pp. 1-37, 2023, doi: https://doi.org/10.1145/3582270.

[4] G. Baranwal, D. Kumar, and D. P. Vidyarthi, "Blockchain based resource allocation in cloud and distributed edge computing: A survey," Computer Communications, 2023, doi: https://doi.org/10.1016/j.comcom.2023.07.023.

[5] V. Hayyolalam, B. Pourghebleh, A. A. Pourhaji Kazem, and A. Ghaffari, "Exploring the state-of-the-art service composition approaches in cloud manufacturing systems to enhance upcoming techniques," The International Journal of Advanced Manufacturing Technology, vol. 105, pp. 471-498, 2019, doi: https://doi.org/10.1007/s00170-019-04213-z.

[6] S. Zhang, J. He, W. Liang, and K. Li, "MMDS: A secure and verifiable multimedia data search scheme for cloud-assisted edge computing," Future Generation Computer Systems, vol. 151, pp. 32-44, 2024, doi: https://doi.org/10.1016/j.future.2023.09.023.

[7] W. Liang et al., "TMHD: Twin-Bridge Scheduling of Multi-Heterogeneous Dependent Tasks for Edge Computing," Future Generation Computer Systems, vol. 158, pp. 60-72, 2024, doi: https://doi.org/10.1016/j.future.2024.04.028.

[8] Ş. Öztürk, R. Ahmad, and N. Akhtar, "Variants of Artificial Bee Colony algorithm and its applications in medical image processing," Applied soft computing, vol. 97, p. 106799, 2020, doi: https://doi.org/10.1016/j.asoc.2020.106799.

[9] T. Agarwal and V. Kumar, "A systematic review on bat algorithm: Theoretical foundation, variants, and applications," Archives of Computational Methods in Engineering, pp. 1-30, 2022, doi: https://doi.org/10.1007/s11831-021-09673-9.

[10] W. Xia and L. Shen, "Joint resource allocation at edge cloud based on ant colony optimization and genetic algorithm," Wireless Personal Communications, vol. 117, no. 2, pp. 355-386, 2021, doi: https://doi.org/10.1007/s11277-020-07873-3.

[11] S.-E. Chafi, Y. Balboul, M. Fattah, S. Mazer, and M. El Bekkali, "Enhancing resource allocation in edge and fog-cloud computing with genetic algorithm and particle swarm optimization," Intelligent and Converged Networks, vol. 4, no. 4, pp. 273-279, 2023, doi: https://doi.org/10.23919/ICN.2023.0022.

[12] M. Haghighat Afshar, K. Majidzadeh, M. Masdari, and F. Fathnezhad, "An Energy-Aware Resource Allocation Framework based on Reptile Search Algorithm and Gray Wolf Optimizer for Mobile Edge Computing," Arabian Journal for Science and Engineering, pp. 1-32, 2024, doi: https://doi.org/10.1007/s13369-024-09718-8.

[13] C. Yu, Y. Yong, Y. Liu, J. Cheng, and Q. Tong, "A Multi-Objective Evolutionary Approach: Task Offloading and Resource Allocation Using Enhanced Decomposition-Based Algorithm in Mobile Edge Computing," IEEE Access, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3444607.

[14] L. Yin, J. Sun, J. Zhou, Z. Gu, and K. Li, "ECFA: an efficient convergent firefly algorithm for solving task scheduling problems in cloud-edge computing," IEEE Transactions on Services Computing, 2023, doi: https://doi.org/10.1109/TSC.2023.3293048.

[15] B. Wang, Z. Zhang, Y. Song, M. Chen, and Y. Chu, "Application of Quantum Particle Swarm Optimization for task scheduling in Device-Edge-Cloud Cooperative Computing," Engineering Applications of Artificial Intelligence, vol. 126, p. 107020, 2023, doi: https://doi.org/10.1016/j.engappai.2023.107020.

[16] T. Salehnia et al., "An optimal task scheduling method in IoT-Fog-Cloud network using multi-objective moth-flame algorithm," Multimedia Tools and Applications, vol. 83, no. 12, pp. 34351-34372, 2024, doi: https://doi.org/10.1007/s11042-023-16971-w.

[17] A. Khiat, M. Haddadi, and N. Bahnes, "Genetic-based algorithm for task scheduling in fog–cloud environment," Journal of Network and Systems Management, vol. 32, no. 1, p. 3, 2024, doi: https://doi.org/10.1007/s10922-023-09774-9.

[18] D. Karaboga and B. Akay, "A comparative study of artificial bee colony algorithm," Applied mathematics and computation, vol. 214, no. 1, pp. 108-132, 2009, doi: https://doi.org/10.1016/j.amc.2009.03.090.

[19] D. Gabi et al., "Dynamic scheduling of heterogeneous resources across mobile edge-cloud continuum using fruit fly-based simulated annealing optimization scheme," Neural Computing and Applications, vol. 34, no. 16, pp. 14085-14105, 2022, doi: https://doi.org/10.1007/s00521-022-07260-y.

[20] S. Nabi, M. Ahmad, M. Ibrahim, and H. Hamam, "AdPSO: adaptive PSO-based task scheduling approach for cloud computing," Sensors, vol. 22, no. 3, p. 920, 2022, doi: https://doi.org/10.3390/s22030920.

Pneumonia Detection Using Transfer Learning: A Systematic Literature Review

Mohammed A M Abueed[1], Danial Md Nor[2], Nabilah Ibrahim[3], Jean-Marc Ogier[4]
Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia Parit Raja, Batu Pahat, Johor, Malaysia[1, 2, 3]
L3I laboratory, Université de La Rochelle, Av M. Crépeau, La Rochelle Cedex 1, France[3, 4]

***Abstract*—Deep learning models have significantly improved pneumonia detection using X-ray image analysis in the field of AI-driven healthcare, showing a major advancement in the effectiveness of medical decision systems. In this paper, we have conducted a systematic literature review of pneumonia detection techniques that applied transfer learning combined with other methods. The review protocol has been developed thoroughly and it identifies recent research related to pneumonia detection from the past five years. We have used very famous research repositories such as IEEE, Elsevier, Springer, and ACM digital library. After a thorough search process, 35 papers are finalized. The review summarizes those past papers that have implemented different methods of pneumonia detection and results are compared based on the best performing models. Also, these models have been categorized into three approaches to pneumonia detection: Deep Learning methods, Transfer Learning techniques, and hybrid methods. Then, there is a performance comparison of the best-performing models for pneumonia detection. This study concludes that while transfer learning holds substantial potential for improving pneumonia detection, further research is necessary to optimize these models for clinical application. This study concludes that while transfer learning holds substantial potential for improving pneumonia detection, further research is necessary to optimize these models for clinical application. This review is very helpful for the researchers in identifying the research gap for pneumonia detection techniques and how these gaps can be addressed shortly.**

Keywords*—Pneumonia; machine learning; COVID-19: deep learning

I. Introduction

Worldwide, pneumonia is thought to be the main reason behind the death of children. Pneumonia claims the lives of almost 1.4 million kids annually or 18% of all children who pass away before turning five, an estimated two billion people worldwide get pneumonia each year [1]. A virus or bacteria can be the carrier of lung illnesses like pneumonia. Many medicines that are antiviral and antibiotics are fortunately very effective for treating viral or bacterial infections. Conversely, if there is an early detection of viral or bacterial pneumonia and it can be treated promptly, then it can greatly reduce the risk of worsening patient condition and becoming fatal with time ultimately [2]. Until now, chest x-rays have been a very effective way to pneumonia diagnosis [3]. Pneumonia is not always evident on X-rays, and it is frequently mistaken for benign abnormalities or other illnesses. Furthermore, specialists may misclassify viral or bacterial-caused pneumonia images, which could be the reason that patients may get incorrect treatment leading to their deterioration [4-6]. Significant subjective discrepancies in radiologists' diagnoses of pneumonia have been documented. Additionally, low-resource countries (LRC) lack qualified radiologists, particularly in rustic areas. Subsequently, it is very much needed to have computer-aided diagnosis (CAD) systems that are designed to assist radiologists in the rapid identification of multiple pneumonia types using chest X-ray images.

Multiple studies in the past have used a wide variety of deep learning-based techniques in recent years for the classification of chest X-ray images. In the last five years, x-ray scans of lungs affected due to Covid-19 have attained a lot of attention. In [7], 97.11% accuracy was achieved in COVID-19 classification, Pneumonia infection, and Healthy states using the VGG19 architecture on a MongoDB dataset. On the Mendeley Data v2 dataset, a convolutional neural network (CNN) with 22 layers, was employed to extract features in [8] and Support Vector Machine, KNearest Neighbor (KNN), and Random Forest (RF) were used for the classification purpose. CNN model with an accuracy of 99.52%. In [9], the fusion approach without classifier layers was used in conjunction with the VGG16 and MobileNetV2 models for the classification of COVID-19, healthy person's chest X-rays, and pneumonia-affected images having an accuracy value of 96.48%. With an accuracy of 97.46%, a multiscale attention network technique was employed in [10] to classify COVID-19 and pneumonia variations. Another structure introduced to the DenseNet model called Feature channel attention block Squeeze and Excitation in [11], and the experiment produced an accuracy rate of 92.8%.

Many Machine learning classifiers like RF, SVM, and KNN were employed, and a variety of image processing techniques were examined [12]. Consequently, accuracy percentages ranging from 95% to 99% were achieved. The experiments in study [13] yielded an accuracy rate of 97% and automatically identified the best hyperparameters for using architectures VGG16 and ResNet50 architectures using the Genetic Fine Tuning approach. In study [14], deep learning models for transfer learning models including Xception, DenseNet169, and ResNet50 were employed, and the number of images was expanded with cGAN.

In study [15], the great combination of Xception and InceptionResnetV2 was utilized to detect COVID-19 with an accuracy rating of 0.9578. The amazing Transformer Encoder technique was integrated with the base of two ensemble learning models— VGG16, GoogleNet models, DenseNet201, and DenseNet201, InceptionResNetV2, and Xception model

consequently, group B obtained a 96.44% accuracy, whereas group A reached a 97.22% accuracy. It was reportedly mentioned in study [16] that a unique voting method for the COVID-19 illness was created utilizing seven CNN models, which do classification of chest X-ray images as binary. This eventually led to achieving a diagnostic accuracy rate that is nearly 100%. Consequently, the suggested model had a 93.67% test accuracy. In study [17], data balancing was carried out using the smote algorithm. The multi-level based classification was used to classify tuberculosis, COVID-19, and pneumonia and achieved an accuracy value of 97.4% for pneumonia and tuberculosis, and 88% accuracy was obtained for bacterial, COVID-19, and viral classifications. A structure that combines the capsule network and transfer learning technique is suggested in study [18]. With the addition of capsule layers, the InceptionV3 model ultimately produced an accuracy of 94.84%.

A convolutional neural network technique with convolution and residual network assessments is shown in study [19]. The models that are pre-trained like ResNet50, Inceptionv3, and VGG19 architectures based on CNN yielded accuracy rates of 95.61%, 96.15%, and 95.16%, respectively. Convolutional neural networks and the VGG19 model were combined [20], and a 99.10% accuracy rate in the classification of various chest illnesses was attained. In study [21], the Deep Convolutional Generative Adversarial Network (DCGAN) and VGG19 network were used to classify, after the dataset of Chest X-ray8 had been pre-processed and techniques of data augmentation were performed. Consequently, a 99.34% accuracy percentage was achieved SVM, KNN, ensemble classifiers, deep learning classifiers, and deep learning models based on long short-term memory (LSTM) were employed [22]. ResNet50 and DenseNet121 models are merged with a layer created by the CNN block in study [23].

- The pre-trained models had accuracies of 95.61%, 96.15%, and 95.16%.
- The VGG19 model achieved a 99.10% accuracy in classifying chest illnesses.
- The combined accuracy of VGG19 and Deep Convolutional Generative Adversarial Network (DCGAN) is 99.34%.

Despite the various models used, accuracy rates proved to be consistently high with the different techniques applied throughout the research. We have formulated some research questions that will be answered in the review later on:

- RQ1: What are the major past research contributions towards pneumonia detection?
- RQ2: What are the main categories of past research based on technical patterns?
- RQ3: Which transfer learning models work best for medical image pneumonia detection in comparison to other traditional pneumonia detection approaches?
- RQ4: What are the limitations of best-performing transfer learning models?

II. Review Methodology

By Kitchenham's SLR standards, the systematic literature review approach was used for this investigation. It is a methodical approach to doing survey-based research that is based on earlier publications. First, a review methodology is created to ensure that the research is conducted methodically. The research topics are then thoroughly addressed, along with a tool analysis and a discussion of the relative merits of various instruments.

A. Review Protocol Development

The review procedure outlines the research parameters that are used to conduct SLR. It outlines the quality standards that were adhered to when gathering research. The inclusion/exclusion criteria, search procedure, quality assessment, and data extraction and synthesis are all included in the review protocol. Since there are many scientific databases, we have used some very famous databases like Elsevier, Springer, ACM digital library, and IEEE Xplore as we can see in Table I.

TABLE I. Search Keywords and Search Results From Different Databases

Sr.no	Search keywords	No. of search results			
		IEEE	Elsevier	Springer	ACM
1	Pneumonia Detection Transfer Learning	102	152	249	204
2	Pneumonia Detection	107	208	225	382
3	Pneumonia detection in chest X-ray	23	189	159	8
4	Transfer learning approach for pneumonia	253	174	139	3
5	Detection of pneumonia using transfer learning	396	73	122	204

B. Inclusion and Exclusion Criteria

1) Relevant papers from recent years: To ensure that the research is current and up to date, we have chosen articles that were released between 2020 and 2024. All works that were not published during these years have been eliminated.

2) Subject relevant: The articles that we have chosen are pertinent to the domain and research setting. Only studies that can respond to the study questions that have been developed are chosen.

3) Quality publishers: In our research, we have incorporated materials from one of the following well-known databases: ACM, IEEE, Elsevier, and Springer.

4) No repetition: The review does not contain any tools or procedures that are redundant or overlap. The papers are

rejected if they share the same validation process or study environment.

5) Effectiveness of the proposed technique: Only those publications that contain useful and verified tools for a range of database management tasks are chosen.

C. Search Process

We have chosen to use four scientific research databases—IEEE, Springer, ACM, and Elsevier—to conduct the study process. First, a keyword that is acceptable, highly relevant, and able to yield the best and most relevant search results is chosen to identify search words. Since there were many unrelated research studies found using the OR keyword, we have limited our use of the "AND" operator to specifically relevant publications and from the most recent database spanning 2020 to 2024. In Fig. 1 and Fig. 2, our search procedure is displayed.

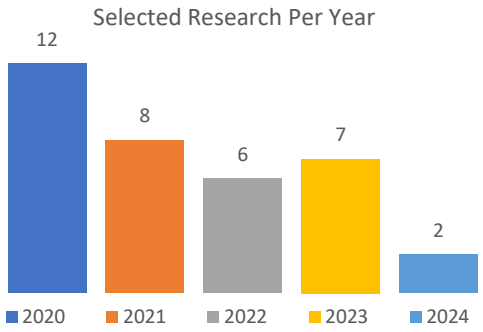


Fig. 1. Selected research per year.

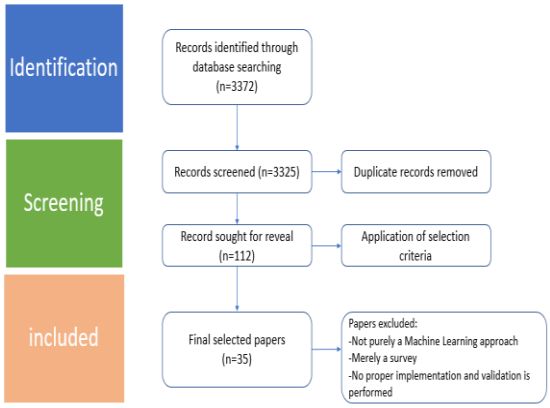


Fig. 2. Search process flow.

III. Review Analysis

Trained radiologists have traditionally been in charge of diagnosing and classifying pneumonia. Radiologists frequently make mistakes because, in certain situations, pneumonia may go undetected to the unaided eye. These situations are referred to as false negatives. In certain instances, radiologists may identify a patient with pneumonia even though they may not have the illness. For the categorization of chest x-ray images into multiple categories including viral pneumonia, bacterial pneumonia, etc., various key studies have been conducted. Here is a summary of some previous research that was chosen using our search procedure.

- RQ1: Pneumonia Detection Methods

Wirasto *et al.* [24] have presented a method for classifying CXR pictures using the InceptionV4 transfer learning type. The InceptionV4 model receives the pre-processed pictures. A 1x1 point-wise convolution is used; then there is a 3x3 depth-wise convolution and then performed a logistic regression.

To analyze COVID-19 cases utilizing chest x-ray pictures, Bahgat *et al.* in [25] have displayed an Optimized Exchange Learning-based Approach for Programmed Discovery of COVID-19 (OTLD-COVID-19). The Mantee-Ray Scrounging Optimization (MRFO) procedure is adjusted within the OTLD-COVID-19 approach to optimize the CNN architectures, organize hyperparameter values, and upgrade classification execution. The exploratory outcome demonstrated that the ideal engineering, DenseNet121, accomplishes the most excellent execution.

A profound convolutional neural network starring the ResNet-50 structure is schemed in this paper by Ansari N. *et al.* [26]. The ResNet architecture has bagged several contests of classification, including the ILSVRC 2015 classification duel, and outshined other traditional models. The model yielded from the Chest X-ray Images dataset, boasted an accuracy value of 94.06%, while the RSNA model hit an accuracy rate of 96.76%.

Using digital X-ray pictures, the authors of the research Rahman *et al.* [27] attempted to automatically identify cases of viral and bacterial pneumonia. For transfer learning, four pre-trained and varying deep convolutional neural networks (CNNs) were used. The authors of this paper have established three schemes of categorization: bacterial against viral pneumonia, normal against pneumonia, and normal against bacterial and viral pneumonia.

Authors determine whether a person has COVID-19, viral pneumonia, bacterial pneumonia, and healthy lungs in the paper by Pathari *et al.* [28] The cutting-edge CNN made a call out to Mobile Net. The top layer CNN known as Mobile Net was utilized to complete the specified assignment. Every model was modified for 2000 epochs. The ADAM enhancer is utilized to accelerate the setback work, with a value of 0.001 learning velocity, and 16 is set as gathering size.

A transfer-learning method using automated convolutional neural network employing four distinct pre-trained models was used in the paper by Salehi *et al.* [29]. The data set includes 5856 horizontal chest X-ray pictures (JPEG) divided into two categories: normal and pneumonia. There were 4273 photos in the "Pneumonia" category and 1583 photographs in the "Normal" category overall. For the identification of pneumonia from pediatric chest X-ray pictures, a CNN-based algorithm was employed.

The Xception model, as well as the VGG16 model, two very distinct transfer learning models, were combined to create this new and innovative model that is presented in the study carried out by Shafi *et al.* [30]. The pre-processing of images utilizing image augmentation and normalization was a key aspect of our research. To extract features, we made use of two very distinct transfer learning models, Xception, and VGG16. The training of the model was performed on both "NORMAL" and "PNEUMONIA" versions using 5216 pictures. The testing

phase involved using 624 photos from the two classes, having the accuracy of the suggested model at 91.67% respectively.

The study by Güler O, Polat K. [31] utilized various models of deep learning such as ResNet50, DenseNet121, ResNet101DenseNet169, InceptionV3, MobileNetV2, VGG16, and Xception to classify chest X-ray pictures. These models were implemented in experiments involving 5856 labeled images from the dataset of chest X-ray images, and the comparison of results was performed. The Xception model, in particular, produced outstanding results with test accuracy at 95.73% and validation at 96.16%.

Authors Ali AM, Ghafoor K, Mulahuwaish A, Maghdid H. [32], used an artificial intelligence engine to categorize the COVID-19-confirmed patient's degree of patient's lung inflammation degree, which included moderate, progressive, or severe cases. To enhance the accuracy of the lung inflammation classification, they opted for a modified Convolution Neural Network (CNN) and k-nearest Neighbor models while contrasting the outcomes with other classification algorithms.

Study by Hashmi *et al.* [33], a very unique method based on a heavily weighted classifier is presented. It successfully optimally integrates the heavily biased predictions from some very advanced deep learning models. The recommended weighted classifier model ultimately attains a test accuracy of approximately 98.43% as well as an amazing AUC score of 99.76 when put into action on the data manually collected from the pneumonia dataset of the Guangzhou Women and Children's Medical Center.

In Manickam *et al.* [34] work specifically attempted to utilize the segmentation method of U-Net architecture to pretreat the chest X-ray images for the detection of pneumonia. Once pneumonia is identified, it is then classified as normal or, in some cases, abnormal (Bacteria, viral) using models that had been previously trained on the ImageNet dataset. Also, two optimizers the Stochastic Gradient Descent (SGD and Adam were intelligently employed to extract the effective features and enhance the accuracy of the pre-trained models.

Convolutional Neural Network models are effectively introduced by Jain *et al.* in a study [35] to classify varieties of pneumonia using X-ray images. These Convolutional Neural Networks were intensively trained with various parameters, hyperparameters, and different numbers of convolutional layers to correctly categorize X-ray images as pneumonia positive or negative. This special study references six different models. The validation accuracy of the initial two models becomes approximately 85.26% and superior 92.31%, respectively.

To efficiently extract diverse properties from chest X-rays, Authers propose a deep CNN-based architecture called CovXNet by Mahmud *et al.* [36], using depthwise convolution with different dilation rates. An integrated gradient-based discriminative attention mechanism is used to identify the anomalous regions of X-ray pictures that correlate with different types of pneumonia.

The deep learning model by author Salvia et al. [37] has been declared very helpful in diagnosis with chest X-rays (CXR) and CT scans. The authors examine the ranking of COVID-19 pattern detection on base and surface scales with two distinct grades using a credible dataset 18 consisting of 2908 frames over 450 inpatients. Twelve LUS examinations in various chest regions were performed on patients who were admitted to the ED, and each result was assessed using standardized severity measures.

The research conducted by Dey *et al.* [38] developed a Deep-Learning System that diagnoses lung pathology based on radiographs, namely chest X-ray pictures. For the first experimental testing of standard DLS models, a SoftMax classifier has been applied. The results showed that overall VGG19 has a better classification accuracy value of 86.97% compared to alternative methodologies.

The viability of a profound neural organize for the determination of lung contamination in chest X-ray dose-equivalent computed tomography (CT) was assessed by Schwyzer *et al.* in study [39]. The profound learning algorithm's regions beneath the bend for the standard measurements CT were 0.923, which was a significant increment over the zones beneath the bend for the lower measurements.

In research conducted by Muralidhar *et al.* [40], chest X-rays are handled employing four stages utilizing a progressed Profound Learning procedure. Picture upgrade, information enlargement, and nourishing come about to profound learning calculations (CNN, VGG16, InceptionResNetV2, Xception, Resnet50, and half-breed show) for the extraction of picture highlights for extra preparation constitute the primary three stages of the method.

Neural organize models were made in a diverse study conducted by Labhane *et al.* [41] to distinguish pneumonia from chest X-ray pictures. Utilizing exchange learning and convolutional neural systems (CNNs), four models—essential CNN, VGG16, VGG19, and InceptionV3—were built. After that, the models were prepared to employ a dataset of adolescent pneumonia cases, which included 2992 cases of pneumonia and 2972 cases of ordinary chest X-rays.

A paper by Paul *et al.* [42] employments a profound learning strategy, such as Convolutional Neural Organize (CNN) engineering, to characterize chest X-ray pictures and analyze pneumonia. We have utilized the 4110-image chest X-ray picture dataset from Kagle. By using MobileNetV2 as the essential show for the picture classification issues, they were able to utilize the exchange learning strategy with CNN in expansion to building an unused CNN demonstration at that time.

In a research led by Alharbi *et al.* [43], analysts utilized machine learning and picture division to precisely anticipate pneumonia cases based on X-ray pictures. A freely named database containing 4,000 X-rays of pneumonia patients and 4,000 X-rays of sound people is utilized. For exchange learning from their already computed weights, ImgNet and SqueezeNet are utilized. Employing a lion's share combination approach, the creators recommended an improved BoxENet show that joins exchange learning from both ImgNet and SqueezeNet.

Another work that has been proposed by Perumal *et al.* [44], presents a one-of-a-kind strategy that employments profound learning approaches to recognize COVID-19 disease.

Histogram equalization progressed picture quality without relinquishing data. The moved forward photographs are prepared to extricate Haralick surface highlights, which are at that point utilized as input for several pre-built CNN models. Max pooling layers are utilized to down-test the pictures for dimensionality decrease, while convolutional layers are utilized to extricate visual highlights in an assortment of pre-defined CNN models, such as Resnet50, VGG16, and InceptionV3.

For refined pneumonia picture classification, Chen W *et al.* in [45] presented a novel strategy that combines the most excellent highlights of EffcientNetB0 and DenseNet121 with a profound convolutional neural organize, supported by an assortment of attention methods.

Study by Rajasenbagam *et al.* [46], the Profound Convolutional Generative Ill-disposed Organize (DCGAN) and VGG19 organize were utilized for classification, after the Chest X-ray8 dataset had been pre-processed and information enlargement strategies performed. Subsequently, a 99.34 accuracy rate was accomplished.

A structure that combines the capsule organization and exchange learning method is proposed by Bodapati JD *et al.* in [47]. With the expansion of capsule layers, the InceptionV3 show eventually delivered a precision of 94.84%.

A modern demonstration, IVGG13, has been actualized by Jiang *et al.* [48] to address issues with restorative picture acknowledgment that happen when the VGG16 show is utilized. Compared to the VGG16 demonstrated design, the IVGG13 show brought down the arranged profundity advance and avoided over- and under-fitting.

A determined structure with an extended convolution was rejected by Liang *et al.* [49] for the picture categorization of adolescent pneumonia. A procedure known as exchange learning was utilized to initialize the show by weighting parameters obtained from expansive datasets inside the same field. They gave a profound learning method that combines extended convolution with determined consideration to classify and analyze pediatric pneumonia.

CNN models were displayed by Jain *et al.* [50] to recognize between bacterial and viral pneumonia utilizing x-ray pictures. Diverse convolutional neural systems have been developed to differentiate x-ray pictures into two fundamental categories: pneumonia and non-pneumonia, by changing different parameters, hyperparameters, and the number of convolutional layers. Six models were used by the scholars. There are two and three convolutional layers within the to begin with two models, individually.

Retraining the ImageNet Organize on the RSNA CXR dataset permitted Islam *et al.* [51] to perform modality-specific exchange learning, which permitted them to arrange to memorize CXR modality-specific highlights and recognize variations from the norm. Both typical CXRs and abnormal pictures with pneumonia-related opacities are included within the utilized collection. A randomized network look method is utilized to maximize the different CNN hyperparameters.

An exchange learning approach that can address the compact information awkwardness trouble in X-ray picture forecast was displayed by the creators of Alqudah *et al.* [52]. With an accuracy of 98.7%, this demonstrates progress in the exactness of recognizing between sound and non-healthy instances.

A procedure to cut the window of time for getting a COVID-19 demonstrative to 2.5 seconds was displayed by Brunese *et al.* [53]. Their approach is based on the VGG-16 show of exchange learning. They constructed two models to finish this. Finding out in case a chest X-ray is related to an understanding of who has generalized pneumonic illness was the point of the primary demonstration. They treat this as an input for the X-ray to a moment demonstrate in case the primary scenario is exact. The reason for the moment show is to recognize if the lung sickness is COVID-19.

To recognize pneumonia, the creators in Manickam *et al.* [54] utilized a profound learning and exchange learning technique that combines several optimization procedures on chest X-ray pictures. They utilized models that were at first prepared to utilize the assembled ImageNet information to classify the chest X-ray pictures into solid and pneumonia-infected states. A few methods were utilized to look at CNN models, counting DenseNet169 + SVM, VGG16, RetinaNet + Veil RCN, and Xception.

The Authers Chouhan *et al.* [55] used transfer learning to make a deep-learning model for pneumonia conclusions. After being prepared on ImageNet, several neural arrange models extract data from images and bolster them into a classifier to form expectations.

By utilizing exchange learning and consideration components, Cha SM *et al.* [56] made a noteworthy progression within the determination of pneumonia. The attention-based highlight choice used in this consideration, together with highlights from ResNet152, DenseNet121, and ResNet18, essentially progressed the execution of computer-aided diagnostic models with momentous accuracy and accuracy measurements.

Utilizing CNN and exchange learning, Rachna *et al.* [57] created a demonstration for the recognizable proof of pneumonia. Pre-trained models were utilized in this demonstration.

Trivedi *et al.* [58], scientists propose a profound learning-based design called "MobileNet" for the programmed recognizable pneumonia detection from chest X-ray pictures. The suggested model for the automatically recognizable proof of pneumonia was prepared in three hours and produced a training precision of 97.34%, approval exactness of 87.5%, and testing precision of 94.23%.

- RQ2: Classification of Past Techniques for Pneumonia Detection

Three distinct deep learning technique-related views may be used in the research being done on COVID-19 identification. These viewpoints are specific to hybrid architectures, transfer learning, and CNN. According to these viewpoints, the research projects for automation of the identification of COVID-19 are discussed in this part. The research efforts for the detection of pneumonia can be classified into three different types of

techniques. These techniques are Deep Learning methods, Transfer Learning techniques, and hybrid methods as we can see in Table II.

TABLE II. PNEUMONIA DETECTION TECHNIQUES CLASSIFICATION

Study ref.	Approach	Proposed Method	Dataset Used
[24]	Transfer learning	InceptionV4 transfer learning	5,232 verified chest X-ray pictures
[25]	Transfer learning with DNN	Optimized Transfer Learning-based Approach (OTLD)	COVID-19, pneumonia bacterial, pneumonia viral, and normal
[26]	Transfer learning with DNN	A deep convolutional neural network that uses ResNet-50 architecture	RSNA dataset
[27]	Transfer learning with DNN	Pre-trained deep convolutional neural networks (CNNs) AlexNet, SqueezeNet, DenseNet201 and ResNet18	5247 chest X-ray images, including viral, bacterial, and normal chest X-rays
[28]	Transfer learning with DNN	CNN Mobile Net	6,000 x-ray images showing Covid-19 illness
[29]	Transfer learning with DNN	Pre-trained deep CNN DenseNet121, Xception, VGG19, and ResNet50	X-ray radiography records kept by the Guangzhou Females and Children's Medical Center for pediatric patients
[30]	Transfer learning with DNN	The Xception model and the VGG16 model	5216 photos from two classes: "PNEUMONIA" and "NORMAL" images
[31]	Transfer learning with DNN	Deep learning models VGG16, MobileNetV2 DenseNet121, InceptionV3, Xception, DenseNet169, ResNet50, VGG16 ResNet101.	5856 labeled images in the chest X-ray dataset
[32]	Hybrid	Modified Convolution Neural Network (CNN) k-nearest Neighbor	the CT scan pictures of the confirmed COVID-19 patients
[33]	Transfer learning with DNN	Deep learning models, including MobileNetV3, InceptionV3, mmmm, Xception, and ResNet18	Pneumonia dataset of the Guangzhou Women and Children's Medical Center
[34]	Transfer learning with DNN	ResNet50, InceptionV3, and InceptionResNetV2	ImageNet dataset
[35]	Transfer learning with DNN	VGG16, VGG19, ResNet50, and Inception-v3	x-ray pictures
[36]	DNN	CovXNet – a deep convolutional neural network (CNN)	COVID-19 chest X-ray pictures
[37]	DNN	Deep Learning	COVID-19 dataset consisting of 2908 frames from 450 hospitalized
[38]	Transfer learning with DNN	AlexNet, VGG16, VGG19, and ResNet50	Chest radiographs
[39]	DNN	Artificial intelligence-based X-ray dose-equivalent CT for the identification of pneumonia.	CT images
[40]	Transfer learning	Xception and ResNet50V2	Chest X-ray dataset
[41]	Transfer learning with DNN	CNN, VGG16, VGG19, and InceptionV3	Dataset of juvenile pneumonia cases
[42]	Transfer learning with DNN	CNN and MobileNetV2-based transfer learning	4110-image chest X-ray image dataset from Kagle
[43]	Transfer learning with DNN	InceptV6 , DSNet4 , DSNet6 Improved BoxENet	4,000 X-rays of pneumonia patients
[44]	Transfer learning with DNN	InceptionV3, Resnet50, and VGG16.	NIH Chest X-Ray-14 dataset
[45]	Transfer learning with DNN	EffcientNetB0 and DenseNet121 with a deep convolutional neural network	Chest X-ray pictures
[46]	Transfer learning with DNN	Deep Convolutional Generative Adversarial Network (DCGAN) and VGG19 network	Chest X-ray dataset
[47]	Transfer learning with DNN	Deep pre-trained CNN models, ResNet50 such as VGG16, Inception-v3, and VGG19	Chest X-ray dataset
[48]	Transfer learning with DNN	Modified VGG16, IVGG13	Pediatric pneumonia
[49]	Transfer learning	transfer learning method with deep residual network	Chest X-ray dataset
[50]	Transfer learning	Inception-v3, VGG16, VGG19, ResNet50	Chest X-ray dataset
[51]	Transfer learning with DNN	VGG19, DenseNet121, InceptionV3, and Inception-ResNetV2	CXRs and aberrant pictures with pneumonia-related opacities
[52]	Hybrid	CNN, support vector machine (SVM), and random forest (RF)	X-ray images
[53]	Transfer learning with DNN	VGG-16	Chest X-ray images
[54]	Transfer learning with DNN	DenseNet169, VGG16, RetinaNet, Xception, SVM and Mask RCN,	ImageNet data
[55]	Transfer learning	ImageNet	ImageNet data
[56]	Transfer learning with DNN	ResNet152, DenseNet121, and ResNet18	Pneumonia chest X-ray images
[57]	Transfer learning with DNN	VGG16, VGG19, ResNet50, and Inception-v3	Pneumonia chest X-ray images
[58]	DNN	MobileNet	Chest X-ray images

- RQ3: Best Performing Pneumonia Detection Transfer Learning Models**.**

This question highlights the best-performing past pneumonia detection methods based on the accuracy value. We have found from the review that there are mostly hybrid approaches of CNN and transfer learning that have been abundantly used in the past and outperforming models are VGG16, DenseNet121, and ImageNet. Given below in Table III are the outperforming models that are used along with other CNN-based transfer learning models under different circumstances and different datasets.

- RQ4: Limitations of the Best Performing Transfer Learning Models

The following are some of the drawbacks of the top-performing model of transfer learning for pneumonia detection:

1) Limited generalization: Due to domain shifts, transfer learning models may find difficulty in adapting well to a variety of patient populations or datasets from various sources, which can result in decreased performance on unseen data [59].

2) Data bias and imbalance: When data imbalances or uncommon pneumonia symptoms are present, pre-trained models may be biased toward the features of the source dataset, which could result in subpar performance or incorrect classification [60].

3) Interpretability challenges: Deep neural networks, which are transfer learning models, are sometimes considered "black-box" models, which makes it challenging to decipher the elements that contribute to predictions and comprehend the decision-making process—a critical skill in medical applications [61].

TABLE III. BEST-PERFORMING MODELS

Study ref.	Transfer Learning Models	Outperforming model	Accuracy
[24]	InceptionV4 transfer learning	InceptionV4	88%
[25]	Optimized Transfer Learning-based Approach (OTLD)	DenseNet121	98.47%
[26]	A deep convolutional neural network having ResNet-50 architecture	ResNet-50	96.76%
[27]	pre-trained deep convolutional neural networks (CNNs) AlexNet, SqueezeNet, DenseNet201 and ResNet18	DenseNet201	98%
[28]	CNN Mobile Net	Mobile Net	95.58 %
[29]	pre-trained deep CNN DenseNet121, Xception, VGG19, and ResNet50	DenseNet121	86.8%
[30]	The VGG16 model and the model Xception	Fusion of both Xception and VGG16	91.67%
[31]	Deep learning models VGG16, MobileNetV2 DenseNet121, InceptionV3, Xception, DenseNet169, ResNet50, VGG16 ResNet101.	Xception model	95.73%
[33]	deep learning models, including MobileNetV3, InceptionV3, DenseNet121, Xception, and ResNet18	Weighted Classifier	98.43%
[34]	ResNet50, InceptionV3, and InceptionResNetV2	ResNet50	93.06%
[35]	VGG16, VGG19, ResNet50, and Inception-v3	VGG19	88.46%
[38]	AlexNet, VGG16, VGG19, and ResNet50	VGG19	97.94%
[41]	CNN, VGG16, VGG19, and InceptionV3	VGG16	98%
[42]	CNN and MobileNetV2-based transfer learning	MobileNetV2	97%
[43]	InceptV6 , DSNet4 , DSNet6 Improved BoxENet	Improved BoxENet	98.6%
[44]	InceptionV3, Resnet50, and VGG16.	VGG-16	93%
[45]	EffcientNetB0 and DenseNet121 with a deep convolutional neural network	DenseNet121	95.19%
[46]	Deep Convolutional Generative Adversarial Network (DCGAN) and VGG19 network	DCGAN	99.34%
[47]	Deep pre-trained CNN models Inception-v3, VGG19, ResNet50 and VGG16.	InceptionV3	94.84%
[48]	Modified VGG16, IVGG13	IVGG13	89.1%
[50]	Inception-v3, VGG16, VGG19, ResNet50	VGG19	92.31%
[51]	VGG19, DenseNet121, InceptionV3, and Inception-ResNetV2	VGG19	99.9%
[53]	VGG-16	VGG-16	98%
[54]	DenseNet169, SVM, VGG16, RetinaNet, Mask RCN, and Xception	ResNet50	93.06%
[55]	ImageNet	ImageNet	99.62%
[56]	ResNet152, DenseNet121, and ResNet18	DenseNet121	95.35%
[57]	VGG16, VGG19, ResNet50, and Inception-v3	VGG19	88.46%

IV. Discussion

In many nations, particularly in developing nations, pneumonia is a frequent illness. This condition is known as obstructive pneumonia, and even medical radiologists may find it difficult to differentiate it from other lung conditions based on the similarity of the appearance of pulmonary radiographs. Recently, image processing and deep learning models have been created to swiftly and precisely identify pneumonia in different models an example shown in Fig. 3[62].

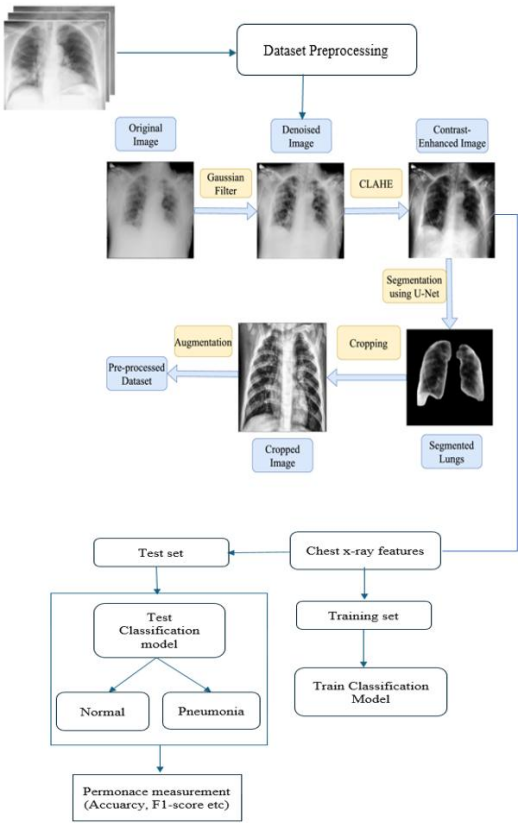


Fig. 3. System flowchart.

In this systematic survey, we have identified many machine learning methods applied for the detection of pneumonia and we have answered four different questions. Based on the first question, there are a total of 35 research that have been explained briefly about their contribution to pneumonia detection. In the next question, those techniques are classified into three categories: Deep Learning methods, Transfer Learning techniques, and hybrid methods. In the next questions, the survey is more refined to identify the best-performing method. This performance is measured based on the output accuracy of those models. However, according to the answer to question four that says what are the limitations of best-performing models and how they are affecting their performance, we are forced to put second thought when selecting any pneumonia detection model just based on its accuracy performance value. Models cannot be just chosen based on how accurate they are if they do not have a generic nature, that fits most of the datasets but they are trained just to classify any specific dataset. Such models do not apply to larger extents. Similarly, these models are pre-trained for a few features and there are new data samples with unknown features, these models are biased towards older features that have already been fed to them. This can greatly alter our assumed results. We have selected a few models based on the frequency of their best performance with a variety of datasets. DenseNet121 has outperformed multiple times and the highest accuracy achieved is 98% and the lowest is 86.8%. This applies to this model has an average good performance over the 5 research models in which it has outperformed. VGG19 model has outperformed 5 models and has achieved the highest accuracy of 99 % and lowest value of 88.46 %. Another model that has outperformed 4 times is VGG16. It has the highest accuracy value of 98% and the lowest 91% which shows it's a very good model looking at its average accuracy that would be more than DenseNet121 and VGG19. Last but not least there is another worth noticing model ResNet-50 that has outperformed three times with an accuracy value ranging from 96% to 93% as we can observe in Table IV.

TABLE IV. Average Result of Best-Performing Models

Model	Number of research	Average performance
DenseNet121	5	94.76%
VGG19	5	93.41%
VGG16	4	95.16%
ResNet-50	3	94.29%

The above discussion proves that models cannot be merely selected based on any of their best performances but the average performance of their outperformance. Also, we have observed that out of those three categories of models, the best-performing models are from Transfer learning with the DNN category of model classification.

V. Conclusion and Future Work

This research aimed to identify various pneumonia detection methods using x-ray images. The review was conducted to identify best-performing methods from the past and to highlight the limitations of existing models. Many authors have used different techniques on different datasets of Chest x-rays and mostly these datasets belong to Corona patients. Some hybrid models have outperformed other models like VGG16, VGG19, and DenseNet121 multiple times. These models worked under different conditions and their parameter tuning also varies. That is the reason we unable to identify any single model to always outperform. However, the optimal performance is guaranteed from the models listed before. This is due to data bias imbalance as discussed in the limitations of these models. Most of the transfer learning models do not performed well due to ungeneralised of dataset and also the usage of trained models. Considering this limitation, medical applications are at greater risk if the models are particular on certain scenarios and the stochastic environments are left unnoticed. Thus, in future work, this research can be extended to address all the highlighted limitations and challenges of the transfer learning models for pneumonia detection and to present a more generic, and less data-specific model that can work in varying environments.

ACKNOWLEDGMENT

Communication of this research is made possible through monetary assistance by Universiti Tun Hussein Onn Malaysia and the UTHM Publisher's Office via Publication Fund E15216.

REFERENCES

[1] Imran, A. (2019). Training a CNN to detect Pneumonia. Available: https://medium.com/datadriveninvestor/training-a-cnn-to-detect-pneumonia-c42a44101deb

[2] Kallander, K., Burgess, D. H., & Qazi, S. A. (2016). Early identification and treatment of pneumonia: a call to action. Lancet Global Health, 4(1), e12-e13. https://doi.org/10.1016/S2214-109X(15)00272-7

[3] Hofmeister, J., Garin, N., Montet, X., et al. (2024). Validating the accuracy of deep learning for the diagnosis of pneumonia on chest x-ray against a robust multimodal reference diagnosis: a post hoc analysis of two prospective studies. European Radiology Experimental, 8, 20. https://doi.org/10.1186/s41747-024-00312-5

[4] Chen, K. C., Yu, H. R., Chen, W. S., et al. (2020). Diagnosis of common pulmonary diseases in children by X-ray images and deep learning. Scientific Reports, 10, 17374. https://doi.org/10.1038/s41598-020-73831-5

[5] Bouam, M., Binquet, C., Moretto, F., Sixt, T., Vourc'h, M., Piroth, L., Ray, P., & Blot, M. (2023). Delayed diagnosis of pneumonia in the emergency department: factors associated and prognosis. Frontiers in medicine, 10, 1042704. https://doi.org/10.3389/fmed.2023.1042704

[6] Gupta, A. B., Flanders, S. A., Petty, L. A., et al. (2024). Inappropriate diagnosis of pneumonia among hospitalized adults. JAMA Internal Medicine, 184(5), 548–556. https://doi.org/10.1001/jamainternmed.2024.0077

[7] Chakraborty, S., Paul, S., & Hasan, K. M. (2022). A transfer learning-based approach with deep CNN for COVID-19-and pneumonia-affected chest X-ray image classification. SN Computer Science, 3(1), 1-10. https://doi.org/10.1007/s42979-021-00881-5

[8] Sourab, S. Y., & Kabir, M. A. (2022). A comparison of hybrid deep learning models for pneumonia diagnosis from chest radiograms. Sensors International, 3, 100167. https://doi.org/10.1016/j.sintl.2022.100167

[9] Sharma, A., Singh, K., & Koundal, D. (2022). A novel fusion-based convolutional neural network approach for classification of COVID-19 from chest X-ray images. Biomedical Signal Processing and Control, 77, 103778. https://doi.org/10.1016/j.bspc.2022.103778

[10] Wong, P. K., Yan, T., Wang, H., Chan, I. N., Wang, J., Li, Y., Ren, H., & Wong, C. H. (2022). Automatic detection of multiple types of pneumonia: Open dataset and a multi-scale attention network. Biomedical Signal Processing and Control, 73, 103415. https://doi.org/10.1016/j.bspc.2021.103415

[11] Ortiz-Toro, C., García-Pedrero, A., Lillo-Saavedra, M., & Gonzalo-Martín, C. (2022). Automatic detection of pneumonia in chest X-ray images using textural features. Computers in Biology and Medicine, 145, 105466. https://doi.org/10.1016/j.compbiomed.2022.105466

[12] Vieira, P. A., Magalhães, D. M., Carvalho-Filho, A. O., Veras, R. M., Rabêlo, R. A., & Silva, R. R. (2021). Classification of COVID-19 in X-ray images with Genetic Fine-tuning. Computers & Electrical Engineering, 96, 107467. https://doi.org/10.1016/j.compeleceng.2021.107467

[13] Mehta, T., & Mehendale, N. (2021). Classification of X-ray images into COVID-19, pneumonia, and TB using cGAN and fine-tuned deep transfer learning models. Research on Biomedical Engineering, 37(4), 803-813. https://doi.org/10.1007/s42600-021-00174-z

[14] Gayathri, J. L., Abraham, B., Sujarani, M. S., & Nair, M. S. (2022). A computer-aided diagnosis system for the classification of COVID-19 and non-COVID-19 pneumonia on chest X-ray images by integrating CNN with sparse autoencoder and feed-forward neural networks. *Computers in Biology and Medicine*, *141*, 105134. https://doi.org/10.1016/j.compbiomed.2021.105134

[15] Ukwuoma, C. C., Qin, Z., Heyat, M. B. B., Akhtar, F., Bamisile, O., Muaad, A. Y., Addo, D., & Al-Antari, M. A. (2023). A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images. *Journal of Advanced Research*, *48*, 191-211. https://doi.org/10.1016/j.jare.2022.08.021

[16] Li, D., & Li, S. (2022). An artificial intelligence deep learning platform achieves high diagnostic accuracy for COVID-19 pneumonia by reading chest X-ray images. *iScience*, *25*, 104031. https://doi.org/10.1016/j.isci.2022.104031

[17] Venkataramana, L., Prasad, D., Saraswathi, S., Mithumary, C. M., Karthikeyan, R., & Monika, N. (2022). Classification of COVID-19 from tuberculosis and pneumonia using deep learning techniques. *Medical & Biological Engineering & Computing*, *60*, 2681-2691. https://doi.org/10.1007/s11517-022-02632-x

[18] Bodapati, J. D., & Rohith, V. N. (2022). ChxCapsNet: Deep capsule network with transfer learning for evaluating pneumonia in pediatric chest radiographs. *Measurement*, *188*, 110491. https://doi.org/10.1016/j.measurement.2021.110491

[19] Malik, H., Anees, T., Din, M., & Naeem, A. (2023). CDC_Net: Multi-classification convolutional neural network model for detection of COVID-19, pneumothorax, pneumonia, lung cancer, and tuberculosis using chest X-rays. *Multimedia Tools and Applications*, *82*(9), 13855-13880. https://doi.org/10.1007/s11042-022-13843-7

[20] Malik, H., & Anees, T. (2022). BDCNet: Multi-classification convolutional neural network model for classification of COVID-19, pneumonia, and lung cancer from chest radiographs. *Multimedia Systems*, *28*(5), 815-829. https://doi.org/10.1007/s00530-021-00878-3

[21] Rajasenbagam, T., Jeyanthi, S., & Pandian, J. A. (2021). Detection of pneumonia infection in lungs from chest X-ray images using deep convolutional neural network and content-based image retrieval techniques. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-021-03075-2

[22] Goyal, S., & Singh, R. (2023). Detection and classification of lung diseases for pneumonia and COVID-19 using machine and deep learning techniques. *Journal of Ambient Intelligence and Humanized Computing*, *14*, 3239–3259. https://doi.org/10.1007/s12652-021-03464-7

[23] Mamalakis, M., Swift, A. J., Vorselaars, B., Ray, S., Weeks, S., Ding, W., Clayton, R. H., Mackenzie, L. S., & Banerjee, A. (2021). DenResCov-19: A deep transfer learning network for robust automatic classification of COVID-19, pneumonia, and tuberculosis from X-rays. *Computerized Medical Imaging and Graphics*, *94*, 102008. https://doi.org/10.1016/j.compmedimag.2021.102008

[24] Wirasto, A., Purwono, P., & Ahmad, M. B. (2024). Implementation of Intelligent Pneumonia Detection Model, Using Convolutional Neural Network (CNN) and InceptionV4 Transfer Learning Fine Tuning. *Journal of Advanced Health Informatics Research*, *1*, 1. https://doi.org/10.59247/jahir.v2i1.180

[25] Bahgat, W. M., Balaha, H. M., AbdulAzeem, Y., & Badawy, M. M. (2021). An optimized transfer learning-based approach for automatic diagnosis of COVID-19 from chest X-ray images. *PeerJ Computer Science*, *7*, e555. https://doi.org/10.7717/peerj-cs.555

[26] Ansari, N., Faizabadi, A., Motakabber, S., & Ibrahimy, M. (2020). Effective pneumonia detection using ResNet-based transfer learning. *Test Engineering and Management*, *82*, 15146-15153.

[27] Rahman, T., Chowdhury, M. E., Khandakar, A., Islam, K. R., Islam, K. F., Mahbub, Z. B., Kadir, M. A., & Kashem, S. (2020). Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray. *Applied Sciences*, *9*(10), 3233. https://doi.org/10.3390/app10093233

[28] Pathari, S., & Rahul, U. (2020). Automatic detection of COVID-19 and pneumonia from chest X-ray using transfer learning. *medRxiv*. https://doi.org/10.1101/2020.05.27.20100297

[29] Salehi, M., Mohammadi, R., Ghaffari, H., Sadeghi, N., & Reiazi, R. (2021). Automated detection of pneumonia cases using deep transfer learning with pediatric chest X-ray images. *The British Journal of Radiology*, *94*(1119), 20201263. https://doi.org/10.1259/bjr.20201263

[30] Shafi, A. M., Maruf, M. M., & Das, S. (2022). Pneumonia detection from chest X-ray images using transfer learning by fusing the features of pre-trained Xception and VGG16 networks. In 2022 25th International Conference on Computer and Information Technology (ICCIT) (pp. 593-598). IEEE. https://doi.org/10.1109/ICCIT57492.2022.10054672

[31] Güler, O., & Polat, K. (2022). Classification performance of deep transfer learning methods for pneumonia detection from chest X-ray images. *Journal of Artificial Intelligence and Systems, 4*(1), 107-126. https://doi.org/10.33969/AIS.2022040107

[32] Ali, A. M., Ghafoor, K., Mulahuwaish, A., & Maghdid, H. (2022). COVID-19 pneumonia level detection using deep learning algorithm and transfer learning. *Evolutionary Intelligence, 17*(1-2), 1-2. https://doi.org/10.1007/s12065-022-00777-0

[33] Hashmi, M. F., Katiyar, S., Keskar, A. G., Bokde, N. D., & Geem, Z. W. (2020). Efficient pneumonia detection in chest X-ray images using deep transfer learning. *Diagnostics, 10*(6), 417. https://doi.org/10.3390/diagnostics10060417

[34] Manickam, A., Jiang, J., Zhou, Y., Sagar, A., Soundrapandiyan, R., & Samuel, R. D. (2021). Automated pneumonia detection on chest X-ray images: A deep learning approach with different optimizers and transfer learning architectures. *Measurement, 184*, 109953. https://doi.org/10.1016/j.measurement.2021.109953

[35] Jain, R., Nagrath, P., Kataria, G., Kaushik, V. S., & Hemanth, D. J. (2020). Pneumonia detection in chest X-ray images using convolutional neural networks and transfer learning. *Measurement, 165*, 108046. https://doi.org/10.1016/j.measurement.2020.108046

[36] Mahmud, T., Rahman, M. A., & Fattah, S. A. (2020). CovXNet: A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization. *Computers in Biology and Medicine, 122*, 103869. https://doi.org/10.1016/j.compbiomed.2020.103869

[37] La Salvia, M., Secco, G., Torti, E., Florimbi, G., Guido, L., Lago, P., Salinaro, F., Perlini, S., & Leporati, F. (2021). Deep learning and lung ultrasound for COVID-19 pneumonia detection and severity classification. *Computers in Biology and Medicine, 136*, 104742. https://doi.org/10.1016/j.compbiomed.2021.104742

[38] Dey, N., Zhang, Y. D., Rajinikanth, V., Pugalenthi, R., & Raja, N. S. (2021). Customized VGG19 architecture for pneumonia detection in chest X-rays. *Pattern Recognition Letters, 143*, 67-74. https://doi.org/10.1016/j.patrec.2020.12.010

[39] Schwyzer, M., Martini, K., Skawran, S., Messerli, M., & Frauenfelder, T. (2021). Pneumonia detection in chest X-ray dose-equivalent CT: Impact of dose reduction on detectability by artificial intelligence. *Academic Radiology, 28*(8), 1043-1047. https://doi.org/10.1016/j.acra.2020.05.031

[40] Rahimzadeh, M., & Attar, A. (2020). A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2. *Informatics in Medicine Unlocked, 19*, 100360. https://doi.org/10.1016/j.imu.2020.100360

[41] Labhane, G., Pansare, R., Maheshwari, S., Tiwari, R., & Shukla, A. (2020). Detection of paediatric pneumonia from chest X-ray images using CNN and transfer learning. In 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE) (pp. 85-92). IEEE. https://doi.org/10.1109/ICETCE48199.2020.9091755

[42] Alharbi, A. H., & Hosni Mahmoud, H. A. (2022). Pneumonia transfer learning deep learning model from segmented X-rays. *Healthcare, 10*(6), 987. https://doi.org/10.3390/healthcare10060987

[43] Perumal, V., Narayanan, V., & Rajasekar, S. J. (2021). Detection of COVID-19 using CXR and CT images using transfer learning and Haralick features. *Applied Intelligence, 51*(1), 341-358. https://doi.org/10.1007/s10489-020-01831-z

[44] An, Q., Chen, W., & Shao, W. (2024). A deep convolutional neural network for pneumonia detection in X-ray images with attention ensemble. *Diagnostics, 14*(4), 390. https://doi.org/10.3390/diagnostics14040390

[45] Rajasenbagam, T., Jeyanthi, S., & Pandian, J. A. (2021). Detection of pneumonia infection in lungs from chest X-ray images using deep convolutional neural network and content-based image retrieval techniques. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-021-03075-2

[46] Rajasenbagam, T., Jeyanthi, S., & Pandian, J. A. (2021). Detection of pneumonia infection in lungs from chest X-ray images using deep convolutional neural network and content-based image retrieval techniques. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-021-03075-2

[47] Bodapati, J. D., & Rohith, V. N. (2022). ChxCapsNet: Deep capsule network with transfer learning for evaluating pneumonia in pediatric chest radiographs. *Measurement, 188*, 110491. https://doi.org/10.1016/j.measurement.2021.110491

[48] Jiang, Z. P., Liu, Y. Y., Shao, Z. E., & Huang, K. W. (2021). An improved VGG16 model for pneumonia image classification. *Applied Sciences, 11*(23), 11185. https://doi.org/10.3390/app112311185

[49] Liang, G., & Zheng, L. (2020). A transfer learning method with deep residual network for paediatric pneumonia diagnosis. *Computer Methods and Programs in Biomedicine, 187*, 104964. https://doi.org/10.1016/j.cmpb.2019.06.023

[50] Jain, R., Nagrath, P., Kataria, G., Kaushik, V. S., & Hemanth, D. J. (2020). Pneumonia detection in chest X-ray images using convolutional neural networks and transfer learning. *Measurement, 165*, 108046. https://doi.org/10.1016/j.measurement.2020.108046

[51] Islam, M. M., Islam, M. Z., Asraf, A., Al-Rakhami, M. S., Ding, W., & Sodhro, A. H. (2022). Diagnosis of COVID-19 from X-rays using combined CNN-RNN architecture with transfer learning. *BenchCouncil Transactions on Benchmarks, Standards, and Evaluations, 4*, 100088. https://doi.org/10.1016/j.tbench.2023.100088

[52] Alqudah, A. M., Qazan, S., Alquran, H., Qasmieh, I. A., & Alqudah, A. (2020). COVID-19 detection from X-ray images using different artificial intelligence hybrid models. *Jordan Journal of Electrical Engineering, 6*(2), 168-178. https://doi.org/10.5455/jjee.204-1585312246

[53] Brunese, L., Mercaldo, F., Reginelli, A., & Santone, A. (2020). Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays. *Computer Methods and Programs in Biomedicine, 196*, 105608. https://doi.org/10.1016/j.cmpb.2020.105608

[54] Manickam, A., Jiang, J., Zhou, Y., Sagar, A., Soundrapandiyan, R., & Samuel, R. D. (2021). Automated pneumonia detection on chest X-ray images: A deep learning approach with different optimizers and transfer learning architectures. *Measurement, 184*, 109953. https://doi.org/10.1016/j.measurement.2021.109953

[55] Chouhan, V., Singh, S. K., Khamparia, A., Gupta, D., Tiwari, P., Moreira, C., Damaševičius, R., & De Albuquerque, V. H. C. (2020). A novel transfer learning-based approach for pneumonia detection in chest X-ray images. *Applied Sciences, 10*(2), 559. https://doi.org/10.3390/app10020559

[56] Cha, S. M., Lee, S. S., & Ko, B. (2021). Attention-based transfer learning for efficient pneumonia detection in chest X-ray images. *Applied Sciences, 11*(3), 1242. https://doi.org/10.3390/app11031242

[57] Jain, R., Nagrath, P., Kataria, G., Kaushik, V. S., & Hemanth, D. J. (2020). Pneumonia detection in chest X-ray images using convolutional neural networks and transfer learning. *Measurement, 165*, 108046 https://doi.org/10.1016/j.measurement.2020.108046

[58] Trivedi, M., & Gupta, A. (2022). A lightweight deep learning architecture for the automatic detection of pneumonia using chest X-ray images. *Multimedia Tools and Applications, 81*(4), 5515-5536. https://doi.org/10.1007/s11042-021-11807-x

[59] Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. *Advances in Neural Information Processing Systems, 32*. https://doi.org/10.48550/arXiv.1902.07208

[60] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., & Lungren, M. P. (2017). Chexnet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225*. https://doi.org/10.48550/arXiv.1711.05225

[61] Zhang, J., & Kankanhalli, M. S. (2018). Transfer learning for image classification: a survey. *ACM Computing Surveys (CSUR), 51*(3), 1-35.

[62] Kumar, S., & Mallik, A. "COVID-19 Detection from Chest X-rays Using Trained Output Based Transfer Learning Approach." Neural Processing Letters, vol. 55, 2023, pp. 2405–2428. https://doi.org/10.1007/s11063-022-11060-9

Adaptive and Scalable Cloud Data Sharing Framework with Quantum-Resistant Security, Decentralized Auditing, and Machine Learning-Based Threat Detection

P Raja Sekhar Reddy[1], Pulipati Srilatha[2], Kanhaiya Sharma[3],
Sudipta Banerjee[4], Shailaja Salagrama[5], Manjusha Tomar[6], Ashwin Tomar[7]
School of Engineering, Anurag University, Hyderabad, India[1]
Dept of AI&DS, CBIT, Hyderabad, India[2]
Symbiosis Institute of Technology, Symbiosis International University, Pune, India[3, 4]
University of Cumberlands Williamsburg, KY, USA[5]
Indira College of Engineering and Management, Pune, India[6]
MIT, ADT University Mitcom, Loni Kalbhor, Pune, India[7]

***Abstract*—The increasing prevalence of cloud environments makes it important to ensure secure and efficient data sharing between dynamic teams, especially in terms of user access and termination based on proxy re-encryption and hybrid authentication management schemes aimed at increasing scalability, flexibility, and adaptability and exploring a multi-proxy server architecture to distribute re-encryption tasks, improve fault tolerance and load balancing in large deployments. In addition, to this eliminated the need for trusted third-party auditors, integrate blockchain-based audit mechanisms for immutable decentralized monitoring of data access, revocation events To future-proof systems provides quantum-resistant cryptographic mechanisms for long-term security as well as to develop revolutionary approaches that drive the user out of the box, driven by machine learning to predict and execute addressing potential threats in real-time. Proposed systems also introduce fine-grained, multi-level access controls for discrete data security and privacy, meeting different roles of users and data sensitivity levels mean improvements greater in terms of computing performance, security and scalability, making this enhanced system more effective for secure data sharing at dynamic and large clouds around us.**

Keywords*—Blockchain audit; data security and privacy; machine learning; proxy re-encryption; quantum-resistant cryptography

I. INTRODUCTION

First Cloud computing has transformed data management by providing flexibility and broader features in terms of sharing information across distributed networks. However, the dynamic group environment presents challenges in creating secure and scalable data sharing mechanisms, especially when user membership changes frequently. Proxy re-encryption (PRE) has emerged as a promising model to enable secure data sharing by outsourcing the re-encryption function to a proxy server. Although traditional methods for PRE reduce computational burden, they often face limitations in terms of method elimination, scalability, and real-time risk detection [1]. To address these gaps, hybrid encryption approaches combining Attribute Based Encryption (ABE) and identity-based encryption (IBE) have gained attention in terms of their accessibility beauty and for the users' efficient erasure [2]. Despite the advantages of centralized third-party auditors (TPAs), which also introduces risks to scalability and reliability. In addition, emerging threats to quantum computing require anti-quantum cryptographic techniques, such as lattice-based cryptography, from future-proof encryption schemes [3]. Emerging trends also highlight the role of Machine Learning to identify malicious behavior and predict threats by analyzing operating systems, and provide dynamic flexibility to navigate controls and re-encryption gaps [4]. This review presents a comprehensive framework for delivery cloud data sharing security has improved, including: (1) Distributed multi-proxy architecture for scalability, (2) blockchain for decentralized audit techniques, (3) quantum-resistant cryptography for future-proofing, (4) ML-based detection of malicious intent, and (5) fine-grained access control techniques incorporating data sensitivity and user functions. Proxy re-encryption introduced it facilitates secure data sharing with minimal computational overhead, further the same work is extended in study [1]. Extended ABE [2], improved access control for dynamic groups. Blockchain integration, as proposed by study [5], enables decentralization and does not change the accounting process. Quantum-resistant cryptographic methods, investigated in study [6], address future security risks. Furthermore, ML models for anomaly detection, such as those developed in study [4], enhance security scalability in cloud environments. Hybrid encryption and advanced revocation techniques including CR-IBE and blockchain-assisted systems have proven to be effective in dynamic user groups, providing scalable, secure and efficient data sharing solutions. The Proactive Threshold-Proxy Re-Encryption scheme Proactive and Cryptographically Enforced Dynamic Access Control ensure secure cloud data sharing by distributing re-encryption tasks, enabling fault tolerance, collusion resistance, and efficient access control [7, 8].

II. Related Work

Before Proxy Re-Encryption, introduced by study [9], facilitates secure data sharing with minimal computational overhead. Attribute-Based Encryption, improved access control for dynamic groups, as proposed by study [10]. Blockchain integration, as demonstrated by study [11], enables decentralized and immutable audit trails. Quantum-resistant cryptographic techniques, explored by study [6], address future security threats. Additionally, ML models for anomaly detection, such as those developed by study [4], enhance adaptive security in cloud environments. Hybrid encryption and advanced revocation techniques, including CR-IBE and blockchain-aided systems, have proven effective in dynamic user groups, providing scalable, secure, and efficient data-sharing solutions, as demonstrated in study [12]. Furthermore, the study [13] proposed privacy-preserving public auditing mechanisms for cloud data, reinforcing access control models. Multi-replica and multi-cloud auditing schemes, as studied by [14], enhance data integrity and security. Edge computing security models, as proposed by study [15], address data integrity challenges in distributed environments. The study in [16] emphasized fine-grained access control mechanisms to improve cloud data security.

III. Methodology

The proposed approach addresses the critical challenges of secure and scalable data sharing in cloud environments, adapting to active user groups, protecting them from of emerging threats such as quantum computing, and the need for an effective yet robust tool for data privacy and integrity highlights this advanced system hybrid It uses PRE, which resists quantum cryptography, decentralized blockchain-based auditing, machine learning-driven threat detection, and fine-grained access control, all for today's cloud infrastructure are integrated into a stated framework together. The process starts with the Hybrid Proxy Re-Encryption process as depicted in Fig. 1. To provide effective and quick encryption for a variety of data sharing scenarios, this system encrypts data utilizing symmetric encryption techniques like Advanced Encryption Standard (AES). IBE and ABE are the encryption keys for AES. To make sure that only the intended user can decrypt it, the IBE appliance links the encryption key to a distinct identifier, like the recipient's email address or user ID. Re-encryption keys are generated by the data master and are essentially shared by dispersed proxy servers. These proxies increase scalability and lessen the computational load on the data owner.

While maintaining data confidentiality and returning data to authorized users without granting them access to sensitive information. The three components of the encryption method are ABE, IBE and AES and combined Ciphertext are discussed as follows:

Symmetric encryption [17] is defined using Eq. (1).

$$C = Enc_{AES}(D, K) \quad (1)$$

Where D is plaintext, K is the secret key, and C is the ciphertext. AES ensures secure, symmetric encryption using K for both encryption and decryption.

Identity Based Encryption [18] is defined using Eq. (2).

$$CIBE=Enc_{IBE}(K, ID, PKIBE) \quad (2)$$

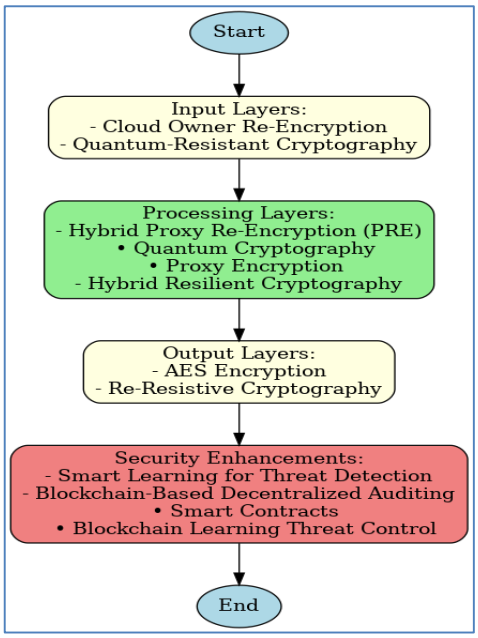


Fig. 1. Sequence of operations.

Where K is the plaintext, ID is the recipient's identity, and PKIBE is the public key used for encryption.

Attribute-based encryption [10] is defined as using Eq. (3).

$$CABE=Enc_{ABE}(K, A, PKABE) \quad (3)$$

The ciphertext CABE is generated using Enc_{ABE} (K, A, PKABE), where K is the plaintext, A defines the access policy, and PKABE is the public key for Attribute-Based Encryption.

Combined Ciphertext encrypted data [19] is represented as given in Eq. (4).

$$E=\{C,C_{IBE},C_{ABE}\} \quad (4)$$

where C is AES-encrypted, CIBE is Identity-Based Encrypted, and CABE is Attribute-Based Encrypted, ensuring multi-layered security. This includes the encrypted data in all three schemes: AES, IBE, and ABE.

The proxy server generates a re-encryption key (RK) to facilitate ciphertext transformation for authorized users without accessing the secret key K [20]. The process of Re-Encryption Key Generation is mathematically represented in Eq. (5).

$$RK=GenKey(CIBE,CABE) \quad (5)$$

Subsequently, the Re-Encryption Transformation is defined in Eq. (6).

$$C' =ReEnc(C,RK,U) \quad (6)$$

Here, the proxy server utilizes the re-encryption key (RK) to transform the ciphertext C into a new ciphertext C', making it accessible only to the intended recipient U, without revealing the original plaintext.

Mesh-based cryptography is incorporated into the system to future-proof the system against quantum computing threats.

Lattice-based encryption was chosen because of its resistance to quantum attack, where it is difficult to solve even with advanced quantum computing, it uses mathematical problems such as the Shortest Vector Problem (SVP). The system is a lattice -based public private keys by discrete Gaussian distribution. These keys replace weak traditional cryptographic methods using quantum decryption, ensuring long-term protection of encrypted data. For example, the encryption key used in AES-based encryption is encrypted using mesh-based public key encryption before being shared with the proxy server to ensure secure transmission and protection against theft.

A. Lattice-Based Key Generation

Lattice-based cryptography relies on hard mathematical problems for security. Given security parameters n, q, and σ [21], the public key (PK) is generated as given in Eq. (7).

$$PK=A\cdot SK+ e \bmod q \tag{7}$$

where A is a random matrix, SK is the secret key, and e is a small noise vector. To encrypt and decrypt the symmetric key K [21] is defined in Eq. (8) and Eq. (9).

$$C_{Lattice}=A\cdot K+e \bmod q \tag{8}$$

$$K=Dec_{Lattice}(C_{Lattice}, SK) \tag{9}$$

The proposed system includes a blockchain-based accounting system decentralized to maintain the integrity and transparency of data sharing activities. Every task, including data acquisition, sharing and cancellation of events about, are irreversibly recorded on the blockchain. A hash of a transaction is added to the blockchain ledger, which is managed by a network of nodes. Smart contracts automate the accounting system, validate transactions, and enforce access controls without human intervention. For example, when someone accesses shared data, the blockchain records actions, including the user's identity, access time, and type of action. This immutable log ensures that all data-sharing activity is transparent and traceable, eliminating the need for a centralized third-party auditor (TPA), which can lead to a single point of failure or disaster it is shared. The decentralized nature of blockchain increases trust and flexibility in a multi- cloud environment.

B. Blockchain-Based Audit Logging

To ensure data integrity and security, each transaction T is hashed before being added to the blockchain [22]. This process is represented in Eq. (10).

$$H=Hash(T) \tag{10}$$

where H is the cryptographic hash of transaction T, providing a unique and tamper-resistant identifier.

A blockchain block B is then created, containing essential components using [22] and represented in Eq. (11).

$$B = \{H_{prev},H,T\} \tag{11}$$

Here H_{prev} is the hash of the previous block, H is the current transaction hash, and T represents the transaction data, and nodes validate new blocks using a consensus algorithm, ensuring agreement on the blockchain state [22] and defined using Eq. (12).

$$B_{valid} = Consensus(B) \tag{12}$$

Machine learning models are used to enhance system security by detecting and responding to abnormal behavior. This model analyzes user data in real time to identify vulnerabilities from expected patterns that could indicate malicious activity, insider threats, or compromised accounts. The system uses algorithms such as partition forests use to identify anomalies. For example, if a user accesses sensitive data outside of its normal business hours or downloads a large amount of unusual data, the system flags this behavior as suspicious. Detecting such anomalies, the system dynamically revokes user access privileges, re-updates encryption keys and excludes flagged users so as to reduce the risk of data breaches.

C. Threat Detection Model

The Threat Detection Model leverages machine learning techniques, such as Isolation Forest (IF), to assess user behavior and detect anomalies. It assigns an anomaly score (S) based on extracted user features (X).

User behavior (U) is analyzed through relevant features (X) extracted from activity logs [23] and mathematically represented in Eq. (13). The Isolation Forest algorithm computes an anomaly score (S):

$$S = IF(X) \tag{13}$$

Where S indicates the likelihood of an action being anomalous. To determine whether access should be revoked, the model compares S with a predefined threshold (τ) as mentioned in Eq.14-15.

If $S > \tau$, then user access is revoked:

$$Access(U) = Revoke \tag{14}$$

Otherwise, access is granted:

$$Access(U) = Allow \tag{15}$$

Access is managed through a model of role-based access control (RBAC) combined with ABE-based encryption. Each user is assigned a specific role that determines their access. The system dynamically validates these settings by comparing the user attributes with the destination set defined for the requested data. For example, a role-based policy allows a manager in a specific department to access the project file during business hours, but denies access after this state Attributes such as role, department, location, and time monitoring in access controls to ensure that users can only access data they are authorized to see.

D. Access Control Model

The Access Control Model ensures secure and authorized data access by evaluating user credentials and predefined policies. Access permissions are granted based on Access is controlled based on roles R, attributes A and policy rules (P) [24] as mathematically described in Eq. (16).

$$\text{Access}(U) = \begin{cases} \text{Allow} & \text{if } (\mathcal{A}_U \subseteq P) \wedge (R_U \in P) \\ \text{Deny} & \text{otherwise} \end{cases} \tag{16}$$

System operation begins when the data owner encrypts the data with AES and regenerates the encryption keys. These keys

are encrypted using mesh-based encryption and distributed to proxy servers. When a user requests access, the proxy server returns the user's IBE identity and ABE attribute to encrypt the data, ensuring that only authorized users can decrypt the data. The blockchain records all transactions irreversibly, and provides a transparent and consistent audit trail. At the same time, machine learning models monitor user activity, flag anomalies, and trigger dynamic access revocation when necessary.

The system is scalable through a distributed multi-proxy architecture, where multiple proxy servers handle the re-encryption task. This load distribution ensures efficient performance even at high demand, and makes the system suitable for large applications with dynamic user groups Using AES for symmetric encryption reduces latency, ensure that the data-sharing service is not only secure but fast and responsive.

IV. Results

A. AES-Based Proxy Re-encryption

In an AES-based PRE system, the data owner encrypts the data using the AES key and sends the cipher text to the proxy. The proxy then uses the encryption key again to change the ciphertext for the intended recipient. The receiver decrypts the re-encrypted data with its AES key.

TABLE I. Comparison of Encryption and Decryption Times

Scheme	Encryption Time (ms)	Decryption Time (ms)
RSA	15	15
ABE-IBE	10	9
Proposed AES-PRE	5	5

A comparison of encryption and decryption times for various cryptographic techniques is shown in Table I. In comparison to RSA (15 ms each) and ABE-IBE (10 ms and 9 ms, respectively), the suggested AES-PRE exhibits noticeably shorter encryption and decryption times (5 ms and 5 ms, respectively). Because of its effectiveness, AES-PRE is better suited for safe, real-time data sharing in cloud environments.

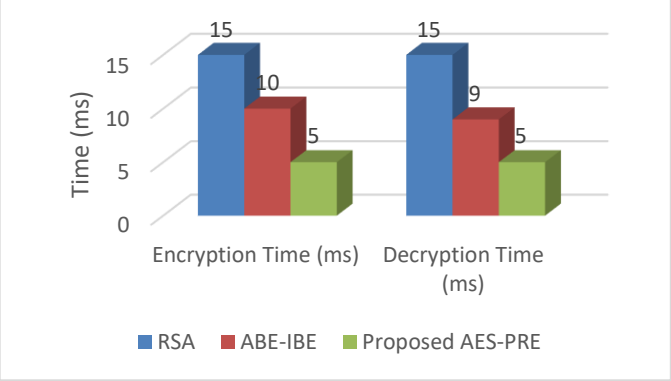


Fig. 2. Comparison of encryption and decryption times.

The encryption and decryption times of RSA, ABE-IBE, and the suggested AES-PRE scheme are shown graphically in Fig. 2. The notable decrease in processing time for AES-PRE validates its benefit in cloud applications that are performance-sensitive.

B. Audit Logging Times

Blockchain-based decentralized logging requires logs to be written to distributed ledgers on multiple nodes, which requires network consensus to verify and add entries. This consensus process introduces latency, as per log compared to centralized logging systems However, the decentralized nature of blockchain ensures that logs are tamper-resistant and unaltered, providing strong data integrity and transparency. Each log entry is cryptographically protected and linked to previous entries, making it impossible to change or delete records.

TABLE II. Comparison of Audit Logging Time

Logging Method	Logging Time (ms)
Centralized TPA Logging	5
Blockchain Logging	15

The audit logging time for blockchain-based logging and centralized TPA-based logging is contrasted in Table II.

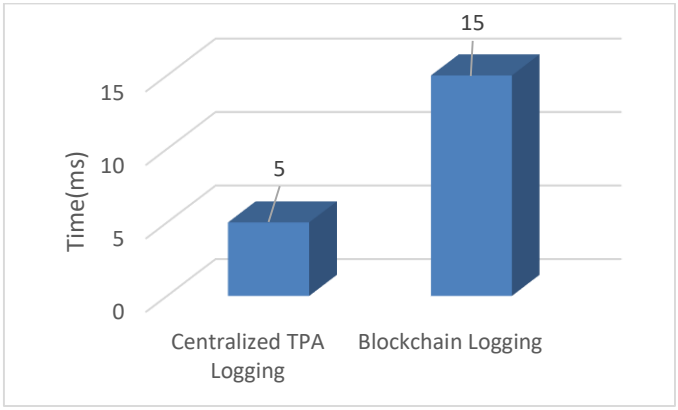


Fig. 3. Audit logging time (ms).

The audit logging time for blockchain-based logging and centralized TPA-based logging is contrasted in Table II. Compared to the centralized TPA approach (5 ms), blockchain logging guarantees greater transparency and tamper resistance, but it takes longer (15 ms) because of consensus validation. The audit logging times for blockchain-based and centralized TPA logging are shown in Fig. 3. Although a little slower, the blockchain-based method offers better security and integrity, which makes it a more dependable option for cloud-based data sharing.

C. Anomaly Detection Accuracy

The proposed machine learning-based anomaly detection, such as the separation forest algorithm, works by extracting anomalies from data through a tree-based algorithm that identifies patterns more efficiently than traditional rule-based methods unlike algorithm a it is based on the law, which is predetermined. Relying on threshold conditions, random forest can adapt to complex data distributions, increasing accuracy in detecting new or previously undetected anomalies.

A comparison of the accuracy of anomaly detection between rule-based detection and the suggested ML-based detection

method is shown in Table III. With an accuracy of 92%, the ML-based system outperforms rule-based techniques, which only attain 80% accuracy. This enhancement demonstrates how well machine learning works to dynamically identify security threats.

TABLE III. DETECTION ACCURACY

Detection Method	Detection Accuracy (%)
Rule-Based Detection	80
ML-Based Detection	92

Fig. 4 compares the accuracy of rule-based and ML-based approaches for anomaly detection. The ML-based approach's improved accuracy shows that it can adjust to changing security threats more successfully than static rule-based methods.

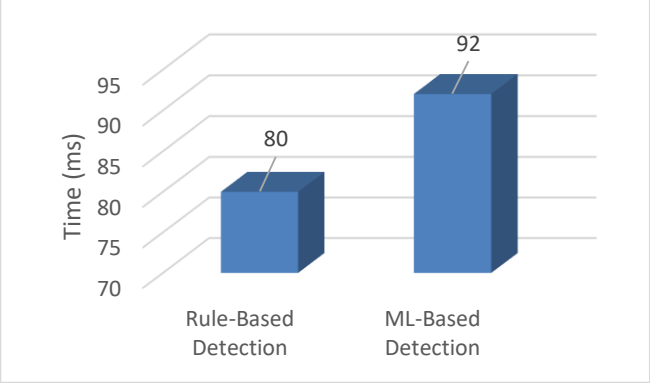


Fig. 4. Anomaly detection accuracy (%).

D. Access Control Flexibility and Security

This frame work -role-based access control (RBAC) offers flexibility in making access methods to change dynamically This also allows precise control With RBAC where group managers can update the user access rights without impacting the complete groups. This granular access techniques will strengthen the systems overall security. The adaptability of various access control models is assessed in Table IV. In contrast to conventional group-based access models, which receive a score of 3, the suggested Role-Based Access Control (RBAC) model receives the highest flexibility rating of 5. This suggests that RBAC enhances security and usability in cloud environments by enabling more dynamic and granular access permissions.

TABLE IV. ACCESS FLEXIBILITY

Access Control Model	Access Flexibility (1-5)
Group-Based	3
Proposed RBAC	5

The flexibility of various access control models is contrasted in Fig. 5. RBAC's greater flexibility rating indicates that it can efficiently handle changing user roles and permissions, guaranteeing security and convenience of access control.

E. Computational Overhead

Quantum computing can break the traditional cryptographic algorithms like AES or RSA, to protect from possible threats we need quantum-resistant cryptographic techniques like lattice-based encryption. This algorithm demands for greater processing requirements when compared with traditional encryption methods complex keys and complex mathematical computations required for encrypt and decrypt of quantum algorithms.

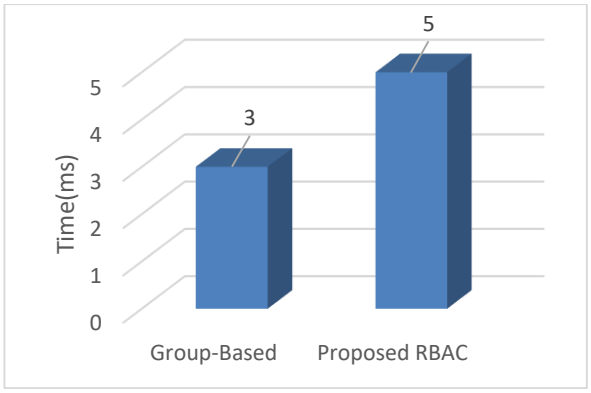


Fig. 5. Access control flexibility.

TABLE V. COMPUTATIONAL OVERHEAD

Cryptographic Scheme	Computational Overhead (%)
AES (Conventional)	5
Lattice-Based (Quantum-Resistant)	30

The computational overhead of the quantum-resistant lattice-based encryption technique and conventional AES encryption is contrasted in Table V. Although lattice-based encryption is more secure, it comes with a 30% overhead, while AES only has a 5% overhead. This demonstrates how using quantum-resistant cryptography involves a trade-off between increased security and computational efficiency.

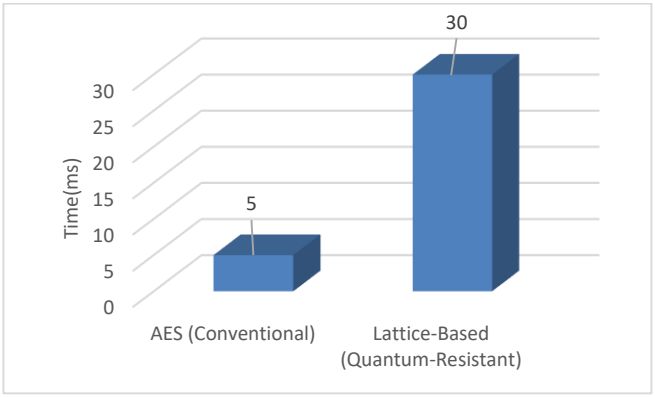


Fig. 6. Computational overhead (%).

The computational overhead of AES and lattice-based encryption is depicted in Fig. 6. Lattice-based encryption is a vital option for future-proofing cloud security systems because its resistance to quantum attacks justifies its higher overhead.

AES is the fast and efficient algorithm which reduces the time taken for secured data transfer. For unmatched data consistency and transparency, I require blockchain audit recording. In contrast with rule-based approaches, the ML model achieves the high true positives and high detection accuracy. For greater flexibility always preferable to use role-based application control (RBAC) which provides greater flexibility.

V. DISCUSSION

The proposed AES-based proxy re-encryption scheme is faster than RSA and ABE-IBE for encryption and decryption, making it suitable for environments where low latency is required. This increased level of security and transparency makes blockchain ideal for applications that require high levels of accountability and accountability. This approach also reduces reliance on rules that intensity down and provides much higher detection accuracy by better capturing subtle patterns in large data sets evaluates the outliers based on statistical features. As a result, it provides more reliable results, especially in dynamic and changing environments where models are not stable. Additionally, it permits accurate control. Group managers can modify user access rights with RBAC without affecting the entire group. The overall security of the system will be strengthened by these granular access techniques. Compared to conventional encryption techniques, this algorithm requires more processing power because it requires complicated keys and mathematical calculations to encrypt and decrypt quantum algorithms. The figures and graphs show quantitatively proves that proposed scheme is better than conventional methods in supporting for safe, secured and scalable cloud data exchange.

VI. CONCLUSION

To tackle the issues of real time threat detection, scalability and quantum threats too. The current work explored the architecture for safe and adaptable data exchange in cloud environment. For enhanced security this work also incorporates quantum-resistant cryptography, besides also suggests to use Role-Based Access Control (RBAC) for flexible, fine-grained access and incorporates machine learning-based anomaly detection for proactive threat detection and revocation. The work offers reliability, flexibility and enhanced access control through a scalable, adaptable framework. Further exploration of the current work will focus on reducing the overhead of quantum cryptography and improving blockchain logging.

VII. FUTURE WORK

While the proposed structure significantly enhances secure and scalable cloud data sharing, there are many areas for further discovery and improvements such as designing, lightweight lattice-based encryption and Post quantum cryptographic schemes such as Code-Based, Multivariate, and Hash-Based Cryptography could further enhance performance. Using Adaptive ML models that constantly learns from new attack patterns in real-time to reduce false positives and false negatives. Exploring layer -2 solutions like side chains, lightening network, shading can help optimize block chain efficiency.

REFERENCES

[1] S. Mhiri, A. Egio, M. Compastié, and P. Cosio, "Proxy Re-Encryption for Enhanced Data Security in Healthcare: A Practical Implementation," in *Proc.* 19th Int. Conf. on Availability, Reliability and Security (ARES 2024), Vienna, Austria, Jul. 30–Aug. 02, 2024, pp. 1–11, doi: 10.1145/3664476.3670874.

[2] R. Sharma and B. Joshi, "H-IBE: Hybrid-identity based encryption approach for cloud security with outsourced revocation," in Proc. 2016 Int. Conf. on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, India, 2016, pp. 1192–1196, doi: 10.1109/SCOPES.2016.7955629.

[3] S. Ling, K. Nguyen, H. Wang, and J. Zhang, "Server-Aided Revocable Predicate Encryption: Formalization and Lattice-Based Instantiation," The Computer Journal, vol. 62, no. 12, pp. 1849–1862, Dec. 2019, doi: 10.1093/comjnl/bxz079.

[4] R. Agrawal, M. Imran, and H. Khan, "Machine Learning-Based Malicious User Detection in Cloud Computing," J. Inf. Secur. Appl., vol. 47, pp. 233–239, 2019.

[5] S. Wang, N. Luo, B. Xing, et al., "Blockchain-based proxy re-encryption access control method for biological risk privacy protection of agricultural products," Sci. Rep., vol. 14, p. 20048, 2024, doi: 10.1038/s41598-024-70533-0.

[6] N. Bindel, M. Brenner, and M. Naveed, "Lattice-based proxy re-encryption in the quantum-safe era," IEEE Trans. Inf. Forensics Secur., vol. 16, pp. 2273–2285, 2021.

[7] S. Qi and Y. Zheng, "Crypt-DAC: Cryptographically Enforced Dynamic Access Control in the Cloud," IEEE Trans. Dependable Secur. Comput., vol. 18, no. 2, pp. 765–779, Mar.–Apr. 2021, doi: 10.1109/TDSC.2019.2908164.

[8] R. Raghav, N. Andola, K. Verma, S. Venkatesan, and S. Verma, "Proactive threshold-proxy re-encryption scheme for secure data sharing on cloud," J. Supercomput., vol. 79, no. 13, pp. 14117–14145, Sep. 2023, doi: 10.1007/s11227-023-05221-3.

[9] H. Yu, X. Lu, and Z. Pan, "An Authorized Public Auditing Scheme for Dynamic Big Data Storage in Cloud Computing," IEEE Access, vol. 8, pp. 151465–151473, 2020.

[10] V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-Based Encryption for Fine-Grained Access Control of Encrypted Data," in Proc. 13th ACM CCS, 2006.

[11] Y. Zhang and Y. Mao, "Blockchain-Based Public Auditing for Dynamic Data Sharing in Cloud Environments," J. Cloud Comput., vol. 8, no. 1, pp. 21–32, 2019.

[12] P. R. S. Reddy and K. Ravindranath, "CR-IBE Based Data Sharing and Revocation in the Cloud," J. Discrete Mathematical Sciences and Cryptography, 2024.

[13] X. Chen, Y. Zhang, and M. Li, "Privacy-Preserving Public Auditing for Secure Cloud Storage," IEEE Trans. Comput., vol. 64, no. 5, pp. 1223–1235, 2015.

[14] X. Yang, M. Wang, X. Wang, G. Chen, and C. Wang, "Multi-Replica and Multi-Cloud Data Public Audit Scheme Based on Blockchain," IEEE Access, vol. 8, pp. 144809–144822, 2020.

[15] B. Li, Q. He, F. Chen, H. Jin, Y. Xiang, and Y. Yang, "Auditing Cache Data Integrity in the Edge Computing Environment," IEEE Trans. Parallel Distrib. Syst., vol. 32, no. 5, pp. 1210–1223, 2021.

[16] C. Wang, K. Ren, S. Yu, and W. Lou, "Toward Publicly Auditable Secure Cloud Data Storage Services," IEEE Network, vol. 24, no. 4, pp. 19–24, 2010.

[17] J. Herranz, "Attribute-based encryption implies identity-based encryption," IET Inf. Secur., vol. 11, no. 6, 2017, doi: 10.1049/iet-ifs.2016.0490.

[18] D. Boneh and M. Franklin, "Identity-Based Encryption from the Weil Pairing," in Proc. CRYPTO 2001, Lecture Notes in Computer Science, vol. 2139, pp. 213–229, 2001.

[19] S. Yu, C. Wang, K. Ren, and W. Lou, "Achieving Secure, Scalable, and Fine-Grained Data Access Control in Cloud Computing," in Proc. IEEE INFOCOM, pp. 1–9, 2010.

[20] G. Ateniese, K. Fu, M. Green, and S. Hohenberger, "Improved proxy re-encryption schemes with applications to secure distributed storage," ACM Trans. Inf. Syst. Secur., vol. 9, no. 1, pp. 1–30, Feb. 2006, doi: 10.1145/1127345.1127346.

[21] O. Regev, "On Lattices, Learning with Errors, Random Linear Codes, and Cryptography," in Proc. 37th Annu. ACM Symp. Theory Comput. (STOC '05), 2005, pp. 84–93.

[22] Z. Zheng, S. Xie, H. Dai, X. Chen, and H. Wang, "An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends," IEEE Access, vol. 6, pp. 12399–12421, 2017.

[23] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in Proc. 8th IEEE Int. Conf. Data Min. (ICDM), pp. 413–422, 2008.

[24] V. C. Hu, D. R. Kuhn, and D. F. Ferraiolo, "Attribute-based access control," *Computer*, vol. 48, no. 2, pp. 85–88, Feb. 2015.

ALE Model: Air Cushion Impact Characteristics of Seaplane Landing Application

Yunsong Zhang, Ruiyou Li Shi*, Bo Gao, Changxun Song, Zhengzhou Zhang
State Grid Electric Power Space Technology Company Limited, Beijing, 102213, China

***Abstract*—Seaplane landing is a strong nonlinear gas-liquid-solid multiphase coupling problem, and the coupling impact characteristics of air cushion are very complicated, and it is difficult to maintain the stability of the air-frame. In this paper, The ALE method is used to study the landing of seaplane at different initial attitude angles and velocities. Firstly, a comparative study of the structure entry model and the air cushion effect model of flat impact water surface is conducted to verify the reliability of the numerical model in this paper, and the influence of the velocity, the water shape and the air cushion are accurately analyzed. Then, a seaplane landing is systematically studied, and the vertical acceleration, attitude angle, aircraft impact force and flow field distribution are analyzed. The results show that the air cushion has a great influence on the landing of seaplane. The smaller the initial horizontal velocity, the more obvious the cushioning effect of the air cushion. Cavitation causes a secondary impact on the tail and produces a pressure value exceeding the initial value, which may cause damage to the aircraft structure. The air cushion has a buffering effect on the seaplane, the pitch angle increases at a slower rate and the pressure value at the monitoring point decreases. The larger the initial attitude angle, the more significant the air cushion. By analyzing the landing rules of seaplane, the range of speed and attitude angle suitable for seaplane takeoff and landing process is given. The results of this paper can provide theoretical guidance for the stability design of seaplane takeoff and landing process.**

***Keywords*—Seaplane; ALE method; multiphase coupling; air cushion**

I. INTRODUCTION

Seaplanes can take off and land on water surface such as rivers, lakes, and seas. They possess notable characteristics, including excellent maneuverability, accessibility, high safety, and scalability. Seaplanes can operate on both water and land, reducing their dependence on specific geographical environments and the need for dedicated airport runways [1, 2]. However, the takeoff and landing process of seaplanes presents a challenging problem characterized by strong nonlinear gas-liquid-solid multiphase coupling. This process needs to consider anti-sinking ability, static stability, surface maneuverability, etc. Moreover, the presence of strong impacts further complicates this problem [3, 4].

Recent research on seaplanes has primarily focused on fluid-structure coupling algorithms, air cushion effects, and structural impact mechanisms. Regarding fluid-structure coupling algorithms, Iwanowski et al. [5] conducted numerical investigations on the horizontal rigid body impacting a water surface and analyzed the influence of compressible air cushions. The governing equations for air and water (modeled as compressible and non-compressible fluids, respectively) are solved using finite difference and fluid volume (VOF) methods. Mori Y et al. [6] combined the discrete element method (DEM) with computational fluid dynamics (CFD) to theoretically derive the stability condition of the drag term and develop a new implicit algorithm. The compatibility of the implicit algorithm with the boundary model composed of symbolic distance function and immersed boundary method is verified by experiments. Hessenthaler et al. [7] integrated several independent analytical solutions into a comprehensive framework, demonstrating its utility in analyzing convergent behavior and introducing novel fluid-structure interaction (FSI) algorithms. Han K et al. [8] validated the applicability of the coupled Boltzmann method (LBM) and discrete element method (DEM) in solving irregular particle transport in turbulence, employing test cases involving polygonal and super-quadratic particle transport in high Reynolds number fluid flows. Oger et al. [9] applied smooth particle hydrodynamics (SPH) method to simulate the solid-liquid coupling problem in a free-surface flow environment, and proposed a new formula of spatial variation resolution with variable smoothing length. Panciroli R et al. [10] investigated the fluid elasticity of an elastic wedge using a coupled finite element method and smooth particle hydrodynamics (FEM-SPH) method. The numerical simulation results agreed well with experimental data, accurately predicting the influence of hydro-elasticity on water impact involving the elastic wedge. Ahmadzadeh M et al. [11] employed the coupled Euler-Lagrange (CEL) method to study the impact of a sphere in free fall motion. Facci et al. [12] utilized the volume of fluid (VOF) method, based on the finite volume method, to simulate free surface multi-phase flows. They numerically analyzed the water impact phenomenon on a moving body and obtained the multi-phase flow field and surface pressure distribution on the body. Servan-Camas et al. [13] applied the SPH-FEM coupling model to analyze the liquid sloshing problem in the navigation body. Aquelet N et al. [14] proposed the Euler-Lagrange coupling algorithm, using a penalty function to calculate the coupling force at the fluid-structure interface and predict the local pressure peak on the structure. Fourey et al. [15] compared two fluid-structure coupling algorithms, parallel interleaving and sequential interleaving [16]. They found that the latter exhibited higher accuracy and stability but lower computational efficiency.

During the takeoff and landing process, the fuselage structure of a seaplane experiences significant friction or collision with the water body, while the presence of an air cushion adds complexity to the dynamics of the structure. Chuang [17] found that objects with small dead corners are more likely to form an air cushion when entering water. The air

*Corresponding Author.

cushion between the model and the water surface plays an important role in the impact process. Song et al. [18] compared different shapes of air cushion, analyzed the structure of air cushion with different volume, and analyzed the influence of different volume of air cushion on the peak value of water entry impact force of buoy. Ermanyuk and Ohkusu [19] designed a flat-bottom disk slamming experiment to study the influence of air cushion on the impact pressure. The presence of air cushion on the surface strongly affects the impact time scale and the shape of the splash jet. Chen Zhen et al. [20] used MSC. Dytran to simulate the mixing of air layer and water surface, and made a detailed analysis on the formation of air cushion. They used the ALE method to observe the presence of a hollow air cushion during the water entry process of flat-bottomed structures, noting that the air cushion can be identified within the air cushion when the peak impact pressure occurs. They also employed a neural network method to fit the impact pressure results and predict the bottom impact pressure. Huera-Huare et al. [21] investigated the effects of different angles on models and their corresponding air cushion effects. They observed that when the water entry angle decreased to less than 5°, an air cushion formed with a significantly lower peak impact pressure than the theoretical value. When the exit angle exceeded 5°, the peak impact pressure followed von Kan's theory. Zhang Jian et al.[22] used numerical simulation to study the influence mechanism of hollow air cushion during the water entry of two-dimensional wedge. Fang et al. [23] studied the air cushion effect and impact load in two-dimensional flat plate water entry problem using the multi-phase Riemann-SPH method based on the PVRS Riemann-solver.

The structural impact problem involves the complex coupling of rigid bodies or deformable bodies with the movement of the surrounding flow field. This complexity is further heightened by the dynamic changes in the free surface and the interaction with air. Adam et al. [24] proposed a new surface tension formula that can deal with multi-phase problems with high density and viscosity ratios. Wang et al. [25] (2022) developed a strategy to eliminate gas phase tensile instability, ensuring computational stability. They also established Riemann models for different materials and obtained a robust gas-solid-liquid contact algorithm. Washino K et al. [26] (2020) proposed an interface capture method based on color function to improve the smoothness of the interface. Shi et al. [27] investigated the effects of head shape parameters, shell thickness, water entry velocity, and angle on the acceleration, pressure, stress, and structural deformation of an elastic underwater vehicle during water entry impact. Based on the fluid volume multi-phase flow model and dynamic grid technology, Liu et al. [28] established a coupling calculation method for the multi-phase flow field and trajectory of a cross-medium vehicle entering water at high speed.

Grid-free method is also often used to simulate the impact of structures [29]. Shao et al. [30] simulated the high-speed impact jet problem, and analyzed the pressure response rule on the wall. Shao et al. [31] established an improved Smoothed Particle Hydrodynamics (SPH) model, and analyzed the influencing factors and flow field changes during the water entry of slender objects. Yang et al. [32] proposed an SPH-EBG algorithm to simulate the impact of dam break flow on elastic plates. Khayyer et al. [33] introduced a full Lagrange particle method under the Material Point Simulation (MPS) framework to simulate the influence of structural elastic response on water entry [34]. Sun et al. [35] improved the SPH method and conducted a simulation study on cylinder water entry problems [36]. Chen et al. [37] utilized MPS method to simulate the water entry problem of a two-dimensional wedge. They investigated the effects of different particle arrangements on calculated results, including vertical hydrodynamic force and free surface changes.

The take-off and landing of seaplane is a strong nonlinear gas-liquid-solid multiphase coupling problem, and it is necessary to consider the air cushion effect and impact effect during the movement. Theoretical analysis can establish clear relationships between changes in physical quantities and flow parameters, offering broad applicability. However, solving the nonlinear problem of air cushion impact is challenging. Experimental analysis often encounters scale effects between experiment models and actual motion due to the intricate topological shapes and external environment. Additionally, model testing requires a lengthy period, and in many cases, the experimental data is incomplete with inadequate repeatability. The grid method can get relatively accurate results when simulating small deformations, however, the VOF method or the level set method needs to be used to track the interface. The particle-type method has made significant progress in simulating severe fluid-structure coupling problems. However, there is a scarcity of full-fluid-structure coupling algorithms that consider the boundary layer effect on the structure's surface. In this paper, ALE method is used to study the landing of a seaplane. The coupled impact dynamic characteristics at different angles and entry speeds are analyzed, and the changes of air cushion are analyzed. The results of this paper have a good guiding significance for the development of seaplane.

II. METHODOLOGY

A. Fundamental Equations

In this paper, the ALE algorithm is used to describe the fluid domain. And the fluid-structure coupling algorithm based on penalty function and Lagrange method are used to simulate the landing process. The Euler coordinate system serves as a fixed fluid coordinate system, unaffected by the object's movement or deformation. The Lagrangian coordinate system acts as a fixed solid coordinate system, with its grid nodes attached to the material nodes. As the solid undergoes deformation, the solid coordinate system adjusts accordingly. The ALE coordinate system is independent of the Eulerian coordinate system and the Lagrangian coordinate system and is not completely fixed on space or solid nodes.

The ALE description introduces a reference domain independent of the material domain and the spatial domain, which always coincides with the grid throughout the calculation process. Fig. 1 shows the mapping relationship between the structural domain and the fluid domain. The mapping expression from the material domain to the spatial domain is as follows:

$$x = x(X, t) \tag{1}$$

The mapping from ALE reference domain to the spatial domain is:

$$x = w(w,t) \tag{2}$$

Since the reference domain always coincides with the grid, the transformation is performed:

$$w = w-1(x,t) \tag{3}$$

The mapping relationship of the material domain and the reference domain:

$$w = w-1[x(X,t),t] = v(X,t) \tag{4}$$

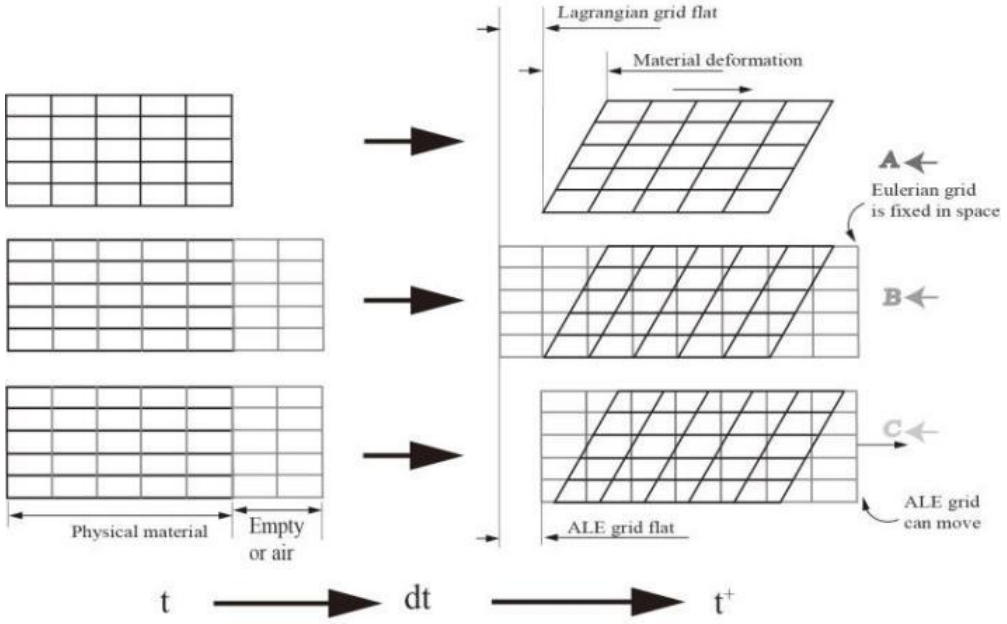


Fig. 1. Mapping between lagrange, euler, and ALE domains.

The expression of the mass equation is:

$$\frac{\partial \rho(\chi,\ t)}{\partial t} = -\rho \frac{\partial u_i}{\partial x_i} - c_i \frac{\partial \rho}{\partial x_i} \tag{5}$$

Where χ is the Lagrange coordinate system, x is the Euler coordinate system, ρ is the fluid density, and c_i is the relative velocity between the structure particle and the grid point in the reference coordinate system.

The equation of motion is:

$$\frac{\partial u_i(\chi,\ t)}{\partial t} = \frac{\partial \sigma_{ij}}{\partial x_j} + \rho b_i - c_j \frac{\partial u_i}{\partial x_j} \tag{6}$$

Where b_i is a unit force, in the Newtonian fluid, the stress tensor is related to the speed, the relationship is related:

$$\sigma_{ij} = \delta_{ij} P + \mu\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \tag{7}$$

Wherein μ is a power viscosity coefficient, P represents pressure.

B. Contact Algorithm

In the contact algorithm, the contact force is proportional to the permeation vector in the time step. In the explicit finite element method, the contact algorithm calculates the interface force due to the influence of the structure on the fluid. These forces act on the fluid and structural nodes to prevent them from crossing the contact interface. The fluid utilizes either an ALE mesh or a Lagrangian grid [38]. In the contact algorithm, as shown in Fig. 2, one surface is designated as the contacting surface, while the second surface is the primary surface. The nodes located on these surfaces are respectively called slave nodes and primary nodes. For fluid-structural coupling problems, the fluid nodes on the interface are considered slave nodes, whereas the structural elements are treated as primary nodes. In this paper, a penalty-based contact method is employed. The force is applied from the slave node, and the force transmitted to the primary element node is scaled using a shape function. The corresponding expression is as follows:

$$F_s = -k_i \cdot d \tag{8}$$

$$F_m^i = N_i \cdot k \cdot d \tag{9}$$

The N_i is the shape function of the surrounding node *i*, (in 2D problems, the values are 1, 2; in 3D problems, the values are 1...4.). The coefficient k represents the stiffness of the spring, and d is a penetrating vector. If the node is completely

coincident from one of the primary nodes (see Fig. 2), the coefficient $k=1$. Applying the Euler-Lagrangian coupling method to calculate the coupling force on the nodes can effectively prevent the large deformation of the grid.

$$F_s = -k_i \cdot d \tag{10}$$

$$F_m^1 = k \cdot d \ \ \alpha\nu\delta \ \ F_m^2 = F_m^3 = F_m^4 = 0 \tag{11}$$

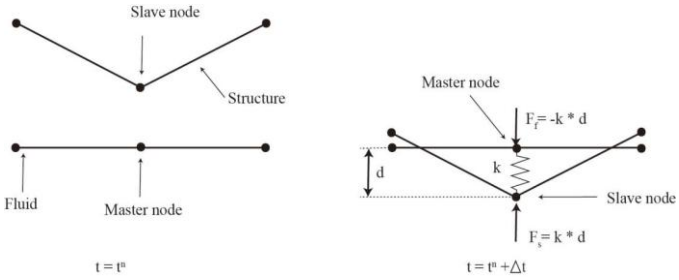


Fig. 2. Schematic diagram of the contact algorithm.

III. Numerical Verification

A. Water Entry

In order to verify the accuracy of the numerical model, this article first simulates the problem of a cylinder entering water. The aluminum solid cylindrical model has a length of 197 mm and a diameter of 50 mm, with a weight of 1.06 kg. It is subjected to an in-water angle of 60° and a speed of 4.35 m/s. Fig. 3 and Fig. 4 present the change in speed and acceleration of the cylinder, and Fig. 5 shows the simulation and experimental water entry process. It is observed that the numerical results before 0.12 s exhibit a strong agreement with the experimental results by Hou [39]. However, after 0.12 s, as the cavity begins to close, the pulsating pressure starts affecting the structure, leading to a decline in speed. Additionally, the simulated acceleration is slightly lower than the experiment.

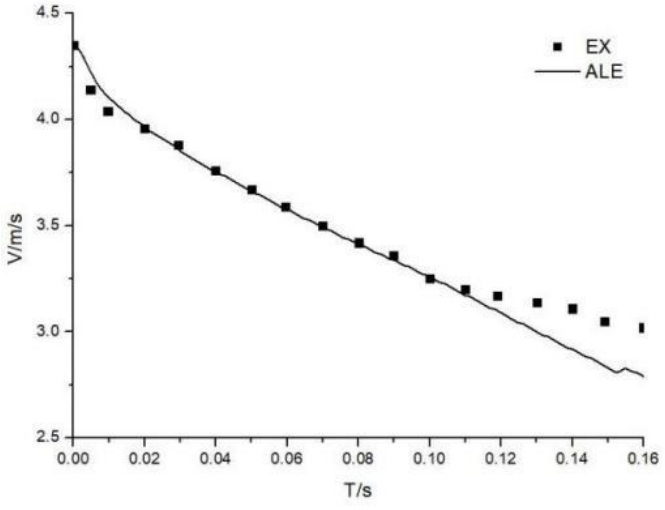


Fig. 3. Speed comparison between simulation and experiment [39]

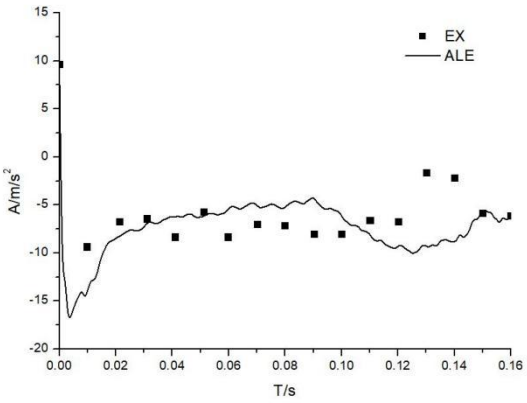


Fig. 4. Acceleration comparison between simulation and experiment [39].

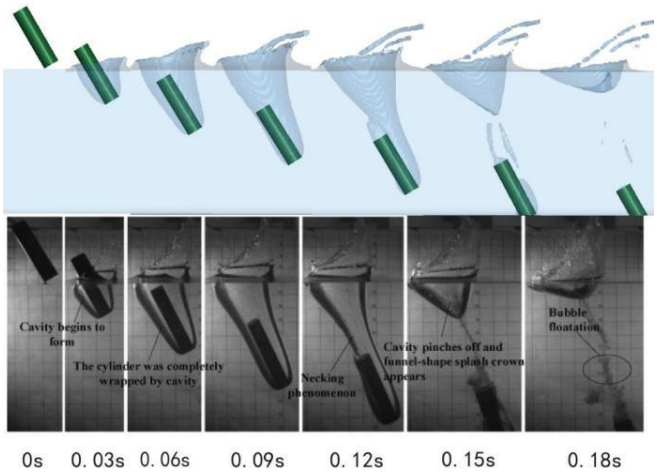


Fig. 5. The water entry of simulation and experiment [39].

B. Verification of Air Cushions

To validate the accuracy of the numerical model for the cushion effect, the same flat plate model as Ma et al. [40] was utilized. The impact plate employed has a mass of 32 kg, a length of 0.25 m, a width of 0.25 m, and a thickness of 0.012 m. By adjusting the initial position of the plate, an impact speed of 5.5 m/s was achieved.

Fig. 6 depicts the pressure curve at the center of the flat plate, and the simulation results exhibit favorable agreement with the experimental findings. The peak pressure slightly surpasses the experimental value, potentially attributed to the utilization of a bubble-generating device during the experiment, causing numerical fluctuations in the results. The pressure at the plate center only occurs momentarily within the structure, then air cushion is formed between the plate and the water surface, resulting in a pressure value of zero. Fig. 7 illustrates the impact force curve in the vertical direction, and the incorporation of the low-speed structure in this study does not exhibit a prominent air cushion effect. Additionally, Fig. 8 demonstrates the formation of an air cushion on the flat impact surface. As the plate continues to impact, the air cushion gradually increases, exerting minimal influence on the plate's impact.

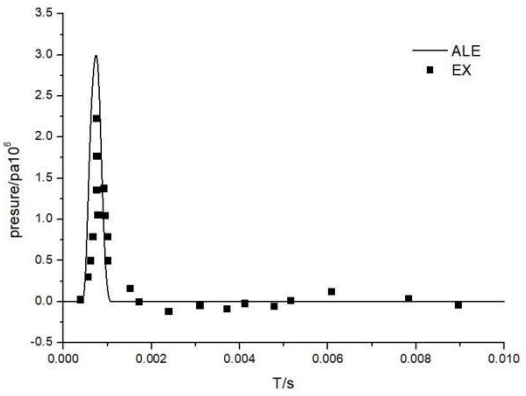


Fig. 6. Pressure distribution on plate center.

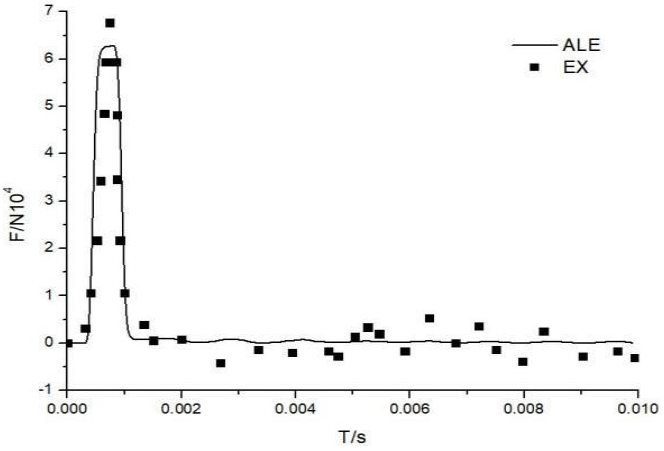


Fig. 7. Vertical impact force distribution of flat plate.

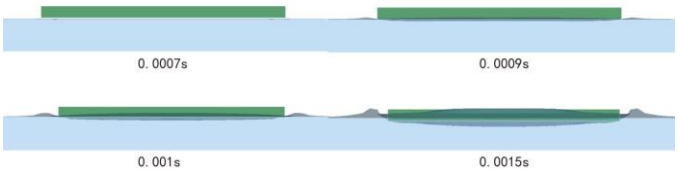


Fig. 8. Air cushion effect during impact.

IV. NUMERICAL MODELS

A. Finite Element Model

Fig. 9 is a finite element model of a seaplane, which is simplified on the basis of the actual model. The aerodynamic effect is ignored, and it is assumed that the seaplane is completely rigid. The weight of the aircraft is 53500kg, the overall length is 37m, and the wing length is 38.4m. The size of water is 200m×50m×20m, the size of air is 200m×50m×200m. The centroid is situated at 30% of the average aerodynamic cord length, precisely aligned with the center of the wing roots. About two million meshes were used. In Fig. 9, the monitoring point P1 is positioned at the center of the aircraft fuselage, located 14.5 m from the aircraft's head. The boundary point P2 is situated 20 m from the seaplane's head, while P3 corresponds to the center position of the seaplane's bottom. The vertical distance between P2 and P3 is 0.6 m, with a horizontal separation of 14.2 m.

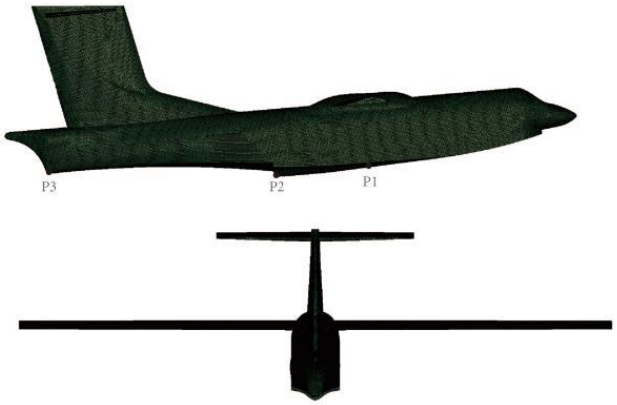


Fig. 9. Aircraft model and grid.

B. Material Parameters

The Mie–Gruneisen equation of state is as follows:

$$P=\frac{\rho_0 C^2\mu[1+(1-\frac{\gamma_0}{2})\mu-\frac{\alpha}{2}\mu^2]}{[1-(S_1-1)-S_2\frac{\mu^2}{\mu+1}-S_3\frac{\mu^3}{(\mu+1)^2}]}+(\gamma_0+\alpha\mu)E \tag{12}$$

Where E is in unit volume, C is the intercept of the u_s-u_p curve, S_1, S_2 and S_3 are the unit-less coefficients of the slope of the u_s-u_p curve. γ_0 is a Gruneisen parameter, α is a first-order correction value of γ_0. The compression ratio is related to the volume, defined as:

$$\mu=\frac{1}{V}-1 \tag{13}$$

An equation can also be approximately:

$$P=\rho_0 C^2\mu+(\gamma_0+\alpha\mu)E \tag{14}$$

The polynomial state equation is selected for the air domain, and the internal energy of the initial volume changes linearly.

$$P=C_0+C_1\mu+C_2\mu^2+C_3\mu^3+(C_4+C_5\mu+C_6\mu^2)E \tag{15}$$

where, C1, C2, C3, C4, C5, C6 are the confident constant respectively.

The thickness of the rigid shell is 0.05m, and the density is 1257kg/m^3. Table I shows the material parameters of the water and air domains.

TABLE I. MATERIAL PARAMETERS

parameter	Density(kg/m^3)	C	S_1	S_2	S_3	γ_0	V_0	E_0
water	1000	1480	1.92	-0.096	0	0.35	0	0
parameter	density(kg/m^3)	C1	C2	C3	C4	C5	C6	E0
air	1.22	0	0	0	0.4	0.4	0	0

C. Landing setting

When the seaplane lands, it should maintain a certain speed and attitude angle. Fig. 10 is a schematic diagram of the initial water entry of the seaplane. Table II shows the settings of speed and attitude angle under different landing conditions.

TABLE II. SEAPLANE LANDING CONDITIONS

Serial number	1	2	3	4	5	6
horizontal velocity (m/s)	35	45	63.9	45	45	45
Vertical velocity (m/s)	1.5	1.5	1.5	1.5	1.5	1.5
angle(°)	5	5	5	5	8	10

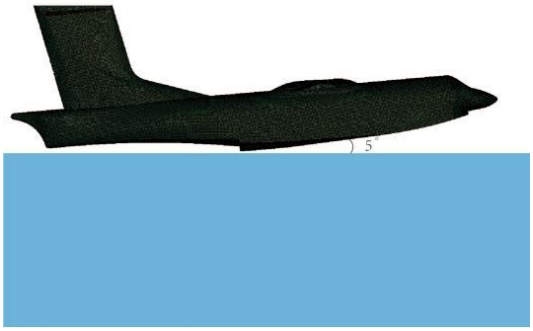


Fig. 10. Diagram of aircraft initial entry conditions.

V. RESULT AND ANALYSIS

A. Velocity Analysis

This paper systematically investigates the aircraft landing process, with an initial vertical velocity of 1.5 m/s. The variations in attitude angles and accelerations are analyzed for different initial horizontal velocities. The changes in the air cushion during the landing process are studied, along with the analysis of variations in impact force and pressure values at the monitoring point.

Fig. 11 illustrates the changes in vertical acceleration for an initial attitude angle of 5°. With an initial horizontal velocity of 35 m/s, the maximum vertical acceleration observed during the landing process is 12.17 m/s². For an initial horizontal velocity of 45 m/s, the maximum vertical acceleration reaches 13.14 m/s². When the initial horizontal velocity is further increased to 63.9 m/s, the maximum vertical acceleration during aircraft landing is 13.42 m/s². Notably, it is observed that the timing and magnitude of extreme acceleration remain similar across different horizontal velocities. Subsequent to the initial impact, the vertical acceleration exhibits relatively small fluctuations under higher horizontal velocity.

Fig. 12 shows the attitude angle change of the seaplane during landing when the initial attitude angle is 5°. For different initial speeds, the attitude angle of the seaplane initially decreases, then increases, and eventually decreases again. During the water impact process, the seaplane exhibits a forward inclination followed by a subsequent upward attitude. With an initial horizontal velocity of 35 m/s, the maximum attitude angle observed is 7.6°. When the initial horizontal velocity is 45 m/s, the maximum attitude angle reaches 7.64°. Finally, for an initial horizontal velocity of 63.9 m/s, the maximum attitude angle during water landing is 7.5°. By comparing the three curves in Fig. 12, it can be observed that for the same initial attitude angle, the seaplane requires a longer time to tilt forward at higher speeds. Therefore, during the landing process, it is essential to choose an appropriate speed. Different speeds lead to varying pitch angles, influencing the position and timing of the seaplane's contact with the water surface, as well as its vertical acceleration.

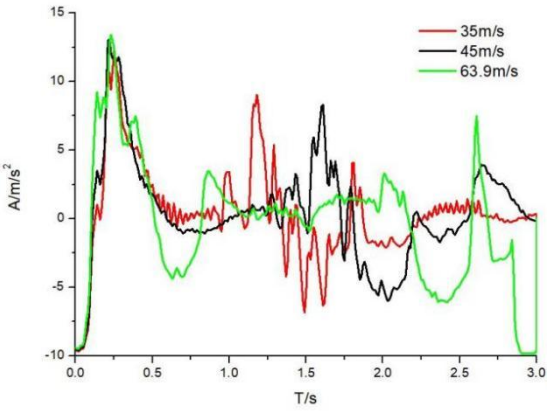


Fig. 11. Vertical acceleration curve.

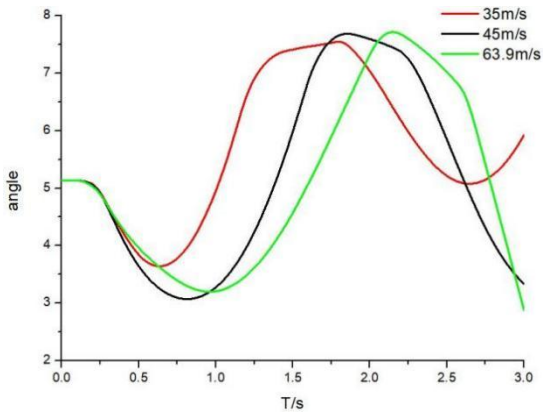


Fig. 12. Attitude angle change of the aircraft.

Fig. 13 shows the landing process of a seaplane. The initial attitude angle of the plane is 5°, the initial horizontal speed is 45m/s, and the initial vertical speed is 1.5m/s. At 0.5s, the tail of the seaplane drew a deep trench on the water surface, and at 1s, the water wave generated by the impact had obviously spread outward. The actual vertical displacement experienced by the seaplane during landing generally ranges from 1 to 2 meters. The maximum vertical displacement is 1.85 meters and the minimum displacement is 0.8 meters, which accords with the actual situation.

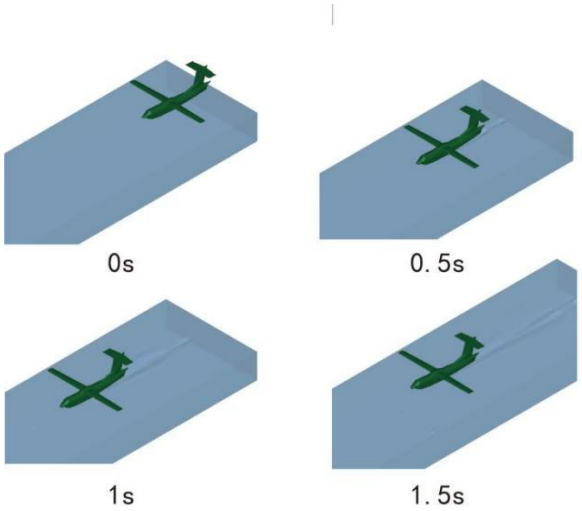


Fig. 13. Landing process of a seaplane.

The air cushion is defined as a hermetically sealed layer of air positioned between the free surface and the underlying structure. During the horizontal propagation of waves, the leading edge of the horizontal plate makes contact with the free surface, while the trailing end remains partially submerged or in contact with the flat plate, resulting in air displacement. This process induces deformation in the water's free surface, which subsequently undergoes reformation parallel to the horizontal plate. Even in tranquil water environments, the forward section of the seaplane's underside remains in contact with the water surface. When this part is not completely detached from the water body, the rear section of the aircraft's underside contacts the water surface, creating a substantial enclosed cavity and forming an air cushion.

Fig. 14 presents the vertical impact force curve for an initial attitude angle of 5°. At an initial horizontal velocity of 35 m/s, the maximum impact force is 1251kN. With an initial horizontal velocity of 45 m/s, the maximum impact force reaches 1289kN. Similarly, at an initial horizontal velocity of 63.9 m/s, the maximum impact force is 1275kN. Comparing the three curves in Fig.14, it can be observed that, when the initial vertical speed is same, the maximum impact force remains similar and is largely unaffected by the horizontal speed. However, it is evident from the figure that, with higher initial horizontal velocities, the vertical impact force of the aircraft exhibits relatively small fluctuations after the initial strong impact with the water surface.

Fig. 15 shows the pressure curve at monitoring point P1 with an initial attitude angle of 5°. At an initial horizontal velocity of 35 m/s, the peak pressure at P1 is 23.68kPa. When the initial horizontal velocity is 45 m/s, the peak pressure at P1 reaches 110.3kPa. Interestingly, at an initial horizontal velocity of 63.9 m/s, the pressure at P1 drops to zero. By comparing the three curves in Fig. 15, it can be inferred that the P1 position is located close to the nose of the aircraft, and as the initial horizontal velocity increases, it becomes more challenging for the P1 position to make contact with the water surface within the time interval of 0-2.75s.

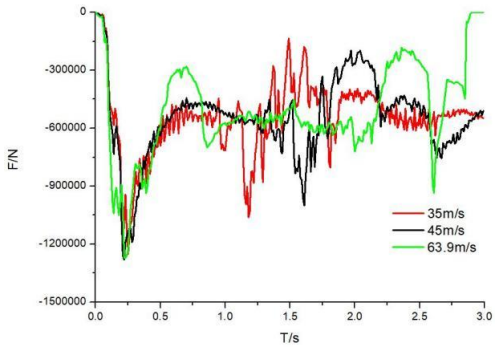


Fig. 14. Vertical impact force variation curve.

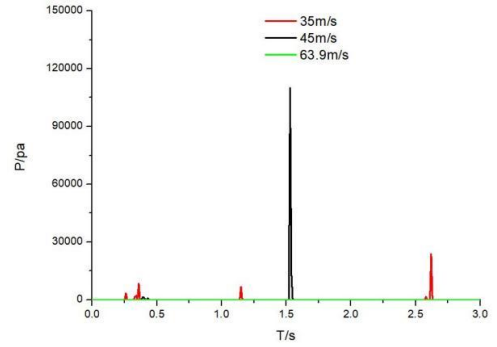


Fig. 15. Pressure curve of P1.

Fig. 16 presents the pressure curve at monitoring point P2 with an initial attitude angle of 5°. At an initial horizontal velocity of 35 m/s, the peak pressure recorded at P2 is 2618kPa. With an initial horizontal velocity of 45 m/s, the peak pressure at P2 reaches 2992kPa. Moreover, at an initial horizontal velocity of 63.9 m/s, the P2 monitoring point exhibits a peak pressure of 3489kPa. It is evident that higher initial horizontal velocities lead to greater peak pressures at P2.

Fig. 17 shows the pressure curve at monitoring point P3 with an initial attitude angle of 5°. At an initial horizontal velocity of 35 m/s, the peak pressure observed at P3 is 4124kPa. When the initial horizontal velocity is 45 m/s, the peak pressure at P3 decreases to 3535kPa. However, with an initial horizontal velocity of 63.9 m/s, the P3 monitoring point exhibits a significantly higher peak pressure of 6767kPa. Comparing the three curves in Fig.17, it becomes apparent that lower initial horizontal velocities result in higher pressure values generated by the second impact with the water surface at P3.

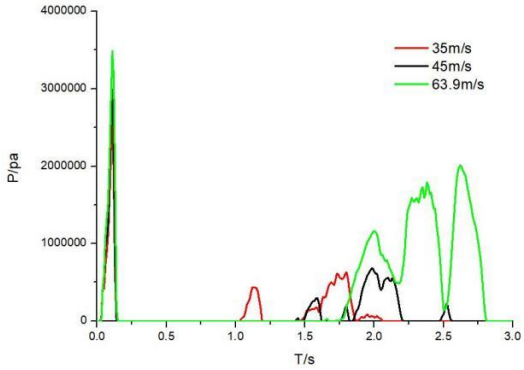


Fig. 16. Pressure curves of P2.

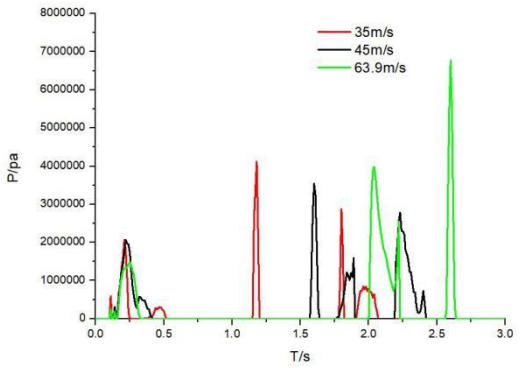


Fig. 17. Pressure curves of P3.

The pressure curves of monitoring points P1, P2, and P3 reveal important insights into the seaplane landing process. During landing, the seaplane initially makes contact with the water surface at monitoring point P2, resulting in a higher pressure value compared to monitoring points P1 and P3. For instance, at an initial horizontal velocity of 35 m/s, contact with the water surface occurs at P2 at 0.11s, leading to a significant pressure surge. At 0.2s, P3 makes contact with the water surface. Affected by the decrease of the impact surface velocity at P2, P3 produces a small pressure value. After 0.25s, P1 experiences a small pressure value influenced by the air cushion formed upon contact with the water surface. At 1.18s, when P3 undergoes secondary contact with the water surface, the pressure increases, surpassing the initial pressure, and an air cushion begins to form, as shown in Fig.18. Subsequently, due to the influence of the air cushion, the elevation angle increase rate decreases, and the elevation angle increases to the maximum value at 1.83 s. At 1.83s, the pressure values at P2 and P3 dropped due to the cushioning of the air cushion, and the tail comes out of the water at 2.6 seconds.

At an initial horizontal velocity of 45 m/s, the interaction between point P2 and the water surface transpires at 0.11s, leading to a notable surge in pressure. Subsequently, at 0.23s, point P3 makes contact with the water surface. Due to the diminished impact surface velocity at P2, P3 registers a diminished pressure value. Following 1.5s, P1 experiences a modest pressure level influenced by the air cushion. By 1.6s, when P3 undergoes secondary contact with the water surface, the pressure exceeds the initial value, resulting in the formation of an air cushion, as shown in Fig. 19. Consequently, influenced by the air cushion, the rate of elevation angle increase diminishes, and the pressure at P3 decreases after 1.6s.

In scenarios with an initial speed of 63.9 m/s, P2 contacts the water surface at 0.1s, generating a substantial pressure. At 0.24s, P3 also makes contact with the water surface, producing a reduced pressure value due to the decreased impact surface velocity at P2. At 2.0s, during P3's secondary contact with the water surface, the pressure surpasses the initial moment's value, and no air cushion is formed when connected with the air, as shown in Fig. 20. Subsequently, the pressure values at monitoring points P2 and P3 remain relatively high, displaying no significant drop or buffering effect. Throughout this period, P1 remains unaffected with a pressure value of zero. The horizontal speed exhibits minimal influence on the vertical impact force. For this specific seaplane structure, lower speeds result in higher pressure values generated by the tail during secondary impact on the water surface, whereas higher initial horizontal velocities yield greater pressure values during the initial contact with the water surface.

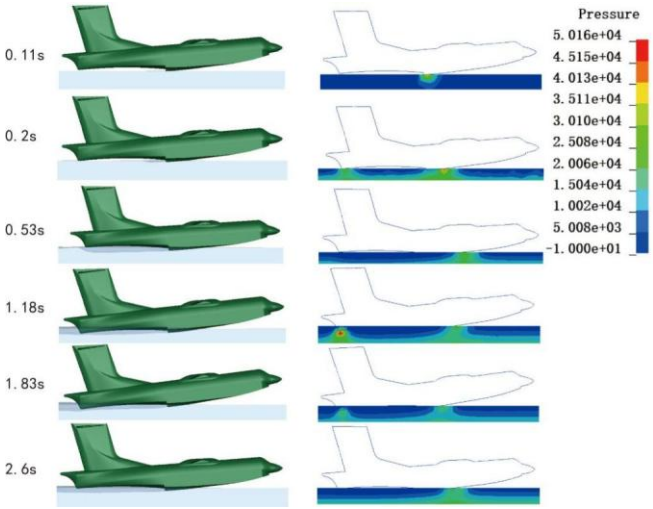


Fig. 18. Field and pressure distribution under 35m/s.

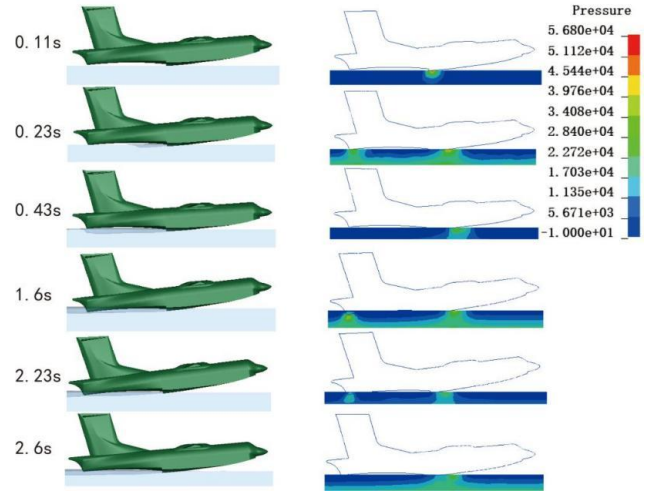


Fig. 19. Field and pressure distribution under 45m/s.

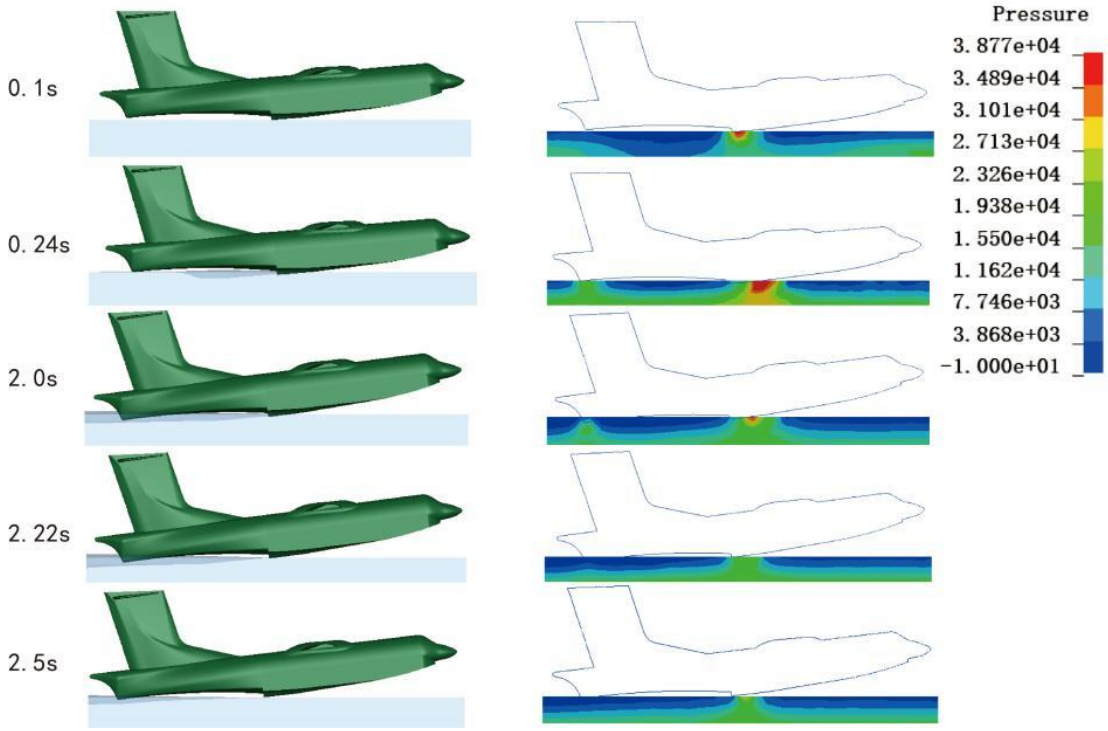


Fig. 20. Field and pressure distribution under 63.9m/s.

B. *Attitude Angles Analysis*

This paper presents a systematic investigation of the seaplane landing process, with a focus on analyzing the changes in attitude angle and acceleration. The study explores the behavior of the air cushion formed during the landing process and examines the variations in impact forces at different stages, as well as pressure at various positions.

Fig. 21 illustrates the vertical acceleration changes of the seaplane at an initial speed of 45m/s. For an initial attitude angle of 5°, the maximum vertical acceleration is measured at 13.2m/s². With an initial attitude angle of 8°, the maximum vertical acceleration reaches 18.7m/s². Moreover, an initial attitude angle of 10° results in a maximum vertical acceleration of 29.7m/s². By comparing the three curves in Fig. 21, it is evident that the vertical acceleration increases as the initial attitude angle becomes larger. Following the initial impact, a smaller attitude angle leads to a narrower range of vertical acceleration for the seaplane, exhibiting a similar trend.

Fig. 22 depicts the changes in attitude angle for an initial level speed of 45m/s, highlighting variations under different attitude angles. The seaplane initially experiences a decrease in attitude angle, followed by an increase, and subsequently a decrease. During the impact, the seaplane tilted forward at a certain angle and then its attitude angle increased. At an initial attitude angle of 5°, the minimum and maximum attitude angles are measured at 3.07° and 7.64°, respectively. For an initial attitude angle of 8°, the minimum and maximum attitude angles are 1.59° and 8.65°, respectively. Similarly, an initial attitude angle of 10° results in a minimum attitude angle of 1.93° and a maximum attitude angle of 8.38°. Notably, no significant rollover occurs within the first three seconds for all three working conditions.

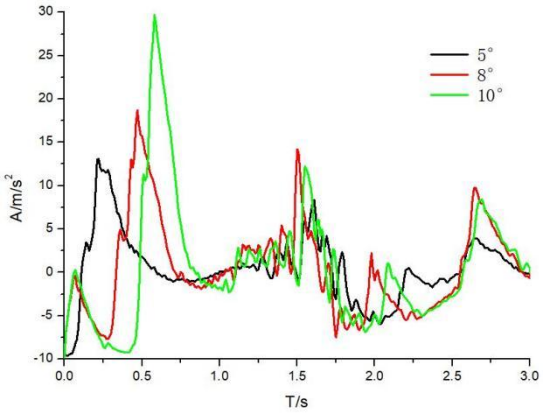


Fig. 21. Vertical acceleration comparison.

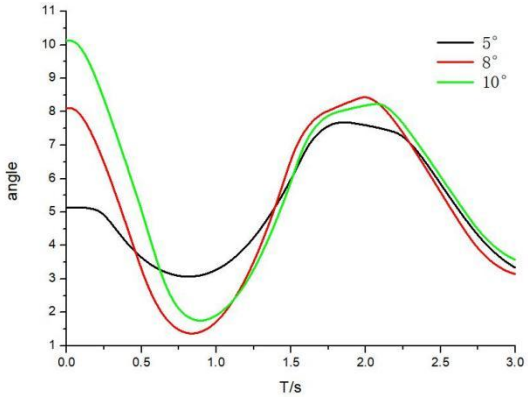


Fig. 22. Posture angle comparison

Fig. 23 shows the pressure curve of the seaplane with the initial horizontal velocity 45m/s. when the initial attitude angle is 5°, the maximum impact force is 1280kN. When the initial attitude angle is 8°, the maximum impact force is 1560kN. When the initial attitude angle is 10°, the maximum impact force is 2296kN. The larger the initial attitude angle, the greater the vertical impact force.

Fig. 24 shows the pressure curve at monitoring point P1 with an initial horizontal velocity 45m/s. At an initial attitude angle of 5°, the peak pressure recorded at P1 is 110.3kPa. With an initial attitude angle of 8°, the peak pressure at P1 reaches 70.27kPa. Furthermore, at an initial attitude angle of 10°, the peak pressure at P1 is measured at 360.2kPa. It is worth noting that a smaller initial attitude angle corresponds to a lower pressure value generated at the P1 position before 1 second. The proximity of the P1 monitoring point to the nose of the seaplane, combined with the front-heavy weight distribution, influences the pressure generated at the contact between P1 and the water surface.

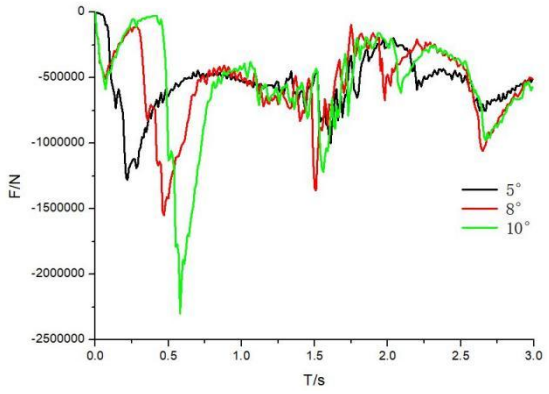


Fig. 23. Comparison of impact force.

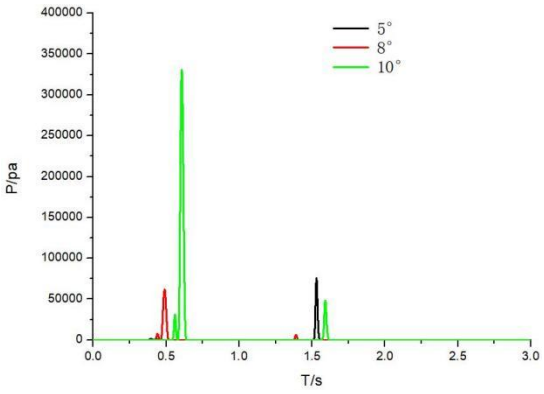


Fig. 24. Pressure comparison of P1.

Fig. 25 illustrates the pressure change curve observed at monitoring point P2 with the initial horizontal velocity 45m/s. For an initial attitude angle of 5°, the peak pressure recorded at P2 is 2992kPa. With an initial attitude angle of 8°, the peak pressure reaches 3121kPa. Similarly, at an initial attitude angle of 10°, the peak pressure at P2 reaches 3544kPa. It is observed that a larger initial attitude angle leads to a higher pressure at P2, with a longer time taken to reach the peak pressure. Fig. 26 shows the pressure curve observed at monitoring point P3 with an initial horizontal velocity of 45m/s. At an initial attitude angle of 5°, the peak pressure recorded at P3 is 3535kPa. With an initial attitude angle of 8°, the peak pressure at P3 reaches 3631kPa. Furthermore, at an initial attitude angle of 10°, the peak pressure is 4406kPa. It is evident that an increase in attitude angle results in a higher pressure generated during the secondary impact of the P3 contact with the water surface.

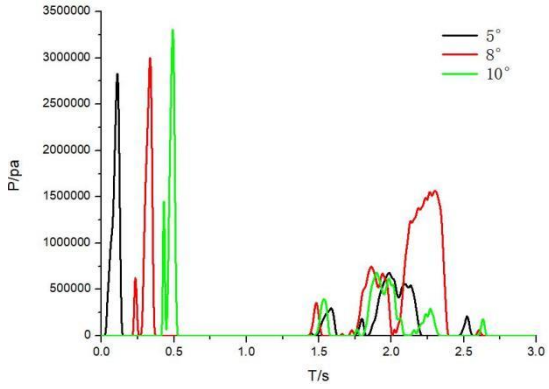


Fig. 25. Pressure comparison of P2.

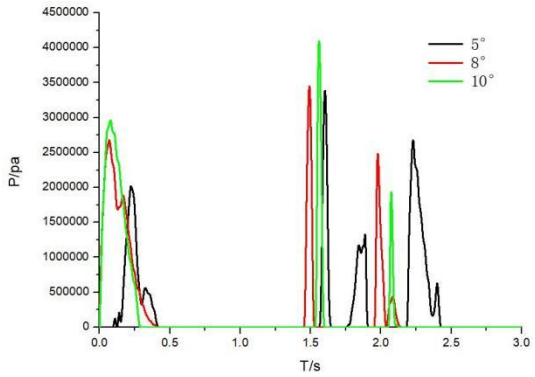


Fig. 26. Pressure comparison of P3.

The examination of pressure curves at monitoring points P1, P2, and P3 reveals that a greater initial attitude angle results in the seaplane's contact position with the water surface being closer to P3. Specifically, with an initial attitude angle of 5°, contact with the water surface occurs at P2 at 0.11s, resulting in a notable pressure surge. Subsequently, at 0.2s, P3 makes contact with the water surface, producing a reduced pressure value due to the decreased impact surface velocity at P2. Post 0.25s, P1 experiences a modest pressure level influenced by the air cushion formed upon contact with the water surface. By 1.18s, during P3's secondary contact with the water surface, the pressure value exceeds the initial moment's pressure, initiating the formation of an air cushion (see Fig. 18). Consequently, the rate of elevation angle increase diminishes due to the influence of the air cushion, reaching its maximum at 1.83s. At this point, the pressure values at P2 and P3 decrease at 1.83s due to the buffering effect of the air cushion, and the tail emerges from the water surface at 2.6s.

For an initial attitude angle of 8°, P2 contacts the water surface at 0.49s, followed by P3's secondary contact at 1.49s, resulting in an elevated pressure value and the formation of an air cushion (Fig. 27). Influenced by the cushioning air cushion,

the attitude angle increase rate diminishes. At 1.98s, the attitude angle starts decreasing, the air cushion connects with the air, leading to its disappearance, and the pressure values at P2 and P3 are high. The tail disengages from the water surface at 2.6s. With an initial attitude angle of 10°, P2 contacts the water surface at 0.51s, and P1 registers a significant pressure due to its instantaneous air cushion. At 1.56s, the pressure rises as P3 undergoes secondary contact with the water surface, initiating the formation of an air cushion (Fig. 28). At 2.07s, the tilt angle begins decreasing, the air cushion connects with the air, resulting in its disappearance, the pressure value at P3 decreases, and the monitoring point P2, positioned at the contact edge, experiences a floating pressure value. The tail disengages from the water surface at 2.6s.

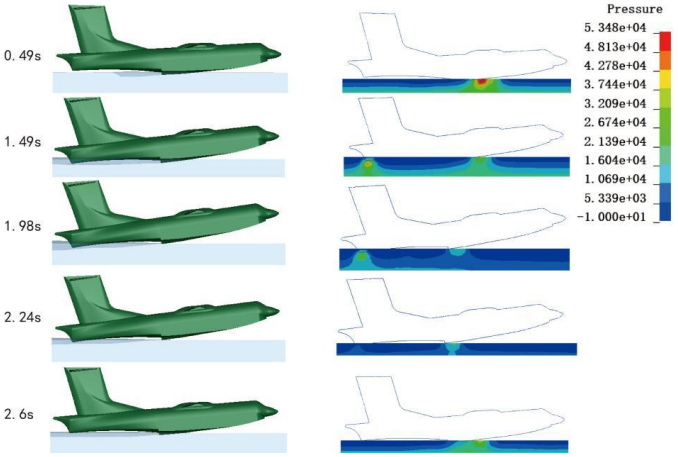


Fig. 27. Flow field and pressure distribution with attitude angle 8°.

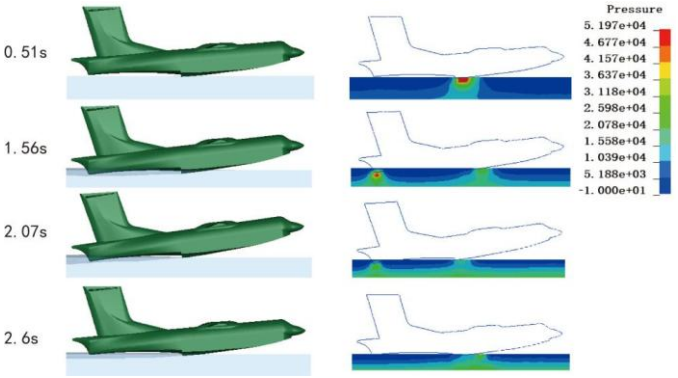


Fig. 28. Flow field and pressure distribution with attitude angle 10°.

VI. Conclusion

This paper presents a numerical model for studying air-cushion-coupled impact during seaplane landing in a hydrostatic environment. The accuracy of the model was verified by comparison with the cylinder water entry and flat plate impact experiments. Subsequently, the characteristics of air-cushion-coupled impact are investigated under various conditions, including different attitude angles, landing speeds, and impact loads. The conclusions are as follows:

- During the seaplane landing process, the attitude angle shows an obvious peak value within 0 to 3 seconds. Under different initial attitude angles, the maximum vertical acceleration of the seaplane increases as the initial attitude angle increases. Furthermore, the higher the initial attitude angle, the greater its change rate. The second impact of the aircraft produced greater pressure than the first impact.
- A reduced initial horizontal velocity of the seaplane leads to a diminished increase rate of pitch angle, accentuated cushioning effects of the air cushion, and a lower peak pressure at the monitoring point. When the initial attitude angle is large, the cushioning effect is more obvious, which reduces the peak pressure of the monitoring point.
- To mitigate seaplane structure damage and optimize air cushion utilization, maintaining an optimal landing speed between 30 m/s and 40 m/s is recommended. Furthermore, to prevent rollover and mitigate adverse air cushion effects, the suggested range for the attitude angle is 6° to 8°.

Acknowledgment

This work is supported by State Grid Electric Power Space Technology Company Limited technology projects. Grant No. 52950024000B

References

[1] Morabito M G. A Review of Hydrodynamic Design Methods for Seaplanes [J]. Journal of Ship Production and Design, 2021, 37;159-180, https://doi.org/10.5957/JSPD.11180039.

[2] Zhang X, Huang J, Huang Y, Huang K, Yang L, Han Y, Wang L, Liu H, Luo J, Li J. Intelligent Amphibious Ground-Aerial Vehicles: State of the Art Technology for Future Transportation[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(1); 970-987., https://doi.org/10.1109/TIV.2022.3193418.

[3] Washio S. Recent Developments in Cavitation Mechanisms: A Guide for Scientists and Engineers[J]. Recent Developments in Cavitation Mechanisms, 2014.

[4] Tan R, Samuel R T, Cao Y. Nonlinear Dynamic Process Monitoring: The Case Study of a Multiphase Flow Facility[J]. Computer Aided Chemical Engineering, 2017, 40: 1495-1500, https://doi.org/10.1016/B978-0-444-63965-3.50251-8.

[5] Iwanowski B, Fujikubo M,Yao T. Analysis of Horizontal Water Impact of a Rigid Body with the Air Cushion Effect[J]. Journal of the Society of Naval Architects of Japan,1993, 1993(173), 293-302, https://doi.org/10.2534/jjasnaoe1968.1993.293.

[6] Mori Y, Sakai M.Development of a robust Eulerian–Lagrangian model for the simulation of an industrial solid–fluid system[J]. Chemical Engineering Journal,2021, 406: 126841, https://doi.org/10.1016/j.cej.2020.126841.

[7] Hessenthaler A, Balmus M, Röhrle O, Nordsletten D. A class of analytic solutions for verification and convergence analysis of linear and nonlinear fluid-structure interaction algorithms[J]. Computer Methods in Applied Mechanics and Engineering, 2020, 362: 112841, https://doi.org/10.1016/j.cma.2020.112841.

[8] Han K, Feng Y T, Owen D R. Numerical simulations of irregular particle transport in turbulent flows using coupled LBM-DEM[J]. Computer Modeling in Engineering & Sciences, 2007, 18(2), 87-100, https://doi.org/10.3970/cmes.2007.018.087.

[9] Oger G, Doring M, Alessandrini B, Ferrant P.Two-dimensional SPH simulations of wedge water entries[J]. Journal of Computational Physics.2006, 213(2): 803-822, https://doi.org/10.1016/j.jcp.2005.09.004.

[10] Panciroli R, Abrate S, Minak G, Zucchelli A.A. Hydroelasticity in water-entry problems: Comparison between experimental and SPH results[J]. Composite Structures, 2012, 94(2), 532-539, https://doi.org/10.1016/j.compstruct.2011.08.016.

[11] Ahmadzadeh M, Saranjam B, Fard A H, Binesh R A. Numerical simulation of sphere water entry problem using Eulerian–Lagrangian method[J]. Applied Mathematical Modelling, 2014, 38(5-6), 1673-1684, https://doi.org/10.1016/j.apm.2013.09.005.

[12] Facci A L, Porfiri M, Ubertini S.Three-dimensional water entry of a solid body: A computational study[J]. Journal of Fluids & Structures, 2016, 66, 36-53, https://doi.org/10.1016/j.jfluidstructs.2016.07.015.

[13] Serván-Camas B, Cercós-Pita J L, Colom-Cobb J, García-Espinosa J.,Souto-Iglesias A. A. Time domain simulation of coupled sloshing–seakeeping problems by SPH–FEM coupling[J]. Ocean Engineering, 2016,123; 383-396, https://doi.org/10.1016/j.oceaneng.2016.07.003.

[14] Aquelet N, Souli M, Olovsson L. Euler–Lagrange coupling with damping effects: Application to slamming problems[J].Computer Methods in Applied Mechanics & Engineering, 2006,195(1-3); 110-132, https://doi.org/10.1016/j.cma.2005.01.010.

[15] Fourey G, Hermange C, Touzé D L,Oger G.An efficient FSI coupling strategy between Smoothed Particle Hydrodynamics and Finite Element methods[J]. Computer Physics Communications, 2017, 217: 66-81, https://doi.org/10.1016/j.cpc.2017.04.005.

[16] Sun P-N, Le Touzé D, Oger G, Zhang A-M . An accurate FSI-SPH modeling of challenging fluid-structure interaction problems in two and three dimensions[J]. Ocean Engineering, 2021, 221: 108552, https://doi.org/10.1016/j.oceaneng.2020.108552.

[17] Chuang S L. Experiments on slamming of wedge-shaped bodies[J]. Journal of Ship Research, 1967,11(3);190-198, https://doi.org/10.5957/jsr.1967.11.3.190.

[18] Song R, Ren C, Ma X, Qiu L, Lin Z. The study on the anti-impact performance of the oscillating buoy with various air cushions[J]. IET Renewable Power Generation,2021, 15(14): 3459-3471, https://doi.org/10.1049/rpg2.12183.

[19] Ermanyuk E V, Ohkusu M. Impact of a disk on shallow water[J]. Journal of Fluids & Structures, 2005,20(3); 345-357, https://doi.org/10.1016/j.jfluidstructs.2004.10.002.

[20] Chen Z, Xiao X. The simulation analysis of the air cushion in the flat structure enters the water[J]. Journal of Shanghai Jiao Tong University, 2005,39(005);670-673, https://doi.org/10.3321/j.issn:1006-2467

[21] Huera-Huarte F J, Jeon D , Gharib M. Experimental investigation of water slamming loads on panels[J]. Ocean Engineering,2011,38(11-12), 1347-1355, https://doi.org/10.1016/j.oceaneng.2011.06.004.

[22] Zhang J, Wang K, Wan Z Q. Research on prediction method of impact load of two-dimensional wedge into water based on air cushion effect[J]. Ship Science and Technology,2016 ,6(2);1672-7649, https://doi.org/10.3404/j.issn.1672-7649.

[23] Fang X L, Ming F R, Wang P P, Meng F Z, Zhang M A . Application of multiphase Riemann-SPH in analysis of air-cushion effect and slamming load in water entry[J]. Ocean Engineering, 2022, 248: 110789, https://doi.org/10.1016/j.oceaneng.2022.110789.

[24] Adami S, Hu X Y , Adams N A. A new surface-tension formulation for multi-phase SPH using a reproducing divergence approximation[J].Journal of Computational Physics, 2010,229(13);5011-5021, https://doi.org/10.1016/j.jcp.2010.03.022.

[25] Wang L, Xu F,Yang Y. Research on water entry problems of gas-structure-liquid coupling based on SPH method[J]. Ocean Engineering, 2022, 257: 111623, https://doi.org/10.1016/j.oceaneng.2022.111623.

[26] Washino K, Chan E L, Kaji T , Matsuno Y, Tanaka T. On large scale CFD–DEM simulation for gas–liquid–solid three-phase flows[J]. Particuology,2021, 59: 2-15, https://doi.org/10.1016/j.partic.2020.05.006.

[27] Shi Y, Pan G, Yim S C, Yan G, Zhang D. Numerical investigation of hydroelastic water-entry impact dynamics of AUVs[J]. Journal of Fluids and Structures,2019, 91: 102760, https://doi.org/10.1016/j.jfluidstructs.2019.102760.

[28] Liu X, Luo K, Yuan X, Qi B X. Numerical study on the impact load characteristics of a trans-media vehicle during high-speed water entry and flat turning[J].Ocean Engineering,2023, 273: 113986, https://doi.org/10.1016/j.oceaneng.2023.113986.

[29] Peng Y X, Zhang A M,Ming F R. A thick shell model based on reproducing kernel particle method and its application in geometrically nonlinear analysis[J]. Computational Mechanics, 2017,62(3);309-321, https://doi.org/10.1007/s00466-017-1498-9.

[30] Shao J R. Li S M, Liu M B. Numerical Simulation of Violent Impinging Jet Flows with Improved SPH Method[J]. International Journal of Computational Methods, 2016, 13(04): 1641001, https://doi.org/10.1142/S0219876216410012.

[31] Shao J R, Yang Y, Gong H F, Liu B M. Numerical Simulation of Water Entry with Improved SPH Method[J].International Journal of Computational Methods,2019,16(02): 1846004, https://doi.org/10.1142/S0219876218460040.

[32] Yang X, Liu M, Peng S, Huang G C. Numerical modeling of dam-break flow impacting on flexible structures using an improved SPH–EBG method[J]. Coastal Engineering,2016, 108: 56-64, https://doi.org/10.1016/j.coastaleng.2015.11.007.

[33] Khayyer A,Gotoh H, Falahaty H, Shimizu Y.Towards development of enhanced fully-Lagrangian mesh-free computational methods for fluid-structure interaction[J]. Hydrodynamic Research and Progress Series B, 2018, 30: 49-61, https://doi.org/10.1007/s42241-018-0005-x.

[34] Xue B, Wang S-P, Peng Y-X, Zhang M A.A novel coupled Riemann SPH–RKPM model for the simulation of weakly compressible fluid–structure interaction problems[J].Ocean Engineering, 2022, 266: 112447, https://doi.org/10.1016/j.oceaneng.2022.112447.

[35] Sun P, Zhang A M, Marrone S, Ming F. An accurate and efficient SPH modeling of the water entry of circular cylinders[J]. Applied Ocean Research, 2018,72;60-75, https://doi.org/10.1016/j.apor.2018.01.004.

[36] Yang F, Gu X, Zhang Q, Zhang Q.A peridynamics-immersed boundary-lattice Boltzmann method for fluid-structure interaction analysis[J]. Ocean Engineering,2022, 264:112528, https://doi.org/10.1016/j.oceaneng.2022.112528.

[37] Chen X, Rao C Q, Wan D C. Numerical simulation of wedge entry problem by MPS method[J]. Journal of computational mechanics, 2017,34 (3);23-27, https://doi.org/10.7511 / jslx201703013.

[38] Souli M, Ouahsine A,Lewin L. ALE formulation for fluid-structure interaction problems[J]. Computer Methods in Applied Mechanics & Engineering, 2012,190(5-7); 659-675, https://doi.org/10.1016/S0045-7825(99)00432-6.

[39] Hou Z, Sun T, Quan X, Zhang Y G,Sun Z, Zong Z.Large eddy simulation and experimental investigation on the cavity dynamics and vortex evolution for oblique water entry of a cylinder[J]. Applied Ocean Research, 2018,81;76-92, https://doi.org/10.1016/j.apor.2018.10.008.

[40] Ma Z H, Causon D M, Qian L, Mingham G C, Mai T, Greaves D, Raby A. Pure and aerated water entry of a flat plate[J]. Physics of Fluids,2016,28(1);016104, https://doi.org/10.1063/1.4940043.

Self-Organizing Neural Networks Integrated with Artificial Fish Swarm Algorithm for Energy-Efficient Cloud Resource Management

Dr. A. Z. Khan[1], Dr.B.Manikyala Rao[2], Janjhyam Venkata Naga Ramesh[3], Elangovan Muniyandy[4], Eda Bhagyalakshmi[5], Prof. Ts. Dr. Yousef A.Baker El-Ebiary[6], Dr. David Neels Ponkumar Devadhas[7]

Assistant Professor, Applied Physics Department, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India[1]

Associate Professor, Dept of CSE, Aditya University, Surampalem, Andhra Pradesh, India[2]

Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India[3]

Adjunct Professor, Department of CSE, Graphic Era Hill University, Dehradun, 248002, India[3]

Adjunct Professor, Department of CSE, Graphic Era Deemed To Be University, Dehradun, 248002, Uttarakhand, India[3]

Department of Biosciences, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India[4]

Applied Science Research Center, Applied Science Private University, Amman, Jordan[4]

Assistant Professor, Dept.of Computer Science and Engineering, Koneru Lakhmaiah Education Foundation, Vaddeswaram, Guntur - 522302, Andhra Pradesh, India[5]

Faculty of Informatics and Computing, UniSZA University, Malaysia[6]

Professor, Department of Electronics and Communication Engineering, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India[7]

***Abstract*—Cloud computing's exponential expansion requires better resource management methods to solve the existing struggle between system performance and energy efficiency and functional scalability. Traditional resource management practices frequently lead systems in large-scale cloud environments to produce suboptimal results. This research presents a brand-new computational framework that unites Self-Organizing Neural Networks (SONN) with Artificial Fish Swarm Algorithm (AFSA) to enhance energy efficiency alongside optimized resource allocation and scheduling improvements. The SONN system groups workload information and automatically changes its structure to support fluctuating demand rates then the AFSA optimizes resource management through swarm-based intelligent protocols for high performance with scalable benefits. The SONN-AFSA model achieves substantial performance gains by analyzing real-world CPU usage statistics and memory usage behavior together with scheduling data from Google Cluster Data. The experimental findings show 20.83% lower energy utilization next to 98.8% prediction rates alongside 95% SLA maintenance and an outstanding 98% task execution rate. The proposed model delivers reliability outcomes superior to traditional approaches PSO and DRL and PSO-based neural networks which achieve accuracy rates above 88% and reach 92% accuracy. The adaptive platform delivers better power management to cloud computations yet preserves operational agility by adapting workload distributions. The learning ability of SONN joined with AFSA optimization segments produces superior resource direction capabilities which yield better service delivery quality. Research will proceed beyond its current scope to study real-time feedback structures as it evaluates multi-objective enhancement through large-scale dataset validation work to boost cloud computing sustainability across various platforms.**

***Keywords*—Energy-efficient cloud resource management; Self-Organizing Neural Networks (SONN); Artificial Fish Swarm Algorithm (AFSA); cloud optimization; swarm intelligence; resource utilization; task scheduling**

I. Introduction

Modern technological development benefits substantially from cloud computing's emergence as a revolutionary solution during the past decades [1]. The internet distribution model known as cloud computing enables remote service access for computing essentials which includes servers and storage systems and databases and networks and applications and various other resources [2]. Amazon Web Services (AWS) together with Microsoft Azure launched their transformative services during early 2000s that resulted in massive cloud adoption [3]. The evolution of cloud computing now delivers three fundamental models named Infrastructure as a Service (IaaS) together with Platform as a Service (PaaS) and Software as a Service (SaaS) [4]. The technology functions as the backbone of current applications by facilitating web hosting alongside big data analytics and machine learning and IoT systems. The fundamental benefit package of cloud computing technologies drives sectoral transformation since business activities and service delivery models have developed new foundations [5].

Cloud resource management stands as a key element of cloud computing because it handles the effective utilization of processing power along with memory and storage capacity and network bandwidth allocation. The demand for cloud services continues to grow significantly because millions of users push the complexity of managing abundant digital data resources to new heights [6]. Complex management systems are necessary for cloud environments due to their shifting requirements which necessitate specific resource distribution for optimal user

performance. The primary purpose of cloud resource management is to achieve optimal resource use by cutting down wasteful usage and promoting system performance optimization per [7]. The primary obstacle in this context lies in controlling resource distributions that exceed genuine operational needs. The delivery of excessive resources leads to waste through both needless power consumption and unutilized resources that drive up operational costs as well as create additional environmental consequences. Under-provisioning of resources causes service quality to decrease because it creates slower response times, increased latency and ultimately subpar user experiences [8]. Stretched computing systems must receive proper resource configurations to maintain affordable cloud resources structures alongside desired service levels. The management success of resources heavily relies on scalability elements. Cloud system resources must automatically adjust through scaling processes to match demand fluctuations for maintaining optimal operational performance.

Cloud providers need to recognize increased user activity peaks alongside the capability to decrease their resource utilization when usage reaches lower levels. Cloud systems utilize dynamic resource allocation strategies to manage workload variations efficiently thus maintaining both performance speed and minimal resource usage [9]. The main focus of cloud resource management requires energy efficiency since data centers present increasing operational costs coupled with rising environmental expenses. Cloud providers face their biggest operational cost in energy usage because their data centers use a sizeable segment of worldwide electrical demand. [10]. The energy demands create major environmental effects that remain substantial when operations use non-renewable power sources. Dynamic workload balancing and virtualization technologies and improved server utilization techniques enable cloud providers to reach these outcomes. By managing power usage through strategic resource allocation cloud providers simultaneously reduce operational costs while pursuing environmental sustainability goals. Cloud infrastructure resource management procedures lead to economic and environmental outcomes that influence billions of daily tasks across the network [11]. Systems with dynamic resource distribution adjust their design according to usage trends to decrease both physical infrastructure demands and corresponding energy usage and costs. Modern cloud service requirements need effective management solutions that maintain effective cost performance through a balance between resource usage and energy efficiency and scalability features [12]. The worldwide power consumption grows because cloud data centers experienced rapid growth with their exploding cloud computing services operations. The vast power usage needs created by the growing demand for computational resources causes crucial environmental problems with large data centers because of their energy demands and cooling requirements [13]. The economic burden to operate these facilities presents maximum challenge to data centers alongside mandatory requirements for energy-efficient resource administration. Thorough power utilization demands substantial pressure on cloud service providers to handle peak performance requirements when they must suppress their usage. The improper distribution of resources generates both economic and environmental challenges through poorly managed energy consumption resulting either from excessive resource spending or from underused resources [14]. Successful service quality preservation combined with adaptable resource management systems which respond to workload variations constitute essential requirements for reducing energy consumption [15]. The critical gap in the previous studies includes the failure in addressing the integration of the adaptive machine learning techniques and swarm intelligence for dynamic resource distribution. Incorporating of both machine learning adaptation with the intelligent swarm-based distribution faces lack a hybridized optimization method. The study develops a new computational framework by integrating Artificial Fish Swarm Algorithm (AFSA) and Self-Organizing Neural Networks (SONNs) to optimize cloud resource management while achieving power consumption minimization. The SONN neural model modifies its learning algorithms and structural arrangements to establish patterns in cloud workload requirements and the fish-based distribution capabilities demonstrated by AFSA underlie resource allocation. The combined adaptive features of Self-Organizing Neural Networks (SONNs) with Artificial Fish Swarm Algorithm (AFSA) optimization capabilities lead to efficient resource distribution while minimizing power usage while maintaining performance quality. The SONN-AFSA hybrid framework orchestrates cloud resources by maximizing energy consumption while achieving efficient task scheduling for cloud environments. Google Cluster Data served as the testing ground for the model through its use of authentic cloud data center workload logs that enabled both the application of the proposed framework and tests of model performance against modern cloud infrastructure. Through the integration of SONN and AFSA models the framework offers dynamic performance adjustments for different cloud platform operations which enable both precise execution and resource optimization regardless of workload variations. The model achieves high resource utilization rates combined with optimized scheduling techniques enabling SLA compliance thus delivering improved quality cloud service outcomes. The proposed approach achieved verification by using the Google Cluster Data which contains actual cloud data center workload traces for validating its practical application to present-day cloud infrastructure. The integration between SONN and AFSA runtime achieves dynamic flexibility which enables scalable efficiency across different cloud platforms with enhanced prediction precision and resource management particularly in situations with varying workload demands. Remains high while service level agreements (SLA) are met thanks to optimized task scheduling along with the model's precise prediction accuracy which leads to improved quality of cloud service delivery. Task scheduling in cloud environments.

- Dynamic adaptability enabled by integrating SONN and AFSA results in predictable scalability while also minimizing resource misallocations in various cloud settings even when fluctuating workloads occur.
- The integration of SONN and AFSA allows for dynamic adaptability, making the model scalable and efficient across different cloud environments, with improved prediction accuracy and resource allocation even under varying workloads.

- The proposed model reduces total energy consumption, demonstrating a significant improvement over traditional and hybrid optimization methods, ensuring sustainability in cloud resource management.
- The model ensures high resource utilization efficiency and meets service level agreements (SLA), improving the quality of cloud service delivery through optimized task scheduling and prediction accuracy.

The rest of the sections of this research have been organized as follows: Review of the existing literature Self-Organizing Neural Networks Integrated with Artificial Fish Swarm Algorithm for Energy-Efficient Cloud Resource Management in Section II. In Section III, proposed research Methodology is explained. The presents the experimental results in Section IV. In Section V, Conclusion and further work is mentioned and the study is concluded.

II. Literature Review

A. Hybrid Machine Learning Approach for Resource Allocation

The paper proposes a hybrid machine learning (RATS-HM) approach for combined resource allocation security and efficient task scheduling in cloud computing to address these challenges according to Bal et al. [16]. The proposed RATS-HM techniques are given as follows: The ICSTS system which incorporates an improved cat swarm optimization algorithm for task scheduling tasks produces reduced make-span times while achieving maximum system throughput. A group optimization-based deep neural network (GO-DNN) serves as a framework for efficient resource allocation through bandwidth and resource load design constraints. NSUPREME functions as a lightweight authentication scheme which provides encryption services for data storage security. The proposed RATSHM technique undergoes simulation with a new setup to demonstrate its superiority against current state-of-the-art methods. Research findings demonstrate that the proposed method outperforms existing approaches by demonstrating better resource utilization alongside lower energy consumption and faster response times. The proposed model demonstrates longer utilization times which require additional improvement.

B. Heuristic Algorithm for Cloud-Based Energy Consumption

Sunil et al.,[17] introduces two energy efficient Virtual machine placement algorithms related to bin packing heuristics focusing the efficiency of the physical machine's energy, Energy Efficient VM Placement (EEVMP) and Modified Energy Efficient VM Placement (MEEVMP), which reduces the total energy usage in the data-center. The reduction in the energy consumption by 53% established using the EEVMP algorithm when compared with the default VM placement algorithm Power-Aware Best-Fit Decreasing algorithm (PABFD) of CloudSim, Average SLA violation of 3.5% and number of VM migrations by 64.47% when compare to PABFD, the MEEVMP algorithm achieves the reduction in energy consumption by 54.24%, average SLA violation by 4.39% and number of VM migrations by 67.713 %.

C. Hybrid Resource Allocation Solution

Shahidinejad et al. [18] proposed a combined solution that manages cloud resource allocation for workloads. The k-means clustering and ICA method served as the resource allocation framework. This research used the decision tree method to determine an efficient resource allocation solution. The researchers ran the cloud workloads through real-world tests to measure the effectiveness of their hybrid solution. The hybrid method demonstrates enhanced capabilities for cloud optimization tasks. The model achieved its performance assessment on a minimal workload while the decision tree technique displayed unstable results. The proposed hybrid solution struggles to handle unpredictable workload fluctuations in real time and depends on static QoS criteria which may restrict its ability to adapt to changing user needs.

D. Secure Sensor Cloud Architecture (SASC)

Nezhad et al., [19] proposes a method that contains three phases, including the first phase as a star structure is constructed in which a specific key that is encrypted is shared between the each child and the parent to secure the communications between them. In second phase, the members of the cluster send their data to the cluster head and also the data is encrypted at the end of the each connection. The third phase included to improve the security of the inter cluster communications with the help of authentication protocol. By this, the cluster heads are authenticated before transmitting the information. The proposed method is also implemented using the NS2 software. The improvement in the energy consumption, end-to-end delay, flexibility and packet delivery rate results in the proposed method compared to other previous methods.

E. Adaptive Heuristic Approach for Energy Efficiency

Yadav et al. [20] developed an adaptive heuristic approach to reduce energy consumption while enhancing system performance. The researchers tested their developed method within the CloudSim and PlanetLab cloud simulation platforms. The new method shows improved performance when measured through energy efficiency along with SLA results. Real-time workload spikes might not be properly managed by the proposed algorithms because they require accurate CPU utilization predictions that prove difficult in fast-changing environments.

F. Optimization Techniques for Load Balancing

The research team of Goyal et al. [21] examined the energy efficiency and load balancing capabilities of the cloud environment through different optimization techniques including the whale optimization algorithm (WOA), cuckoo search algorithm (CSA), BAT, cat swarm optimization (CSO), and particle swarm optimization (PSO). Among the optimization methods WOA demonstrates the best performance efficiency. The integration of AFSA with SONNs remains an underexplored area to resolve convergence and robustness issues. AFSA's dynamical weight and topology optimization capabilities present substantial possibilities for SONN enhancement with faster convergence speed and improved accuracy and greater adaptability to complicated datasets. The gap between swarm intelligence and neural network training holds great promise for innovative research that would combine these two approaches.

G. Research Gap

The previous studies show the key significant contributions in the allocation of resources, task scheduling, and conservation of energy in cloud computing, which includes several limitations. Many of the existing methods were insufficient in adaptability to dynamic workload changes and challenges with flexibility in large scale cloud environments, resulting in the failure of integrating the robust security mechanisms. Further, the static QoS criteria and limited real-time decision-making often restricts the effect of those techniques.

To overcome this limitation, this proposed RATS-HM approach focuses on these gaps by implementing the machine learning with swarm intelligence, which provides more effective and efficient resource allocation model. By implementing the enhanced optimization algorithms, dynamic scheduling ideas and the security protocols, the RATS-HM provides a more comprehensive approach that overcomes the existing problems like resource utilization, energy efficiency and response times.

III. Research Methodology

The research develops an energy-effective cloud resource management method by utilizing Google Cluster Data to supply detailed metrics about CPU and memory usage with task identification numbers and scheduling details. The model integrates two key techniques: Self-Organizing Neural Networks (SONN) and the Artificial Fish Swarm Algorithm (AFSA). AFSA enables the system to allocate tasks which optimize energy conservation without breaking Service-Level Agreement parameters. The research combines these methods together to enhance cloud performance by improving energy efficiency alongside resource utilization and task execution efficiency.

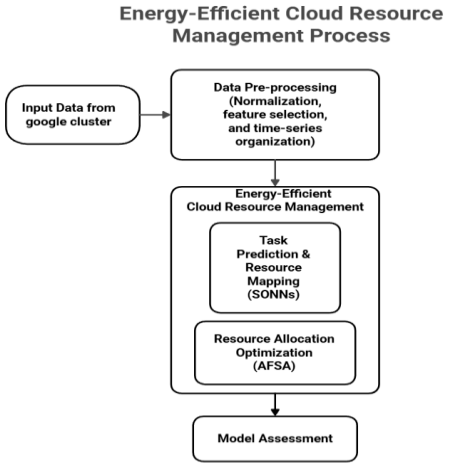


Fig. 1. AFSA-SONNs.

The Fig. 1 represents the work flow Artificial Fish Swarm Algorithm (AFSA).

A. Data Collection

The analysis of workloads and the development of resource management models aim to improve cloud energy efficiency through study of an open-source Google Cluster Data dataset. The Google Cluster Data constitutes an open data repository that presents detailed measurements from Google's production data centers which extend across billions of records during 29 days of analysis [22]. The informative dataset Completion Time includes CPU performance analytics run alongside memory statistics with records of job and task identifiers and their scheduling at various priority threshold levels. The study employs resource prediction capabilities to develop optimal resource management strategies by examining workload patterns through analytic assessment of its attributes. The transparency and practicality of the open-source dataset along with its ability to support collaborative development emerge through billions of production data measurements from Google's data centers which were gathered for 29 days. The dataset includes comprehensive metrics including CPU utilization percentages and memory use alongside task/job identifiers and scheduling events and priority settings. The data characteristics support both workload pattern exploration and resource utilization forecasting and resource allocation optimization. The research utilizes this open-source dataset to guarantee transparency and real-world applicability of swarm intelligence-based neural model development focused on energy-efficient cloud environments while offering reproducibility.

B. Data Pre-Processing

1) *Normalization:* The process of normalizing pixel values across images through blended analysis defines image pre-processing normalization. The mathematical representation of normalization appears in Eq. (1):

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

2) Feature selection: Research approaches for energy-efficient cloud resource management lead to distinctive mathematical formulations in feature selection. Since you're focusing on methods like mutual information and PCA, here are the basic equations for each:

a) Mutual information: Mutual Information serves as a measurement tool to determine the degree at which one feature reveals details about another. The dependency relationship between each target variable feature and the designated outcome (climate emissions) enables selecting proper features through this approach.

For two variables *X and Y*, the mutual information *I (X; Y)*is expressed in Eq. (2):

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \quad (2)$$

b) Principal Component Analysis (PCA): Through PCA we transform our features into principal components which maintain the key information in a set of orthogonal variables.

Through PCA we transform our features into principal components which maintain the key information in a set of orthogonal variables as stated in Eq. (3):

$$\sum = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - x_j)(x_i - x_j)^T \quad (3)$$

C. Artificial Fish Swarm Algorithm (AFSA)

The Artificial Fish Swarm Algorithm (AFSA) uses three fundamental fish behaviors which integrate Foraging with

Swarming to improve cloud environment resource management and following for optimization. Fig. 2 suggest work flow Artificial Fish Swarm Algorithm (AFSA).

1) Foraging: The foraging phase finds optimized resource arrangements through individual fish exploration which achieves both energy optimization and proper allocation of resources for their assigned duties. Fish complete evaluations of their current positions together with alternative options using a fitness function that assesses both energy efficiency and resource distribution and service level agreement compliance. Each fish scans their surrounding positions to find better solutions before reaching points of best arrangement. System exploration procedures show how resource distributions can optimize their distribution patterns while supporting both energy efficiency and tasks.

2) Swarming: Fish distribution uses clustering techniques to organize tasks according to workload patterns that assess resource requirements and task significance. When using assigned grouping all resources can receive tasks that their performance profile matches thus maximizing efficiency. The fish collaboration system provides optimal resource allocation which maximizes system operational performance while minimizing unnecessary resource consumption. Swarm behavior enables the system to allocate resources effectively without causing either excessive loading of individual nodes or resource underutilization.

3) Following: During this phase fish utilize neighborhood detection to locate highly effective neighbors for which they follow. When fish use optimal solutions, they find their performance accelerates the algorithm's convergence toward better configurations. Fish groups enhance system effectiveness as they track top-performing connections among peers to prevent time-wasting refinements of substandard outcomes.

D. Self-Organizing Neural Networks (SONN)

A key component of this research uses Self-Organizing Neural Networks (SONNs) to conduct dynamic workload pattern examinations and resource forecasting that enables power-efficient resource control. Fig. 2 suggest the neural network of Self-Organizing Neural Networks (SONN). Their specific contributions include:

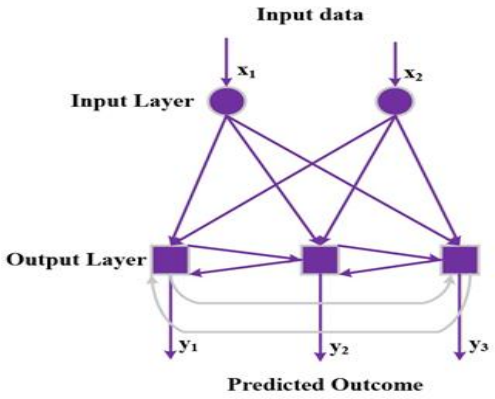


Fig. 2. Self-Organizing Neural Networks (SONN).

1) Workload pattern recognition: SONNs extract workload insights from cloud monitoring metrics which include CPU performance statistics and memory patterns alongside task execution information to form detailed workload patterns. Resource forecasting and understanding utilization become possible through identified patterns.

2) Dynamic adaptation: Traditional neural networks lack SONN's capability to evolve its network topology through adjustments like vegetable or removable of computing elements when confronted by changing data conditions. The adaptable network structure permits processing of sudden workload fluctuations found in cloud systems which demonstrate unpredictable behaviors.

3) Clustering and categorization: Workload categories form through network clustering to assist resource requirement identification. Overy/-based classification provides essential information about workload types that leads to better selection of optimal resource distribution approaches.

4) Feedback-Driven learning: The implementation of feedback systems allows SONNs to enhance their prediction capabilities through ongoing data refinement with repeated improves their accuracy over time. The predicted resource allocation serves as feedback which prompts the network to change its learning process so the system does not repeat inefficient provisioning.

5) Energy-Efficient decision support: The Artificial Fish Swarm Algorithm (AFSA) uses information from SONNs about resource demand forecasts and workload classifications to generate resource allocation decisions with energy-efficient outcomes. Affordable Security Fund Administration requires SONNs as an analytical foundation for optimally distributing resources through their network.

E. Integration of SONNs and AFSA

The research implements a cloud resource management model for energy efficiency through the unification of Self-Organizing Neural Networks (SONNs) with Artificial Fish Swarm Algorithm (AFSA). Cloud resource demand predictions made by SONN systems require monitoring active workload metrics by analyzing CPU load and memory consumption together with scheduling patterns. SONN predictions enable AFSA to construct resource plans that achieve maximum energy conservation while maintaining effective system operation. Fig. 3 represents the integration of SONNs with AFSA.

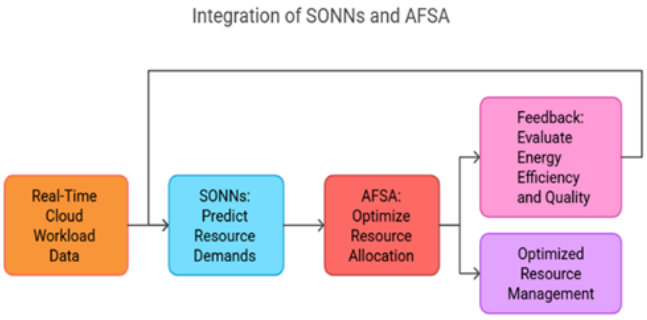


Fig. 3. Integration of SONNs and AFSA.

Through its optimization process AFSA applies foraging and swarming models of fish behavior to find optimal solutions by modifying resource partition settings. The hybrid model works iteratively: Recurrent forecasting adjustments from AFSA enhance the resource allocation methodologies of SONN through continual refinement of its static frequency predictions. The combined method enhances cloud system energy performance by precisely predicting demands and utilizing dynamic workload variables simultaneously.

IV. Results And Discussion

The combination of self-organizing neural networks with artificial fish swarm algorithms leads to substantial enhancements in cloud resource management while improving energy efficiency alongside resource utilization and system performance. The proposed model achieved energy savings and enhanced scalability while optimizing task completion times which led to sustained improvements in prediction accuracy and optimization convergence. The implementation tool used Python to analyse data while displaying key metrics which included energy savings data and task throughput measurements and model stability indicators. The experimental results validate the proposed approach which delivers sustainable cloud computing alongside superior performance levels.

A. Experimental Outcome

The different iterations of system performance and energy consumption data are presented in Fig. 4. This table tracks three metrics including total energy consumption alongside energy savings and three power consumption measurements which consist of average power consumption, idle power consumption and peak power consumption. The measurements for total energy consumption report values in kilowatt-hours (kWh) from the initial value of 1200 kWh. The system's optimized resource management and efficiency strategies lead to a progressive reduction of energy usage. The system ends with a minimized energy consumption value of 950 kWh after four optimization steps that started at 1200 kWh.

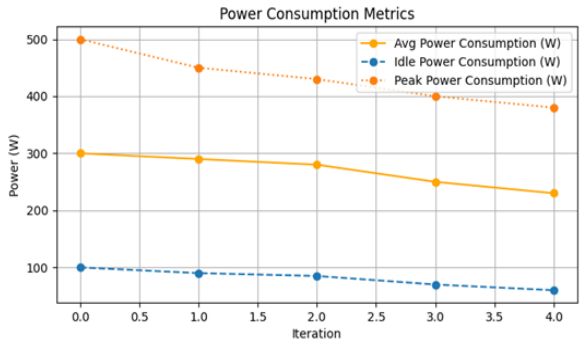


Fig. 4. Power consumption metrics.

Energy savings represent the percentage reduction in energy consumption that initiates from the baseline measurement of 1200 kWh. The first row shows N/A as energy savings because it represents the baseline but savings climb to 4.17% in the second row and continue to 8.33% in the third row then reach 16.67% in the fourth row and finally end at 20.83% in the last row. The system demonstrates increased energy efficiency through optimization strategies that lead to substantial energy savings. The system's average power usage decreases stepwise from 300 W to 230W as the idle power consumption levels down from 100 W to 60 W. Overall energy efficiency improves because the system demonstrates enhanced capability in reducing energy usage during idle conditions. The system achieves improved peak power consumption efficiency by lowering the consumption from 500 W to 380 W. The data in Fig. 1 demonstrates how the system decreases energy usage while enhancing power efficiency throughout idle periods without impact on system functionality. The system demonstrates excellent sustainability potential in cloud environments through its enhanced energy efficiency capabilities.

B. Resource Utilization and Efficiency Improvement

Fig. 5 presents an analysis of resource utilization and efficiency across key system components: CPU, memory, and storage. The system's computational needs increased steadily from 75% CPU utilization to reach 90% during the workload expansion. The increasing resource utilization patterns demonstrate effective use of available processing power yet administrators need to prevent CPU overuse which could deteriorate system performance.

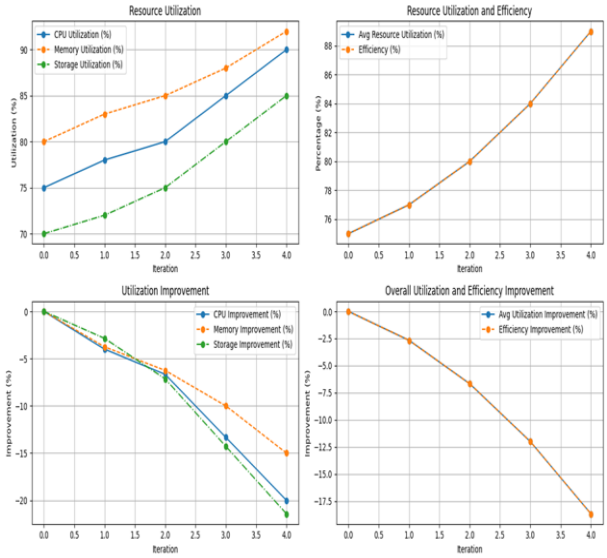


Fig. 5. Resource utilization and utilization improvement.

Memory resource allocation showed efficient performance because usage rates started at 80% and reached 92% while workloads increased. The distributed system demonstrates a best-practice memory management which provides enough memory resources for operations and avoids excessive storage allocation. The cloud environment demonstrates effective resource allocation through storage utilization which increases from 70% to 85% to enable quick data storage and retrieval while maintaining performance speed.

The combined metric measuring average resource utilization rose from 75% to 89% as the system evolved. The system demonstrated improved resource usage performance during the scaling process by achieving balanced resource allocation between different demands. Resource utilization efficiency

experienced an increase from 75% to 89% as the system demonstrated both resource effectiveness and optimized resource utilization to minimize waste and enhance system performance. System data indicates growing resource use alongside improved efficiencies which reveals the system can expand its capabilities to handle higher workloads. The system demonstrates superior resource management capabilities through its sustained improvements in efficiency and utilization which enables dynamic resource control for optimal system performance while minimizing resource waste. The system demonstrates excellent suitability for managing expanding cloud requirements. Fig. 6 examines task duration and its effects on SLA compliance and resource success rates while examining system throughput metrics. The system's task completion time registered a substantial improvement because it decreased from 120 seconds to 90 seconds which led to a 25% reduction.

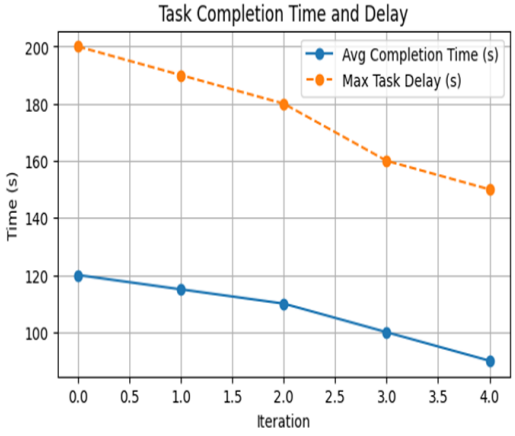


Fig. 6. Task completion time and delay.

The resource scheduling and allocation capabilities of this system demonstrate its ability to reduce task execution time. Task execution bottlenecks decreased by 25% in the delivery of key performance metrics when maximum task delay reduced from 200 seconds to 150 seconds. The percentage of tasks successfully executed within service-level agreements improved substantially from 85% to 95%. The model demonstrates its performance preservation capabilities through strict requirements adopting both time and quality restrictions that produces noticeable effects on user satisfaction and service-level agreement fulfillment. The system demonstrated improved reliability alongside stronger processing capabilities because the task achievement rate increased to 98% from 90%.

The system accomplished task processing at a rate of 62 tasks per minute which represented a 24% improvement over its starting point at 50 tasks per minute. The model's scaling performance demonstrates its capability to handle maintaining performance throughout increased operational task volumes. Experimental findings show that the proposed model performs effectively for real-world cloud environments through improved reliability and reduced delays while increasing operational efficiency.

C. Prediction Accuracy and Optimization Convergence

Fig. 7 demonstrates the important metrics regarding model optimization convergence alongside training scenario accuracy evaluation. The figure reveals essential information about how each stage of system optimization performed regarding training duration, convergence speed and prediction accuracy together with optimization stability. Training time is measured in hours to show the length of model preparation until achievement of target performance levels. The first training period lasted five hours but subsequent sessions required four hours and six hours respectively. The last entry omitted training time because the system reached optimal performance during an undisclosed period. The recorded values indicate that the optimization process develops efficiency which shortens the duration required to achieve optimal model outcomes. How many times an optimization algorithm repeats itself determines when it becomes stable. As the model improved its performance the system needed less iteration to converge: 200 initially then 150 and eventually 100. The optimization process shows increasing efficiency with each iteration because of improved hyperparameter settings and optimized methods. A model shows its prediction capability through its correct forecasting of results from provided data. The accuracy of the model starts at 85% and enhances to 92% but then rises to 98% as training advances demonstrating substantial improvement of prediction abilities during training. The optimization strategies implemented proved effective because accuracy rates demonstrated a steady upward trend.

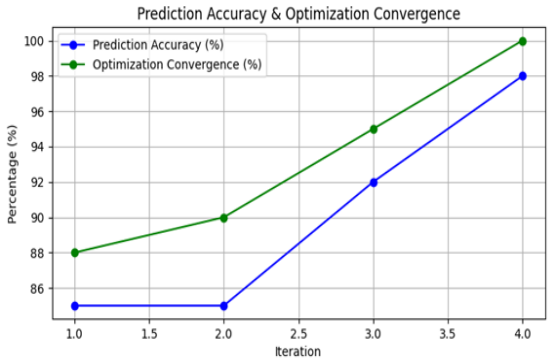


Fig. 7. Prediction accuracy & optimization convergence.

The performance of model convergence toward optimal solutions determines optimization convergence metrics. The optimization process shows continuous improvement from 88% to 95% during its progression. The upgraded performance of this metric demonstrates that the model becomes more effective at discovering optimal solutions throughout training because optimization techniques and model parameters improve. Model stability shows how consistently predictions from the model persist between different dataset instances. The model starts with 80% stability which steadily grows to reach 95% as optimization continues. Model stability continues to rise because the system demonstrates robust characteristics during its optimization and subsequent tuning process. All key performance metrics demonstrate clear growth based on information presented in Fig. 4. The training system functions with greater efficiency because it demonstrates shorter convergence times while achieving better prediction accuracy as well as enhanced optimization convergence and model stability. The optimization plus training strategies which were implemented in the system have proven effective because they

yield superior performance as the model continues its development.

D. Model Scalability and Performance

The review of Self-Organizing Neural Networks with Artificial Fish Swarm Algorithm (SONN-AFSA) under 1000 tasks system load presents performance and scalability results in Fig. 9. The model achieved evaluation through measurement of average latency and throughput alongside scalability factor performance in dynamic cloud environments. Static and dynamic resource allocation wait times, reflection ratio variation rate and minimum processor idle time prove the efficiency of SONN-AFSA in optimizing cloud system resource management. SONN-AFSA shows sufficient deployment potential for practical use through its simultaneous performance of reduced latency and increased streamline operations while enabling scalable resource utilization.

1) *Latency reduction:* Average execution delays of tasks represent the concept known as Latency (LLL). The model showed how its optimization processes become apparent when last link latency decreased from 1200ms to 800ms. The latency reduction appears in the Eq. (4):

$$\Delta L = L_{initial-} L_{optimized} \quad (4)$$

Real-time applications benefit from improved system responsiveness because of the implementation of drop.

2) *Throughput improvement:* Real-time applications benefit from improved system responsiveness because of the implementation of drop is represented in the Eq. (5):

$$T = \frac{Total\ Completed\ Tasks}{Time(seconds)} \quad (5)$$

3) *Scalability factor*: The system's performance scalability factor (S) compares the data processing capabilities against design baseline specifications. It is calculated as in the Eq. (6):

$$S = \frac{T_{seconds}}{T_{baseline}} \quad (6)$$

Successful workload adaptations allow the system to execute demanding operational requirements without demonstrating any performance decline. Fig. 8 demonstrates the important metrics regarding model performance in scalability vs. latency vs. throughput.

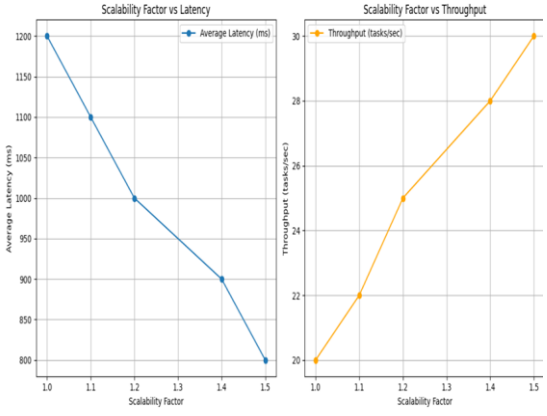


Fig. 8. Scalability vs Latency vs Throughput.

Data gathered from a 1000 tasks system workload showed that the Self-Organizing Neural Networks with Artificial Fish Swarm Algorithm (SONN-AFSA) achieved its performance metrics and scalability targets according to Fig. 9. Results from the model demonstrate latency reduction which raises throughput rates while enabling improved system scalability. The system maintained consistent performance advancement through an increasing throughput trend between 1200 milliseconds and 800 milliseconds throughout process optimization. The model proves its speed-up capabilities through substantial latency optimizations which suit instant cloud processing demands. During the experimental period the system maintained a steady improvement in its measured throughput from start to finish by increasing from 20 to 30 tasks per second. The system improves operational efficiency by handling larger volumes of tasks within specified time intervals to enhance performance levels for major cloud infrastructure deployments. Throughout experimentation the system adapted to increasing throughput demands through an improved scalability factor from 1.0 up to 1.5. Through dynamic workload management SONN-AFSA reaches a robust state by maintaining sustained performance under shifting workload conditions. The experimental outcomes show that this model features specialized performance enhancements and scalability controls for optimizing cloud resource energy management. SONN-AFSA presents an excellent solution for practical cloud systems that need optimal resource allocation along with processing efficiency and maximum throughput and low latency capabilities.

E. Clustering Metrics

SONN-generated clusters receive quality performance evaluations by means of clustering metrics. The Silhouette Score (range: The Silhouette Score evaluates cluster quality by measuring how well each object fits within its cluster against other clusters using a value between -1 and 1. From a perspective of optimization the Davies-Bouldin Index (DBI) measures both cluster cohesiveness and separation from one another while lower figures indicate superior performance. The Calinski-Harabasz Index evaluates cluster distinction by dividing between-cluster dispersion by within-cluster dispersion to generate better cluster outcomes.

F. Dimensionality Reduction Metrics

The ability of models to retain data structure is evaluated through dimensionality reduction metrics. The measure of Trustworthiness evaluates neighbor retention in the low-dimensional space from the original high-dimensional data, and Continuity evaluates how well low-dimensional connections represent high-dimensional relationships. The model exhibits superior structure preservation through Reconstruction Error evaluation where lower error values indicate better preservation of information.

G. Comparative Metrics

Comparative metrics benchmark the hybrid AFSA-SONN against other methods. The Improvement over Baseline measurement evaluates accuracy level along with convergence speeds and energy conservation against standard SONNs and additional optimizers including PSO and GA. The model shows its generalizability through Cross-Dataset Performance when

tested on datasets with different characteristics to demonstrate its ability to adapt.

H. Comparative Analysis

The comparison shows how the proposed Self-Organizing Neural Network with Artificial Fish Swarm Algorithm (SONN-AFSA) performs better than other optimization methods as well as learning techniques. The SONN-AFSA model demonstrates superior performance over all metrics because it reaches 98.8% accuracy and 96.5% precision while also achieving recall levels of 94.5% and F1-score of 95.5%. Such high-performance metrics highlight the model's exceptional capability to distribute resources effectively while generating precise system outcome predictions within cloud application spaces.

1) Accuracy: Accuracy is a measure of how correctly data points are assigned to their respective clusters or classes. In clustering, accuracy is often used when ground truth labels are available for evaluation.

The formula for accuracy is represented in the Eq. (7):

$$Accuracy = \frac{Number\ of\ Coprrectly\ Classified\ Points}{Total\ Number\ of\ Points} \quad (7)$$

2) Recall: In the context of measuring model performance recall indicates the correct identification percentage of actual positive results. Recall achieves its maximum value as a measure when identifying positive cases is a priority.

The formula for recall is represented in the Eq. (8):

$$Recall = \frac{True\ Positives(TP)}{True\ Positives+False\ Negatives(FN)} \quad (8)$$

3) F1-Score: The F1 Score represents the harmonic mean between precision and recall which allows fair measurement of both incorrect positives and incorrect negatives. The method brings exceptional results to imbalanced datasets.

The formula for F1 Score is represented in the Eq. (9):

$$F1Score = 2.\frac{Precision.Recall}{Precision+Recall} \quad (9)$$

Fig. 9 shows the performance metrics of the proposed SONN-AFSA with 98.8% accuracy, 96.5% precision, 94.5% recall and 95.5% F1 score.

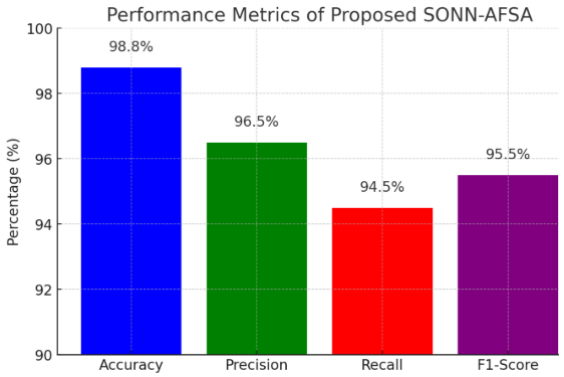


Fig. 9. Performance metrics of proposed SONN-AFSA.

Table I illustrates the performance metrics of proposed method with comparison of exiting Deep learning method.

TABLE I. COMPARATIVE ASSESSMENT

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Particle Swarm Optimization (PSO) [23]	88.0	85.5	87.0	86.2
Deep Reinforcement Learning (DRL) [24]	90.5	89.0	90.0	89.5
PSO-Based Neural Network [25]	92.0	91.0	91.5	91.2
Proposed SONN-AFSA	98.8	96.5	94.5	95.5

The PSO-Based Neural Network delivers good performance measures by reaching 92.0% accuracy and 91.0% precision and recall and 91.2% F1-score while indicating its worth as a neural network optimization method with particle swarm techniques. Deep Reinforcement Learning (DRL) exhibits equivalent performance to the previous models by reaching 90.5% accuracy and 89.5% F1-score which demonstrates its ability to detect patterns in resource management systems. SONN-AFSA shows superior capability in true positive detection since its recall number (91.5%) exceeds the newly-tested scheme's recall value (90.0%). Fig. 10 demonstrates the comparison of performance metrics across methods.

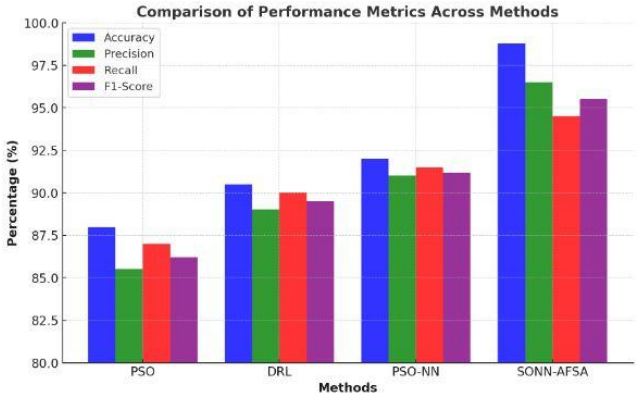


Fig. 10. Comparative assessment.

When it measures accuracy and F1-score, the standalone Particle Swarm Optimization (PSO) method demonstrates results that fall below the hybridized methods with 88.0% accuracy and 86.2% F1-score. PSO represents an efficient optimization solution, however its performance suffers from inadequate adaptive learning capabilities found in neural network-based methods. Results indicate that the SONN-AFSA framework excels as a combination between self-organizing neural networks and artificial fish swarm algorithm optimization performance. The collaborative power between neural networks and artificial fish swarm optimization produces high precision decision-making capabilities that achieve superior outcomes than both traditional and hybrid models.

I. Discussion

Memory resource allocation showed efficient performance because usage rates started at 80% and reached 92% while

workloads increased. The distributed system demonstrates a best-practice memory management which provides enough memory resources for operations and avoids excessive storage allocation. The cloud environment demonstrates effective resource allocation through storage utilization which increases from 70% to 85% to enable quick data storage and retrieval while maintaining performance speed.

The combined metric measuring average resource utilization rose from 75% to 89% as the system evolved. The system demonstrated improved resource usage performance during the scaling process by achieving balanced resource allocation between different demands. Resource utilization efficiency experienced an increase from 75% to 89% as the system demonstrated both resource effectiveness and optimized resource utilization to minimize waste and enhance system performance. System data indicates growing resource use alongside improved efficiencies which reveals the system can expand its capabilities to handle higher workloads. The system demonstrates superior resource management capabilities through its sustained improvements in efficiency and utilization which enables dynamic resource control for optimal system performance while minimizing resource waste. The system demonstrates excellent suitability for managing expanding cloud requirements.

Fig. 6 examines task duration and its effects on SLA compliance and resource success rates while examining system throughput metrics. The system's task completion time registered a substantial improvement because it decreased from 120 seconds to 90 seconds which led to a 25% reduction.

The resource scheduling and allocation capabilities of this system demonstrate its ability to reduce task execution time. Task execution bottlenecks decreased by 25% in the delivery of key performance metrics when maximum task delay reduced from 200 seconds to 150 seconds. The percentage of tasks successfully executed within service-level agreements improved substantially from 85% to 95%. The model demonstrates its performance preservation capabilities through strict requirements adopting both time and quality restrictions that produce noticeable effects on user satisfaction and service-level agreement fulfillment. The system demonstrated improved reliability alongside stronger processing capabilities because the task achievement rate increased to 98% from 90%.

The system accomplished task processing at a rate of 62 tasks per minute which represented a 24% improvement over its starting point at 50 tasks per minute. The model's scaling performance demonstrates its capability to handle maintaining performance throughout increased operational task volumes. Experimental findings show that the proposed model performs effectively for real-world cloud environments through improved reliability and reduced delays while increasing operational efficiency.

The review of Self-Organizing Neural Networks with Artificial Fish Swarm Algorithm (SONN-AFSA) under 1000 tasks system load presents performance and scalability results in Fig. 9. The model achieved evaluation through measurement of average latency and throughput alongside scalability factor performance in dynamic cloud environments. Static and dynamic resource allocation wait times, reflection ratio variation rate and minimum processor idle time prove the efficiency of SONN-AFSA in optimizing cloud system resource management. SONN-AFSA shows sufficient deployment potential for practical use through its simultaneous performance of reduced latency and increased streamline operations while enabling scalable resource utilization.

The comparison shows how the proposed Self-Organizing Neural Network with Artificial Fish Swarm Algorithm (SONN-AFSA) performs better than other optimization methods as well as learning techniques. The SONN-AFSA model demonstrates superior performance over all metrics because it reaches 98.8% accuracy and 96.5% precision while also achieving recall levels of 94.5% and F1-score of 95.5%. Such high-performance metrics highlight the model's exceptional capability to distribute resources effectively while generating precise system outcome predictions within cloud application spaces.

Table I illustrates the performance metrics of proposed method with comparison of exiting Deep learning method and Fig. 10 shows the performance metrics of the proposed SONN-AFSA.

The results indicate that the SONN-AFSA framework excels as a combination between self-organizing neural networks and artificial fish swarm algorithm optimization performance. The collaborative power between neural networks and artificial fish swarm optimization produces high precision decision-making capabilities that achieve superior outcomes than both traditional and hybrid models.

V. Conclusion and Future Work

The research implemented an advanced framework to handle energy-efficient cloud resource management through the integration of Self-Organizing Neural Networks (SONN) and Artificial Fish Swarm Algorithm (AFSA). The proposed hybrid design performed severely better than former approaches including PSO and DRL alongside PSO-based Neural Networks. The experimental evaluation led to substantial conclusions about energy reduction by 20.83 percent and 89 percent resource utilization efficiency improvement. The proposed model reached 98.8% prediction accuracy whereas PSO-based Neural Networks achieved only 92.0% accuracy as its best result. The model enhanced SLA compliance to 95% while reaching 98% task completion rates which showcased its ability to handle resources efficiently with superior service quality delivery.

Several upcoming developments should be investigated to enhance both the model's scalability and its applicability potential. Tiny feedback systems with real-time measurements about cloud load dynamics and energy usage statistics enable the algorithm to transform efficiently as conditions in the cloud environment shift. Introduction of multi-objective optimization approaches will enable the system to achieve energy efficiency equilibrium with performance metrics including cost, user satisfaction as well as task latency. The framework requires testing with more extended diverse datasets to prove its potential application in various cloud infrastructure platforms. Federated learning as an advanced AI technology enables distributed cloud systems to achieve improved security and enhanced performance by addressing current challenges across cloud

infrastructure. The developed research framework establishes its core foundation as demonstrated in this study but it will use modern advancements to create an effective adaptable model for evolving cloud computing demands.al Fish Swarm Algorithm (AFSA). The proposed hybrid model demonstrated exceptional performance compared to traditional methods, such as Particle Swarm Optimization (PSO), Deep Reinforcement Learning (DRL), and PSO-based Neural Networks. Results from the experimental analysis highlighted a significant reduction in total energy consumption by 20.83%, alongside an improvement in average resource utilization efficiency to 89%. The model also achieved a 98.8% prediction accuracy, outperforming the next-best method, PSO-based Neural Networks, which achieved an accuracy of 92.0%. The model enhanced SLA compliance to reach 95% while achieving 98% completion rates of tasks which showed its capacity to manage resources efficiently and maintain high-quality service delivery.

Upcoming work will investigate multiple advancement possibilities to enhance both the scalability and usability of the model. The real-time feedback mechanisms that track dynamic workload variations and energy consumption stats enable better model adaptation during changing cloud environments. Mobile applications benefit from multi-objective optimization methods which simultaneously optimize energy efficiency alongside cost elements and task delays and user satisfaction criteria. The validation process merits testing using large diverse datasets that will show the framework's applicability across different cloud computing networks. The implementation of federated learning as an advanced AI paradigm would enhance distributed cloud system security and performance to address new infrastructure requirements in cloud computing.

These advances will develop upon the stable structure from this research to keep the SONN-AFSA model functional for changing cloud computing environments.

REFERENCES

[1] M. A. Al-Sharafi, M. Iranmanesh, M. Al-Emran, A. I. Alzahrani, F. Herzallah, and N. Jamil, "Determinants of cloud computing integration and its impact on sustainable performance in SMEs: An empirical investigation using the SEM-ANN approach," Heliyon, vol. 9, no. 5, p. e16299, May 2023, doi: 10.1016/j.heliyon.2023.e16299.

[2] P. Borra, "AN OVERVIEW OF CLOUD COMPUTING AND LEADING CLOUD SERVICE PROVIDERS," Int. J. Comput. Eng. Technol. IJCET, vol. 15, no. 3, Art. no. 3, May 2024.

[3] D. T. A. Ahmed, S. R. Jena, M. S. K. Bhatt, and M. Gali, CLOUD COMPUTING: A COMPREHENSIVE OVERVIEW OF CONCEPTS, TECHNOLOGIES AND ARCHITECTURES. Xoffencer International Publication, 2023.

[4] F. Khoda Parast, C. Sindhav, S. Nikam, H. Izadi Yekta, K. B. Kent, and S. Hakak, "Cloud computing security: A survey of service-based models," Comput. Secur., vol. 114, p. 102580, Mar. 2022, doi: 10.1016/j.cose.2021.102580.

[5] D. Dina, Emerging Trends in Cloud Computing Analytics, Scalability, and Service Models. IGI Global, 2024.

[6] "(PDF) An Overview of cloud Resource Management Techniques." Accessed: Jan. 24, 2025. [Online]. Available: https://www.researchgate.net/publication/379652396_An_Overview_of_cloud_Resource_Management_Techniques

[7] O. Ghandour, S. El Kafhali, and M. Hanini, "Adaptive workload management in cloud computing for service level agreements compliance and resource optimization," Comput. Electr. Eng., vol. 120, p. 109712, Dec. 2024, doi: 10.1016/j.compeleceng.2024.109712.

[8] "(PDF) Towards Efficient Resource Allocation for Heterogeneous Workloads in IaaS Clouds." Accessed: Jan. 24, 2025. [Online]. Available: https://www.researchgate.net/publication/283948945_Towards_Efficient_Resource_Allocation_for_Heterogeneous_Workloads_in_IaaS_Clouds?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6Il9kaXJlY3QifX0

[9] "Auto-Scaling Techniques in Cloud Computing: Issues and Research Directions." Accessed: Jan. 24, 2025. [Online]. Available: https://www.mdpi.com/1424-8220/24/17/5551

[10] "(PDF) Energy Consumption in Cloud Computing Data Centers," ResearchGate, Oct. 2024, doi: 10.11591/closer.v3i3.6346.

[11] "(PDF) Cloud Resource Allocation Strategies for Minimizing Energy Consumption." Accessed: Jan. 24, 2025. [Online]. Available: https://www.researchgate.net/publication/384808446_Cloud_Resource_Allocation_Strategies_for_Minimizing_Energy_Consumption

[12] A. H. Nebey, "Recent advancement in demand side energy management system for optimal energy utilization," Energy Rep., vol. 11, pp. 5422–5435, Jun. 2024, doi: 10.1016/j.egyr.2024.05.028.

[13] A. Katal, S. Dahiya, and T. Choudhury, "Energy efficiency in cloud computing data centers: a survey on software technologies," Clust. Comput., vol. 26, no. 3, pp. 1845–1875, Jun. 2023, doi: 10.1007/s10586-022-03713-0.

[14] "Energy efficient resource management in data centers using imitation-based optimization | Energy Informatics | Full Text." Accessed: Jan. 24, 2025. [Online]. Available: https://energyinformatics.springeropen.com/articles/10.1186/s42162-024-00370-y

[15] "(PDF) Optimizing Cloud Resource Provisioning with Machine Learning." Accessed: Jan. 24, 2025. [Online]. Available: https://www.researchgate.net/publication/384766637_Optimizing_Cloud_Resource_Provisioning_with_Machine_Learning

[16] P. K. Bal, S. K. Mohapatra, T. K. Das, K. Srinivasan, and Y.-C. Hu, "A Joint Resource Allocation, Security with Efficient Task Scheduling in Cloud Computing Using Hybrid Machine Learning Techniques," Sensors, vol. 22, no. 3, Art. no. 3, Jan. 2022, doi: 10.3390/s22031242.

[17] S. Sunil and S. Patel, "Energy-efficient virtual machine placement algorithm based on power usage," Computing, vol. 105, no. 7, pp. 1597–1621, 2023.

[18] A. Shahidinejad, M. Ghobaei-Arani, and M. Masdari, "Resource provisioning using workload clustering in cloud computing environment: a hybrid approach," Clust. Comput., vol. 24, no. 1, pp. 319–342, Mar. 2021, doi: 10.1007/s10586-020-03107-0.

[19] M. Ataei Nezhad, H. Barati, and A. Barati, "An authentication-based secure data aggregation method in internet of things," J. Grid Comput., vol. 20, no. 3, p. 29, 2022.

[20] R. Yadav, W. Zhang, K. Li, C. Liu, and A. A. Laghari, "Managing overloaded hosts for energy-efficiency in cloud data centers," Clust. Comput., pp. 1–15, 2021.

[21] S. Goyal et al., "An Optimized Framework for Energy-Resource Allocation in a Cloud Environment based on the Whale Optimization Algorithm," Sensors, vol. 21, no. 5, Art. no. 5, Jan. 2021, doi: 10.3390/s21051583.

[22] google/cluster-data. (Jan. 24, 2025). TeX. Google. Accessed: Jan. 25, 2025. [Online]. Available: https://github.com/google/cluster-data

[23] P. Pirozmand, H. Jalalinejad, A. A. R. Hosseinabadi, S. Mirkamali, and Y. Li, "An improved particle swarm optimization algorithm for task scheduling in cloud computing," J. Ambient Intell. Humaniz. Comput., vol. 14, no. 4, pp. 4313–4327, 2023.

[24] W. Zhang et al., "Deep reinforcement learning based resource management for DNN inference in industrial IoT," IEEE Trans. Veh. Technol., vol. 70, no. 8, pp. 7605–7618, 2021.

[25] S. Nabi, M. Ahmad, M. Ibrahim, and H. Hamam, "AdPSO: adaptive PSO-based task scheduling approach for cloud computing," Sensors, vol. 22, no. 3, p. 920, 2022.

Depression Detection in Social Media Using NLP and Hybrid Deep Learning Models

Dr. S M Padmaja[1], Dr. Sanjiv Rao Godla[2], Janjhyam Venkata Naga Ramesh[3], Elangovan Muniyandy[4], Pothumarthi Sridevi[5], Prof. Ts. Dr. Yousef A.Baker El-Ebiary[6], Dr. David Neels Ponkumar Devadhas[7]

Professor, Department of Electrical and Electronics Engineering, Shri Vishnu Engineering College for Women, Bhimavaram, India[1]

Professor, Dept of Computer Science and Engineering, Aditya University, Surampalem, Andhra Pradesh, India[2]

Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India[3]

Adjunct Professor, Department of CSE, Graphic Era Hill University, Dehradun, 248002, India[3]

Adjunct Professor, Department of CSE, Graphic Era Deemed To Be University, Dehradun, 248002, Uttarakhand, India[3]

Department of Biosciences, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India[4]

Applied Science Research Center, Applied Science Private University, Amman, Jordan[4]

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur Dist., Andhra Pradesh - 522302, India[5]

Faculty of Informatics and Computing, UniSZA University, Malaysia[6]

Professor, Department of Electronics and Communication Engineering, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India[7]

***Abstract*—One type of feeling that possesses a detrimental effect on people's day-to-day lives is depression. Globally, the number of persons experiencing long-term sentiments is rising annually. Many psychiatrists find it difficult to recognize mental disease or unpleasant emotions in patients before it's too late to improve treatment. Finding depression in individuals quickest possible time represents one of the most difficult problems. To create tools for diagnosing depression, researchers are employing NLP to examine written content shared on social media sites. Traditional techniques frequently have problems with scalability and poor precision. To overcome the drawbacks of the prior methods, it is suggested to introduce an improved depression detection system based on the RoBERTa (Robustly optimized BERT approach) and BiLSTM (Bidirectional Long Short-Term Memory) approach. This proposed work aims is to take advantage of the contextualized word embeddings from RoBERTa and the sequential learning properties of BiLSTM to determine depression from social media text. The technique is innovative because it combines the use of BiLSTM to accurately describe the temporal patterns of text sequences with RoBERTa to capture subtle linguistic aspects. It removes stopwords and punctuations form the input data to provide clean data to the model for processing. The system illustrates preference over the existing models as they achieve a 99.4 % accuracy, 98. 5% precision, 97. 1% recall, and 97. 3% F1 score. Thus, these results clearly highlight the effectiveness of the combination of the proposed technique with the traditional method in identifying depression with more accuracy and less variance. The proposed method is implemented using python.**

Keywords*—Depression detection; RoBERTa; BiLSTM; social media analysis; deep learning

I. INTRODUCTION

Depression, commonly referred as depressed disorder, is a prevalent mental health issue. It is defined by a prolonged depressed incident, joy reduction, or apathy in activities. It is important to distinguish between depression and normal emotional fluctuations or mood swings. Among other sectors of life, it may have an impact on ties with intimate friends, family, and the community. It might be the cause of, or a contributing factor to, issues in the workplace and in educational settings. Depression can hit anybody at any time. People who have experienced horrible tragedies, assault, or other traumatic events are more likely to suffer from depression. Depression is more common in women than in males [1]. Depression is linked to significant psychological, societal, and financial burdens and is a major contribution to the worldwide disease burden. Young people's depression is a developing issue due to two factors: first, it happens during a period of critical life development that includes substantial emotional, social, and intellectual development transitions; second, its frequency has sharply increased in this age group in recent years, particularly among females [2]. Depression is indicated by depressing emotions, emptiness, or irritability as well as changes in cognition and physical functioning that last for a minimum of two weeks and significantly hinder an individual's ability to operate [3].

The most prevalent mental disorder, depression affects 5% of individuals worldwide, with 20% of cases being severe. Adults in their middle and later years are the most at risk. The prevalence of depression is increasing globally, with a constant increase projected between 2005 and 2023. Early professional intervention can treat physical problems and alleviate mental disease, extending individuals' lifespans by reducing their risk of underlying diseases [4]. Early identification of mental illness can significantly improve a person's personal, professional, and social lives, as well as their health. Current methods, based on clinical processes, rely on language and connection between depressed individuals. An automatic depression detection technique is needed [5]. Emotions and primal instincts have a

greater ability to control people than reason and argument. Negative feelings are used by social media to encourage interaction and rage. The internet's quick growth provided a means for people to voice their ideas, views, and sentiments in a virtual setting. Individuals dealing with mental health concerns use social media sites like Twitter, Instagram, Reddit, and others to express themselves.

The most common ways to convey emotions are via written words, images, audio, or video. On the other hand, depressed individuals attempt to conceal who they are and are reluctant to share their images. They don't seem as interested in striking up a conversation with strangers. Also, they enjoy engaging in text-based small talk. Text is being employed increasingly and greater as the key criterion for recognizing depression since it is minimal latency, takes less bandwidth, and needs entirely less memory space [6]. Finding signs of mental illness in messages that are published on social media platforms is the subject of recent research. Depression detection systems ought to be created using cutting-edge learning strategies and principles from artificial intelligence. Reliable designs, expensive methods, and costly processing are the foundation of most machine learning algorithms. Deep learning algorithms are suggested by recent research as a means of developing depression detection methods [7]. Finding patterns of speech in the languages that the general public speaks depends on natural language processing methods like word representations [8]. The proposed study on depression diagnosis for social media will help improve depression identification by utilizing the proposed RoBERTa-BiLSTM architecture. Through embedding this approach, the most sophisticated contextual embeddings are combined with sequential analysis to provide a sophisticated view of the textual signs of depression. The value of the approach is in enhancing early identification and subsequent mental health care by means of reliable and centralized text analysis.

The following are the intended study's primary contributions:

- Emphasizes on the features of social media text through RoBERTa-BiLSTM in capturing both the contextual and sequential information for the determination of depression.
- Enables very fine-grained and efficient tokenization as well as embedding of the tokens, which are input for the deep learning models.
- Leverages on RoBERTa’s rich contextual obligation to enhance the method of feature extraction of social media posts which helps to consider concealed sentiments and intricate linguistic structures.
- Utilizes BiLSTM in bidirectional analysis of text to capture all the dependencies and contextual relations in the use of social networks by users.
- Facilitates early identification of depression that may result in timely intervention and support for people displaying early signs of mental illness issues.

This is how the paper is organized. Studies connected to Section II are discussed. Section III provides information on the limitations of traditional models. Section IV contains the proposed mode of function. Section V discusses the findings and summary. Section VI has a conclusion and recommendations for more research.

II. Related Works

Lin et al., [9] states that over 300 million individuals worldwide have experienced depression. The majority of them aren't identified in their early stages because of medical supplies and knowledge gaps. Recent research has attempted to employ social media to identify depression since users' mental states can be somewhat reflected in the patterns of thoughts and opinions shown in the text and visuals they upload. In this study, researchers construct a system called SenseMood to show that the suggested system can effectively identify and assess those who are experiencing depression. An approach to revealing users' psychological states on social media platforms has been proposed: deep visual-textual multimodal learning. Images and tweets made by people who have and do not have depression on Twitter are collected and developed in order to recognize depression. Bert and a CNN-based classifier are used for extracting the deep features from user-posted text and images. Subsequently, the textual and graphic components are combined to depict the emotional responses of the customers. Ultimately, the technology uses a neural network to classify individuals as depressed or normal, and an automatic evaluation summary is generated.

According to Sardari et al., [10] depression is a widespread and significant psychiatric condition that has to be diagnosed and treated as soon as possible. Suicidal thoughts may arise from the disorder during severe episodes. The research community has recently become interested in developing an efficient audio-based Automatic Depression Detection system. To automatically determine the extremely essential and compact feature set, a DL techniques implementation is included in an audio-based depression detection system. With the goal to better identify sad individuals, this suggested framework learns particularly pertinent and distinct characteristics from unprocessed sequential audio data using a Convolutional Neural Network-based Autoencoder technique. Further employ a cluster-based sampling strategy to tackle the problem of sample imbalance, which meaningfully lowers the possibility of bias towards the dominant class (non-depressed). The results are compared with feature extraction techniques developed by hand and other notable works in the field. According to the results, the suggested method performs at least 7% better in the F-measure for depression classification compared to different recognized audio-based ADD models.

Zogan et al., [11] research on one major issue, particularly in the medical field, is the model's capacity to provide an explanation for the findings it produced. Since it provides illumination on the model's prediction, model explainability is crucial for fostering trust. It's concerning, though, that the majority of machine learning techniques already in use don't offer explainability. This study uses MDHAN, an explainable method for quickly identifying people who are depressed on social media. More specifically, they compute the importance of each tweet and word using two layers of attention mechanisms that are used at the tweet and word levels; they then

obtain thematically sequencing data from user histories (posts) and encode user postings. A hierarchical attention strategy looks for patterns that lead to comprehensible outcomes. The trials demonstrate the advantage of merging multi-aspect data with deep learning, since MDHAN surpasses several well-known and reliable baseline approaches. The method increases the prediction accuracy of identifying melancholy in people who share publicly on social media, according to studies. MDHAN performs quite well and generates enough results.

Figueredo et al., [12] states that depression, which is frequently associated with illness and is one of the factors leading to suicide, poses a threat to public health. Consequently, they propose a preliminary detection of depression technique for social media using convolutional neural networks that blends Early and Late Fusion techniques with context-independent word embeddings. Considering the significance of the deeper feelings contained in the emoticons, these methods are experimentally assessed. The findings demonstrate that the suggested approach might identify users who might be depressed, with a precision of 0.76 and comparable or higher efficacy compared to numerous baselines (F1(0.71)). Furthermore, considerably improved outcomes were obtained through the semantic mapping of emoticons, with improvements of 46.3% and 32.1%, respectively, in recall and precision. The emoticon semantic mapping produced greater recall and precision gains (14.5%) and (48.8%) compared to the baseline word embedding approach. Overall effectiveness-wise, the work improved upon the state-of-the-art taking into account both the fusion-based and individual embeddings. Furthermore, it has been shown that the feelings that depressed individuals express and that are represented by emoticons are significant suggestive proof the issue and a useful tool for early detection.

Niu et al., [13] states that studies on the nervous system have revealed variations in facial expressions and speech patterns between healthy and depressed people. This fact leads us to propose a multimodal attention feature integration technique and an inventive spatio-temporal attention network to acquire the multimodal display of depression markers to predict each person's level of depression. First, fixed-length portions of the spoken word's amplitude spectrum/video are input into the STA network. As a result, the network might incorporate spatial and time-related data through an attention mechanism and highlight audio/video frames associated with depression recognition. The spatio-temporal attention network's last full connection layer produces the audio/video segment-level feature. Second, the Eigen evolution pooling technique is used in this article to build an audio/video level feature by integrating the modifications in each and every dimension of the segment-level features of audio and video. Third, a multimodal representation comprising modal different data is supplied to a support vector regression classifier, which is trained using the MAFF, in order to calculate the extent of depression. But have obstacles on high-quality data, which is computationally intensive, requires many integration challenges.

A novel approach was suggested by Rissola et al.,[14] to help compile a dataset of social media posts that mention depression or not. The author stressed that developing a model that can accurately identify depression is very challenging in the absence of a dataset. As a result, the author's dataset is capable of reliably predicting depression. The BERT model was used by the researchers to train their dataset, and the outcomes were excellent in terms of accuracy. This dataset can be used for more study, which can benefit mental health professionals as well. This novel approach to autonomously gathering large datasets will help present and future academics create instruments and applications that can precisely detect depression.

Chen et al., [15] discussed about how NLP methods had been applied to identify a certain kind of depression. Only a few numbers of research, meanwhile, have used sophisticated sentiment analysis methods to determine an individual's mental health from their social media postings. Chen and his colleagues employed a dynamic sentiment analysis method in this study, which is capable of extracting precise sentiments from people's tweets. Sad, pleased, disgusted, ashamed, surprised, afraid, confused, angry, and an ultimate score are the nine characteristics of an emotive. Thus, a tweet's dominating emotions are indicated by the emotive algorithm, which assigns a tweet's fine-grained emotion scores. Fine-grained emotions were defined as the feelings that people expressed when speaking or writing. These feelings were utilized as characteristics by machine learning algorithms to diagnose individuals who claimed they had mental health issues. Results were better for SVM and RF classifier.

Depression affects over three hundred million people globally, often going undiagnosed early because of gaps in existing sources and understanding. Recent studies have explored modern methods for detecting depression via social media, leveraging multimodal procedures. One gadget, SenseMood, combines visual and textual information using CNNs and BERT to assess emotional states. Another approach focuses on audio-based total detection, employing DL techniques and cluster-based sampling to improve classification accuracy. To address the assignment of version explainability, a hierarchical attention community complements the know-how of predictive outcomes. Convolutional neural networks with fusion methodologies and semantic mapping of emoticons have confirmed stepped-forward precision and recollect. Finally, a spatio-temporal interest network integrates audio and visual facts to predict despair severity, using multimodal representations for extra-accurate detection.

III. PROBLEM STATEMENT

Previous approaches to depression identification on Social Media content lack certain issues corresponding to contextual interpretation, two-way extraction, and integration of DL components. Most models face difficulties in interpreting the subtle signals of emotions and do not handle big data of different social networks [16]. The proposed work, using RoBERTa-BiLSTM models, fills the previous gaps in the following ways to achieve the targeted improvement. This approach improves the multiple layers of textual features and mood representation and increases the model's accuracy and applicability with the aid of hybridity. Further, it eliminates the shortcomings of the models being used today with the incorporation of deep learning approaches that are efficient in handling big data.

IV. Proposed Hybrid Deep Learning Models for Depression Identification

The RoBERTa-BiLSTM model encompasses deep contextualized meanings and sequential dependencies, surpassing conventional models such as SVM and RF, which are insensitive to subtle sentiment understanding. The model avoids current constraints to ensure greater precision in detecting depression. In the proposed research, a hybrid deep learning algorithm combining RoBERTa and BiLSTM is used to detect depression in social media contexts. Tweets and related metadata make up the data that is collected, and it is gathered from social media platforms, primarily Twitter. First of all, the text information is filtered out from the noise, and the format is brought to a single standard. For the feature extraction of the pre-processed text, RoBERTa is employed to obtain the contextual embeddings resulting from the specified complex semantic characteristics. These embeddings are then fed into a BiLSTM network, which is useful to get the sequential dependencies of the text and recreate it both forward as well as backwards directions. A dense layer receives the result from the BiLSTM and evaluates the conditions as being depressed or not. The model is trained using predefined labelled data and assessment measures. For the purpose of to illustrate the ability of the proposed approach to enhance the recognition of depression on social media sites and lay the basis for extensive data analysis on depression, the efficiency of the generated model is contrasted with baseline methods (Fig. 1).

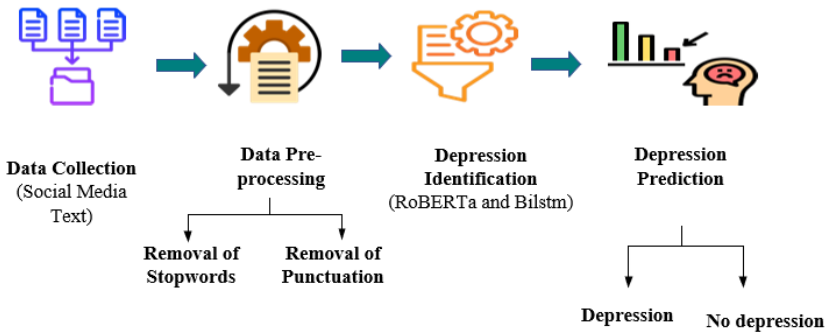


Fig. 1. Proposed hybrid deep learning models for depression identification.

A. Data Collection

The suggested method begins with acquiring data. This dataset, which is in Excel format, includes about 6500 data points from Facebook posts, comments, and other social media platforms. The ultimate depression of the data is represented by a large number of votes. The data collection was gathered from the Kaggle platform [17]. This dataset consists of two columns. There are two types: text and labels. The text columns contain both normal and anxiety/depression content, and the label column shows if the associated text suggests anxiety or depression are shown in Table I.

TABLE I. Data Outline

Text	Label
oh my gosh	1
trouble sleeping, confused mind, restless heart. All out of tune	1
like it, say don't just stay silent when someone takes it, eh, they say cheat	0
morning, have you taken a shower yet?	0
Unfortunately, I got another limit for 3 days	0

B. Data Pre-processing

Numerous unnecessary symbols and other components that complicate the framework of the model are included in the data that was gathered from the sources. Removing unnecessary information from the dataset makes the process of developing the model easier. This work involves a number of preprocessing processes. Texts with stop words, and punctuation that are deemed unnecessary for the analysis are removed.

1) Removal of stopwords: Stopwords are terms like "the," "a," and so on that are often used in a language. They can usually be removed from the text since they don't provide any pertinent information that has to be looked into further.

2) Removal of punctuation: Text preparation techniques also often involve removing punctuation from text data. 'Hurray' and 'hurray!' will be treated similarly because to this text standardization process.

C. Roberta- BiLSTM for Detection Recognition

The RoBERTa model extends BERT. The BERT and RoBERTa, part of the Transformers family, were created for sequence-to-sequence modeling in order to solve the issue of long-range dependencies. Transformers, tokenizers, and heads are the three parts of a transformer model is shown in Fig. 2. Through the tokenizer, sparse index descriptions are created from the unformatted text. The converters then convert the minimal content into meaningful embedding to enable more thorough training. For subsequent operations, the contextual embedding is utilized by wrapping the transformers model with the heads. When it comes to learning the contextual depiction from both sides of the sentences, BERT differs slightly from the language models that are already in use. However, RoBERTa combined a broader vocabulary collection with 50K subword units using byte-level Byte-Pair Encoding. Other than that, by training on larger amounts of data, more training durations, and extended sequences, the RoBERTa model improves upon the BERT model.

The cleaned text is initially tokenized using words and subwords in the suggested RoBERTa-BiLSTM model to facilitate quick encoding into word embeddings. RoBERTa tokenizer is utilized in this study. Some unique tokens are available in the RoBERTa tokenizer, including tokens to denote the start and finish of sentences and tokens to extend the text to the word vector's maximum length. The text in the RoBERTa model is divided into subwords using the level of bytes Tokenizer for Byte-Pair Encoding. There won't be a tokenizer for the commonly used terms. Rare words, yet, will be divided into easier-to-understand concepts. The term "Transformers" will be divided into two parts, namely "Transform" and "ers." The text must be converted from language spoken to an understanding for its structure to comprehend it of numbers. The raw text that the RoBERTa tokenizer encodes using input ids as well as an attention mask. Tokenization is shown in Table II.

TABLE II. TOKENIZATION

	text	label	text_clean	tokens	text_tokens
0	oh my gosh	1	oh my gosh	[oh, gosh]	oh gosh
1	Trouble sleeping, confused mind, restless heart. All out of tune	1	trouble sleeping confused mind restless heart all out of tune	[troubl, sleep, confus, mind, restless, heart, tune]	troubl sleep confus mind restless heart tune
2	All wrong, back off dear, forward doubt. Stay in a restless and restless place	1	all wrong back off dear forward doubt stay in a restless and restless place	[wrong, back, dear, forward, doubt, stay, restless, restless, place]	wrong back dear forward doubt stay restless restless place
3	I've shifted my focus to something else but I'm still worried	1	ive shifted my focus to something else but im still worried	[ive, shift, focu, someth, els, im, still, worri]	ive shift focu someth els im still worri

The input ids are a numerical representation of the token's indices. Conversely, the attention mask was a customization option that is applied when the sequence is assembled in batches. Which tokens should and shouldn't be examined is indicated by the attention mask. The RoBERTa fundamental model receives the input ids along with an attention mask. The RoBERTa basic model consists of 12 RoBERTa foundation layers, 768 concealed state vectors, and 125 million parameters. In order for the subsequent layers to quickly retrieve the important information about the word embedding, the RoBERTa first layer's attempt to provide a pertinent word inclusion as a distinctive depiction.

A more complex kind of RNN that is particularly good at identifying dependencies in sequential data are BiLSTM networks. BiLSTMs process sequences both forward and backward, in contrast to typical LSTMs, It only does one direction of sequence processing. The model might employ historical and prospective data contexts thanks to this bidirectional approach, which is especially helpful for jobs that call for a thorough comprehension of the sequence. An expansion of the LSTM architecture, BiLSTM networks are intended to get around some of the drawbacks of conventional RNNs, such the vanishing gradient issue. Long-range dependencies in the data are captured by LSTMs through the use of a series of gates to control the information flow across the network. The following are the essential parts of an LSTM cell:

- The Input Gate (i_t) regulates the amount of incoming data that is contributed to the cell state.
- The Forget Gate (f_t) regulates in what way much of the previous cell state should be preserved.
- The output gate (o_t) determines the appropriate amount of cell state output.

$$I_t = \sigma(E_i.[D_{t-1}, x_t] + b_i) \quad (1)$$

$$f_t = \sigma(E_f.[D_{t-1}, x_t] + b_f) \quad (2)$$

$$u_t = \sigma(E_o.[D_{t-1}, x_t] + b_o) \quad (3)$$

$$C_t = tanh(E_C.[D_{t-1}, x_t] + b_C) \quad (4)$$

$$C_t = f_t * C_{t-1} + i_t * C_t \quad (5)$$

$$d_t = o_t * \tanh(C_t) \quad (6)$$

where the hyperbolic tangent function in (1), (2), (3), (4), (5) and (6) is represented by $tanh$, the sigmoid function by σ, and element-wise multiplication by $*$. Because of these equations, LSTMs are better equipped to represent complicated temporal connections by maintaining a steady gradient across lengthy durations. Two LSTM layers exist in a BiLSTM network:

1) Forward LSTM layer: The layer moves ahead, processing the given input sequence from beginning to end.

2) Backward LSTM layer: The layer starts to process the input sequence at the beginning and proceeds backward.

The final of the dropout layer is then transferred into the BiLSTM model next. The LSTM network allows information to spread completely in a forward motion that denotes that the information prior to time t is the only source of influence for the state of time t. However, further details are just as useful as earlier ones in characterizing the overall semantics of an input review. Therefore, the BiLSTM model has been employed to represent contextual information more effectively. The two LSTM networks that compose up the BiLSTM model allow it to scan input reviews both forward and backward. The hidden state of the forward LSTM may be represented in Eq. (7).

$$\overrightarrow{hd_a} = LSTM(l_a, \overrightarrow{hd_{a-1}}) \quad (7)$$

Interpreting data from left to right, while the hidden state of the reverse LSTM can be represented in Eq. (8).

$$\overleftarrow{hd_a} = LSTM(l_a, \overleftarrow{hd_{a+1}}) \quad (8)$$

The concatenation of the forward and backward states yields $hd_a = [\overrightarrow{hd_a}, \overleftarrow{hd_a}]$, which is the final overview of the BiLSTM output. Global average pooling and global maximum pooling layers receive the BiLSTM layer's outputs concurrently. The BiLSTM layer's maximum and average values for each feature

are retrieved by the former and latter layers, respectively. Instead of using the dense layer(s), only one global pooling layer (each) is used. Before sending it to the last layer, the concatenate layer combines both the global maximum and global average layers into a single layer.

Detecting depression in social media text analysis using a hybrid RoBERTa-BiLSTM version includes leveraging the strengths of each RoBERTa and BiLSTM to capture deep contextual and sequential statistics from the textual content. RoBERTa, a transformer-based totally version, excels in know-how context via its pre-learned on a full-size corpus of statistics, which allows it to generate rich, nuanced embeddings for every token in the textual content. The technique starts with preprocessing the social media texts, which includes casting off punctuation and stopwords to make sure cleanser records input. These texts are then tokenized using RoBERTa's tokenizer, changing them right into a layout that the version can method. RoBERTa strategies these tokens to provide contextual embeddings, shooting the meanings of phrases in the context of surrounding phrases. This is important in social media text evaluation where slang, abbreviations, and casual language are typical.

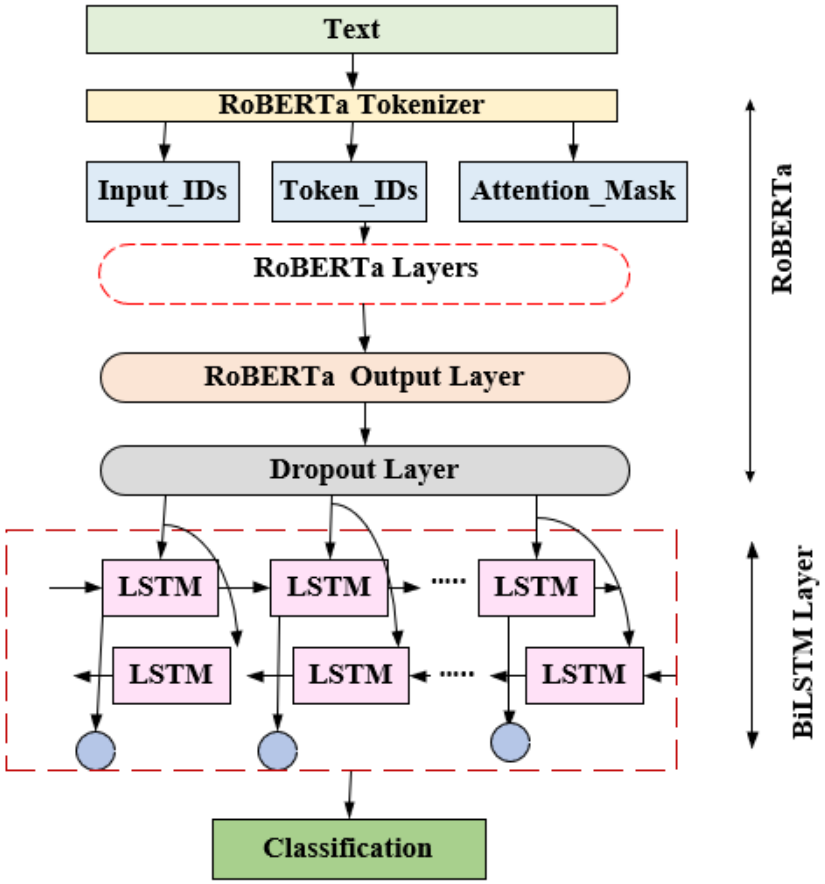


Fig. 2. RoBERTa and BiLSTM method.

The BiLSTM network receives the generated contextual embeddings directly. BiLSTM techniques the series of embeddings in both forward and backward directions, capturing dependencies that span across the entire text. This bidirectional processing is mainly beneficial for knowing the emotional tone and progression inside the textual content, which is important for identifying symptoms of depression. The hidden states generated via the BiLSTM encapsulate records from both beyond and future contexts, imparting a complete know-how of the collection. The model then applies an attention mechanism, which assigns specific weights to exceptional components of the textual content primarily based on their relevance to depression identification. The attention technique enables the version to concentrate on important textual elements which can be more indicative of despair, along with expressions of sadness, hopelessness, or tension. The weighted hidden states are combined to shape a context vector that represents the general emotional and contextual facts of the textual content. This context vector is then used to extract functions that can be fed into a chain of dense layers for category. The dense layers procedure these features to output a prediction indicating whether or not the text shows depressive tendencies. The model is educated on the usage of categorized facts, where the social media texts are annotated with labels indicating the presence or absence of depressive symptoms. During training, the model learns to optimize its parameters to accurately classify new, unseen texts. The pseudocode in Algorithm 1 and flowchart in Fig. 3 outline a technique for detecting depression in social media textual content using RoBERTa-BiLSTM, consisting of facts preprocessing, embedding generation, feature extraction, category, and final predictions.

Algorithm 1: Depression Detection Using RoBERTa-BiLSTM Hybrid Model

```
Input: Social media text samples
Output: Depression prediction for each text sample
Step 1: Data Preparation
Load text samples
Preprocess text:
  - Remove punctuation
  - Remove stopwords
Step 2: Tokenization and Embedding with RoBERTa
tokenized_texts = RoBERTa_tokenizer(preprocessed_texts)
embeddings = RoBERTa_model(tokenized_texts)
Step 3: Hidden State Calculation with BiLSTM
hidden_states = []
for emb in embeddings:
  forward_hidden_state, backward_hidden_state = BiLSTM (emb,
previous_forward_hidden_state,
previous_backward_hidden_state)
  hidden_states. append ((forward_hidden_state,
backward_hidden_state))
Step 4: Feature Extraction
features = extract_features(context_vector)
Step 5: Classification using Dense Layers
depression_prediction = Dense_Layers(features)
Step 6: Training and Evaluation
Train the model using labelled depression data
Evaluate model performance using metrics (accuracy, F1 score,
etc.)
Step 7: Prediction
depression_predictions= predict_depression(new_text_samples)
Output: depression_predictions
End
```

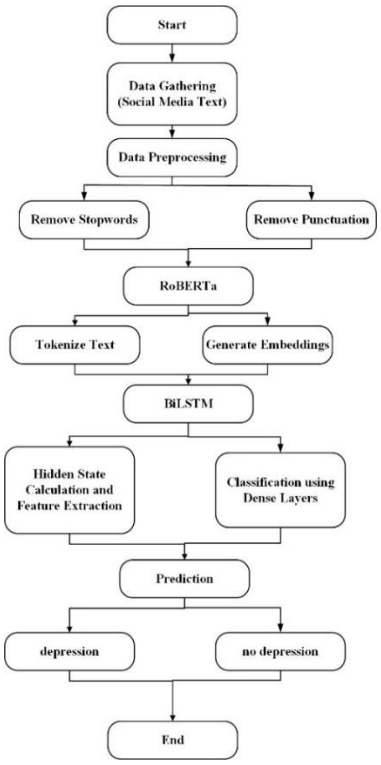


Fig. 3. Flowchart for proposed depression identification method.

V. Results and Discussion

The proposed study using the RoBERTa-BiLSTM hybrid model showed promising results in detecting depression in social media contexts. The model fully processed and analyzed transcript content, using the RoBERTa contextual input and the BiLSTM two-way sequential learning to classify transcripts as "depressed" or "no depressed." indicators Analytical parameters are appeared to perform better than baseline models. The combination of RoBERTa and BiLSTM led to a better understanding of sensory signals and context, resulting in stronger and more reliable predictions, and highlighted the real-world potential of the model in mental health services.

A. Experimental Outcomes

TABLE III. Depression Prediction

text	label	prediction
Hi, I want to tell you... Lately I've been feeling restless, have trouble sleeping, I searched on google it says it's a mild symptom of depression, I used to tell my mom a psychologist friend "don't think too much, it's not important you get depressed easily" then I frequent irregular breathing.	1	0.164200
Yes, the point is that I feel tired, sad, annoyed, restless. It's like the feeling is mixed in my heart and mind, in my brain I'm traveling various things from problems to happiness that I've felt until this moment.	1	0.164194
Every time after sunset, why must this heart be restless as if it can't accept the situation. But with this situation, you can't do anything, if you do it, it can only make things worse	1	0.164191
Anjir looks like he can't watch Dream, I have an appointment at 3 o'clock ””	0	0.095998
w if it doesn't look like it's okay. it's too late for the child w. it's better if it's hard to be alone, so let's be like before	0	0.095998
even though it's just a vanilla latte but why is it strong until the morning, after that it looks like a panda	0	0.095998

The Table III presents the results of a depression detection model using RoBERTa-BiLSTM on social media text samples. The "Text" column contains input text samples, "Label" indicates ground truth (1 for depression, 0 for depression), and

"Prediction" shows the model's confidence score for depression Higher scores (closer to 1) indicate greater confidence in detecting depression. The model demonstrates the ability to distinguish between depressive and no depressive texts, and obtain higher scores for depressive texts.

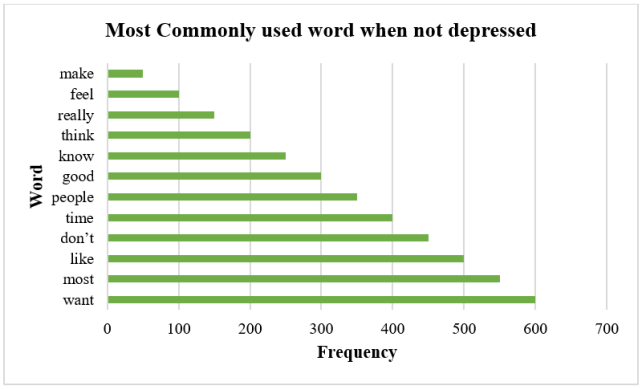


Fig. 4. Not depressed words.

The Fig. 4 indicates the frequency of different words humans use whilst they're no longer feeling depressed. The words are listed from maximum to least frequent: "make," "feel," "certainly," "assume," "know," "humans," "proper," "time," "like," "do not like most," and "want." The maximum frequency is just under seven hundred for "make," and the bottom is around 100 for "need." This figure highlights how regularly those words appear in conversations whilst individuals are not experiencing depression.

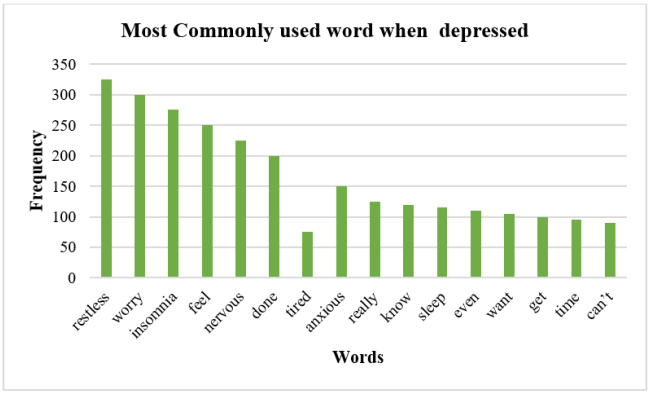


Fig. 5. Depressed words.

The Fig. 5 shows the frequency of particular words used by individuals when feeling depressed. The horizontal axis records terms such as "restlessness," "insomnia," and "fatigue," while the vertical axis measures frequency from 0 to 350 and "restlessness" corresponds to "insomnia," indicating that these terms are mentioned more frequently In the case of depression. This visual representation aids to understand common language patterns associated with depressive states, which can be useful for psychoanalytic or linguistic studies.

B. Training and Testing

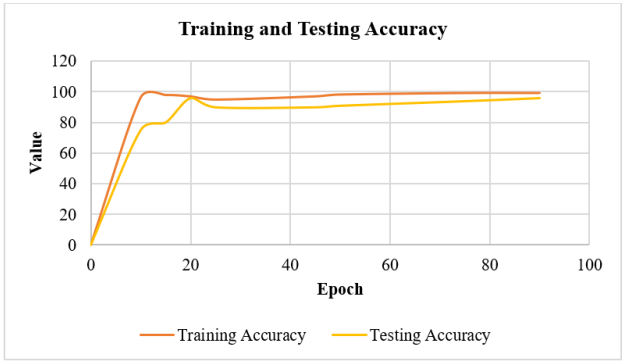


Fig. 6. Training and testing accuracy.

Training and test accuracy measures at different ages demonstrate the learning and generalization capabilities of the model is shown in Fig. 6. Initially, both accuracies start from 0. By time 10, the accuracy of training reaches 96%, the accuracy of testing reaches 75%, indicating initial overfitting and both accuracies improve as training goes on so, with training accuracy standing around 97-99% and testing accuracy of 96 at time 90 peaks at %. This indicates that the model learns the training data well and generalizes well to unseen data, thus confirming the robustness of RoBERTa-BiLSTM hybrid model emphasizes uneasiness found in social media text.

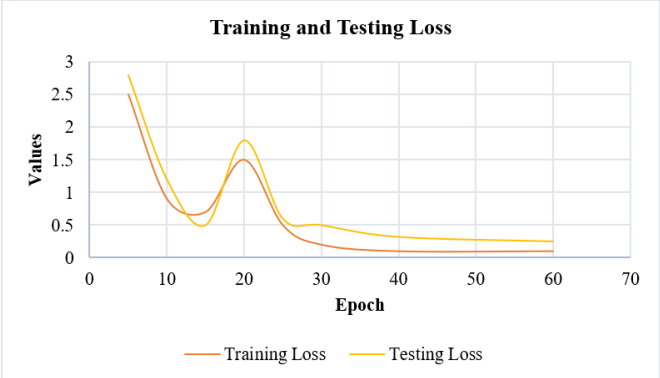


Fig. 7. Training and testing loss.

The training and testing loss show how the error in the model decreases with time is shown in Fig. 7. Initially, at epoch 5, the mean training loss was 2.5 and the mean testing loss was 2.8. By time 10, the loss decreases significantly 0.9 for training and 1.2 for testing, indicating an improvement in model learning. Despite a slight increase at time 20, the loss continues to decrease, with the training loss reaching 0.1 and the testing loss reaching 0.25 at time 60. This steady decrease in loss indicates the accuracy of the model improved and decreased errors, a finding indicative of both effective and general learning for depression identification.

C. Performance Evaluation

When compared to baseline models, the effectiveness evaluation of the suggested hybrid framework for depression detection produced consistent and outstanding results. The

equation in (9), (10), (11) and (12) is used to find the F1-score, recall, accuracy, and precision. T_{pos} means true positive, T_{neg} means true negative, F_{pos} means false positive and F_{neg} means false negative.

$$Accuracy = \frac{T_{pos}+T_{neg}}{T_{pos}+T_{neg}+F_{pos}+F_{neg}} \quad (9)$$

$$Precision = \frac{Tpos}{Tpos+Fpos} \quad (10)$$

$$Recall = \frac{Tpos}{Tpos+Fneg} \quad (11)$$

$$F1\ Score = \frac{2\times precision \times recall}{precision \times recall} \quad (12)$$

TABLE IV. PERFORMANCE COMPARISON

Approach	Accuracy	Precision	Recall	F1-Score
CNN[9]	0.884	0.903	0.87	0.936
CNN AE+SVM [10]	0.71	0.72	0.71	0.71
MDHAN [11]	0.89	0.9	0.89	0.89
BERT [14]	0.87	0.73	0.96	0.81

The Table IV compares the performance of different models in predicting depression. The corresponding kernel support vector machine achieves an accuracy of 83.1% with precision, recall, and F1 scores of around 80-83%. The RNN model performs slightly lower in accuracy 80.5% but better in F1 score 83.8%. The CNN and BiLSTM models significantly improve the accuracy by 98.1% with higher accuracy, recall, and F1 scores. The proposed RoBERTa and BiLSTM models perform the best among them in terms of 99.4% accuracy, good precision 98.5%, recall 97.1%, and F1 score 97.3%, indicating good performance in depression as they are seen in it.

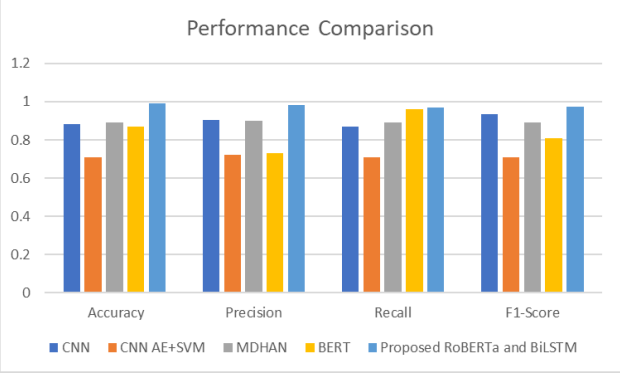


Fig. 8. Performance comparison of various methods.

The Fig. 8 compares four models: The performance comparison graph shows the comparative effectiveness of various models, viz., CNN, CNN AE+SVM, MDHAN, BERT, and the suggested RoBERTa-BiLSTM model, on critical evaluation metrics: accuracy, precision, recall, and F1-score. The suggested RoBERTa-BiLSTM model surpasses all the other models in all metrics, proving to be the best depression detection capability from social media text. Though CNN and CNN AE+SVM exhibit average performance, CNN AE+SVM falls behind because it has less contextual knowledge. MDHAN and BERT are good, but the integration of RoBERTa's contextual embeddings with BiLSTM's sequential learning improves recall and F1-score, and thus the proposed approach is the best.

D. Discussion

The RoBERTa-BiLSTM hybrid model significantly enhances the detection of depression in social media contexts by addressing various limitations of traditional methods. Traditional methods typically involve laborious and subjective manual assessments and clinical interviews. Conventional methods often suffer from limited accuracy and scalability issues [21]. This model uses advanced NLP and deep learning, automates the search process, and provides intuitive and objective solutions. The introduced RoBERTa-BiLSTM model shows considerable strengths compared to the conventional approach by efficiently extracting both contextual and sequential dependencies from social media text. It is superior to common machine learning models such as SVM, RF, and earlier deep learning models including CNN-LSTM. The inclusion of RoBERTa in context captures subtle emotional expressions, while BiLSTM better models' sequential speech dependence, improving sensitivity to depressive symptoms. The results show a notable accuracy of 99.4%, with high accuracy (98.5%), recall (97.1%) and F1 score (97.3%). This implementation demonstrates the model's resilience in differentiating between texts that are sad and those that are not, outperforming both alternative DL techniques and conventional machine learning models. Through guaranteeing accurate and suitable input, preliminary processes that involve the removal of pauses and character input help to develop the model. The accomplishment of this hybrid approach suggests that similar models can be incorporated into mental health care systems, enabling early detection and intervention Future research could extend the model to address or deliver multilingual content other methods such as visual or auditory data will be combined to further enhance the prediction capabilities.

VI. CONCLUSION AND FUTURE SCOPE

For healthcare departments to assist their depressed patients, it is imperative that depression be automatically identified from text. The RoBERTa-BiLSTM hybrid model leads to significant improvements in depression detection in social media contexts. This study achieved 99.4% accuracy, demonstrating the exceptional performance of the model RoBERTa's advanced contextual embeddings with serial capture capabilities dependencies of the BiLSTM Integrating, the model correctly recognizes subtle emotional cues that predict depressive states preliminary steps, including stopword removal and syntactic removal, provide the model's ability to process and analyze contextual information to ensure predictions are accurate and reasonable. The drawbacks of conventional techniques, which usually rely on biased manual analysis, are eliminated by this strategy. Through automating the process of discovery, this model provides a scalable, objective solution that can be incorporated into mental health care systems to facilitate the operation of the early intervention RoBERTa-BiLSTM hybrid model high quality underscores the potential for real-world application in mental health. Subsequent investigations ought to concentrate on several crucial domains to augment the

dependability of the framework. Extending the model to handle multilingual contexts could improve its applicability to different language populations. Furthermore, the integration of multiple inputs, such as visual or auditory feedback, leads to a more comprehensive understanding of depressive symptoms, as emotional symptoms are often expressed through multiple mechanisms. The RoBERTa-BiLSTM model has difficulty with non-verbal signals, which restricts accuracy of depression detection. It is challenged in multilingual environments and potentially drops contextually significant words in preprocessing. The need for high computational requirements discourages real-time use, while risks of misclassifications involve sarcasm and doubtful text, necessitating further tuning for enhanced reliability and explainability. Future research encompasses multilingual adaptation, multimodal data integration (text, audio, images), real-time deployment optimization, simplification of computational complexity, improved explainability, fine-tuning sarcasm detection, contextual understanding improvement, and guaranteeing ethical AI deployment in mental health care.

REFERENCES

[1] "Depressive disorder (depression)." Accessed: Aug. 05, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[2] A. Thapar, O. Eyre, V. Patel, and D. Brent, "Depression in young people," The Lancet, vol. 400, no. 10352, pp. 617–631, 2022.

[3] M. A. Villarroel and E. P. Terlizzi, "Symptoms of depression among adults: United States, 2019," 2020.

[4] C. B. Nemeroff, "The state of our understanding of the pathophysiology and optimal treatment of depression: glass half full or half empty?," Am. J. Psychiatry, vol. 177, no. 8, pp. 671–685, 2020.

[5] E. A. Pataky and U. Ehlert, "Longitudinal assessment of symptoms of postpartum mood disorder in women with and without a history of depression," Arch. Womens Ment. Health, vol. 23, no. 3, pp. 391–399, 2020.

[6] J. M. Havigerová, J. Haviger, D. Kučera, and P. Hoffmannová, "Text-based detection of the risk of depression," Front. Psychol., vol. 10, p. 513, 2019.

[7] L. Squarcina, F. M. Villa, M. Nobile, E. Grisan, and P. Brambilla, "Deep learning for the prediction of treatment response in depression," J. Affect. Disord., vol. 281, pp. 618–622, 2021.

[8] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: a narrative review," NPJ Digit. Med., vol. 5, no. 1, pp. 1–13, 2022.

[9] C. Lin et al., "Sensemood: depression detection on social media," in Proceedings of the 2020 international conference on multimedia retrieval, 2020, pp. 407–411.

[10] S. Sardari, B. Nakisa, M. N. Rastgoo, and P. Eklund, "Audio based depression detection using Convolutional Autoencoder," Expert Syst. Appl., vol. 189, p. 116076, 2022.

[11] H. Zogan, I. Razzak, X. Wang, S. Jameel, and G. Xu, "Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media," World Wide Web, vol. 25, no. 1, pp. 281–304, 2022.

[12] J. S. L. Figuerêdo, A. L. L. Maia, and R. T. Calumby, "Early depression detection in social media based on deep learning and underlying emotions," Online Soc. Netw. Media, vol. 31, p. 100225, 2022.

[13] M. Niu, J. Tao, B. Liu, J. Huang, and Z. Lian, "Multimodal spatiotemporal representation for automatic depression level detection," IEEE Trans. Affect. Comput., vol. 14, no. 1, pp. 294–307, 2020.

[14] E. A. Ríssola, S. A. Bahrainian, and F. Crestani, "A dataset for research on depression in social media," in Proceedings of the 28th ACM conference on user modeling, adaptation and personalization, 2020, pp. 338–342.

[15] X. Chen, M. Sykora, T. Jackson, S. Elayan, and F. Munir, "Tweeting your mental health: An exploration of different classifiers and features with emotional signals in identifying mental health conditions," 2018.

[16] M. O. Nusrat, W. Shahzad, and S. A. Jamal, "Multi Class Depression Detection Through Tweets using Artificial Intelligence," ArXiv Prepr. ArXiv240413104, 2024.

[17] S. Saha, "Students anxiety and depression dataset." Accessed: Aug. 06, 2024. [Online]. Available: https://www.kaggle.com/datasets/sahasourav17/students-anxiety-and-depression-dataset

[18] C.-T. Wu, D. G. Dillon, H.-C. Hsu, S. Huang, E. Barrick, and Y.-H. Liu, "Depression detection using relative EEG power induced by emotionally positive images and a conformal kernel support vector machine," Appl. Sci., vol. 8, no. 8, p. 1244, 2018.

[19] A. H. Uddin, D. Bapery, and A. S. M. Arif, "Depression analysis of bangla social media data using gated recurrent neural network," in 2019 1st International conference on advances in science, engineering and robotics technology (ICASERT), IEEE, 2019, pp. 1–6.

[20] N. Marriwala, D. Chaudhary, and others, "A hybrid model for depression detection using deep learning," Meas. Sens., vol. 25, p. 100587, 2023.

[21] D. Liu, X. L. Feng, F. Ahmed, M. Shahid, J. Guo, and others, "Detecting and measuring depression on social media using a machine learning approach: systematic review," JMIR Ment. Health, vol. 9, no. 3, p. e27244, 2022.

Detecting Chinese Sexism Text in Social Media Using Hybrid Deep Learning Model with Sarcasm Masking

Lei Wang[1], Nur Atiqah Sia Abdullah[2], Syaripah Ruzaini Syed Aris[3]
College of Information Engineering & Computer Science, Hebei Finance University, Baoding, Hebei, China[1]
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia[1, 2, 3]
Knowledge and Software Engineering Research Group (KASERG), Research Nexus UiTM (ReNeU), Office of Deputy Vice Chancellor (Research and Innovation), Universiti Teknologi MARA, Shah Alam, Malaysia[2]
Hebei Provincial Key Laboratory of Financial Technology Application, Baoding, Hebei, China[1]

***Abstract*—Sexist content is prevalent in social media, which seriously affects the online environment and occasionally leads to offline disputes. For this reason, many scholars have researched how to automatically detect sexist content in social media. However, the presence of sarcasm complicates this task. Thus, recognizing sarcasm to improve the accuracy of sexism detection has become a crucial research focus. In this study, we adopt a deep learning approach by combining a sexism lexicon and a sarcasm lexicon to work on the detection of Chinese sexist content in social media. We innovatively propose a sarcasm-based masking mechanism, which achieves an accuracy of 82.65% and a macro F1 score of 80.49% on the Sina Weibo Sexism Review (SWSR) dataset, significantly outperforming the baseline model by 2.05% and 2.89%, respectively. This study combines the irony masking mechanism with sexism detection, and the experimental results demonstrate the effectiveness of the deep learning method based on the irony masking mechanism in Chinese sexism detection.**

***Keywords*—Sexism; chinese; deep learning; sarcasm; masking**

I. Introduction

There is a limitation on the current sexism detection. It is difficult to detect sexist text with a sarcastic sense [1]–[3]. Sarcasm uses irony to mock or express contempt, and there is a discrepancy between an utterance's literal meaning and intended meaning. It is challenging to detect automatically [4]. Sarcasm in long and short text remains a challenge in Natural Language Processing (NLP) and Sentiment Analysis (SA) [5]. In the research of [3], irony limits the performance of sexism classification. According to [1], numerous tweets are labeled under sexism categories due to their sarcastic meaning. However, their model is unable to detect the sarcastic intent. The researchers of [2] also argued that sexist posts with humor, irony, and sarcasm are challenging to spot and often contain no overt expressions of hatred. Therefore, it is necessary to identify sarcastic sexism in the detection of sexism.

For example, 如果她自己够优秀就不会在网络上怨天尤人了 (If she is excellent enough, she will not blame others on the internet) is misclassified sexism [2], which contains a sarcastic sense. It is a sarcastic remark that blames unsuccessful women while insulting others who support gender equality. When there is no explicit presence of harsh language, it might be challenging to recognize sexism in an automatic system.

In the research of [3][5], they discovered that sarcasm limits the performance of the sexism classification. In the research of [6], they reported that sarcasm knowledge is helpful for Spanish sexism classification. The researchers of [7] researched Arabic sexism and sarcasm text classification using deep learning methods and noted that AraBERT performs best for both sexism detection and sarcasm detection. It illustrates that AraBERT is useful for both sexism and sarcasm detection. However, the accuracy of sexism detection is 91.0%, higher than sarcasm detection, which is 88%. It can be concluded that the deep learning method is useful for both sexism and sarcasm detection, and the sarcasm knowledge is useful for improving the performance of sexism detection.

Various deep learning methods, such as Convolutional Neural Networks (CNN) [8], Long Short-Term Memory (LSTM) [9], Bidirectional Long Short-Term Memory (BiLSTM) [10], Bidirectional Gated Recurrent Unit (Bi-GRU) [11], Bidirectional Encoder Representations from Transformers (BERT) [12]-[14], Robustly Optimized BERT Pretraining Approach (RoBERTa) [2], Multilingual BERT (mBERT) [15], have been employed in sexism detection. In the research of [1][8][11][16], deep learning methods outperform traditional machine learning methods, such as Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayes (NB), Decision Tree (DT). Among the deep learning methods, transformers-based deep learning like BERT and RoBERTa usually perform the best [2][3][12][17]. The hybrid approach in the research [10][18]-[21] obtains the best result. Accordingly, deep learning methods, especially transformer-based methods and hybrid approaches, are considered in our research.

The researchers in [6] demonstrated that sarcasm aids in the identification of sexism. They adopted a multi-task learning strategy that simultaneously addresses numerous problems rather than treating them in isolation to enhance Spanish sexism detection. Nonetheless, the strategy is inappropriate for a singular-task situation. The research in [7] demonstrated that deep learning methods are good at detecting both sexism and sarcasm in Arabic. However, they employed the same deep learning model on both the misogyny dataset and the sarcasm

This work was funded by the Science Research Project of Hebei Education Department (QN2024149).

dataset to demonstrate the efficacy of their technology. Consequently, there is a lack of study on sarcasm-based sexism detection in a singular task context, while it is considered a prospective research avenue by [1]–[3][22].

This research aims to improve Chinese sexism detection using a deep learning approach fused with sarcasm. The dataset is the Sina Weibo Sexism Review (SWSR). To leverage sarcasm features in the model, sarcasm-detecting approaches are invested. There are text-only methods and extra information methods for sarcasm detection [23]. In particular, the text-only method concentrates on the utterance itself to detect the sarcasm text, such as exploiting linguistic features [24] and using inconsistent expressions [26]. This research uses sarcastic linguistic features in a sexist lexicon with a masking mechanism. Consequently, it combines them with a deep learning approach. The contributions of this paper are as follows:

1) Proposed Sarcasm-sexism-Nezha-mCNNs model. The model combines a sarcasm lexicon and a sexism lexicon with a transformer-based Nezha and three CNNs. The model's performance in detecting Chinese sexist text is significantly enhanced, achieving an accuracy improvement of 2.05% and a macro F1 score increase of 2.89% compared to the baseline.

2) Constructed a Chinese sarcasm lexicon. A Chinese sarcasm feature lexicon, consisting of 193 words, has been developed. This lexicon combines terms identified in previous studies with words generated through the chi-square test in this paper. It has potential applications in further sarcasm detection endeavors.

3) A sarcasm mask mechanism is proposed. The output from the sarcasm layer is masked with 1 or -1, and then the masked output is multiplied by the output from the sexism lexicon layer as 20% for the final output. This crucial design feature for the proposed model can be adapted for gender-based categorization of sexism in both Chinese and other languages.

The remainder of the paper is organized as follows: Section II reviews related works. Section III outlines the research methodology, including data processing, text representation, classification, evaluation, and experimental settings. Section IV presents the experimental results, while Section V interprets these outcomes and highlights the key findings. Finally, Section VI summarizes our contributions, acknowledges limitations, and suggests directions for future research.

II. Related Work

First, hybrid deep learning approaches for sexism detection are investigated. Next, the detection of sarcasm using Chinese sarcasm linguistic features is explored. The Chinese sarcasm dataset and feature extraction methods are also presented.

A. Hybrid Deep Learning Approach for Sexism Detection

1) Combining more than one deep learning method: In the research of [18], they employed a CNN-LSTM model to detect sexual harassment in Hindi. They substituted all newline characters with a space, eliminated English letters, numerals, hyperlinks, emojis, and special characters, and employed a tokenizer with the Keras library. An accuracy of 93.53% is achieved with CNN-LSTM, surpassing RNN-LSTM.

The approach in a study by [19] utilized a BiLSTM combined with a Temporal Convolutional Network (TCN-BiLSTM) to analyze sexist speech in Arabic social media, surpassing BiLSTM, TCN, and XGBoost. This method benefits languages that provide morphological challenges, such as Arabic, Turkish, and Lithuanian.

The researchers in [20] combined customBERT with a CNN to detect sexism in social media on the dataset of SemEval Task 10. It combines output from BERT, XLM-RoBERTa, and DistilBERT and then fed into CNN. 'Random deletion' improves the model's ability to understand incomplete information. 'Synonym replacement' improves the model's understanding of context. Then, back translation is used for data augmentation. Shapley additive explanation values are used to increase model explainability.

In the research of [22], they employed a semi-supervised model to enhance the EXIST dataset through data augmentation. The pre-trained XLM-R and sentence BERT combined with BiLSTM are utilized for sexism detection, achieving an accuracy of 81.2% and an F1 of 81.1%. They indicated a prospective research direction involving the integration of sarcasm and irony detection into their system.

It illustrates that CNN, LSTM, BiLSTM, and transformer-based models such as BERT, mBERT, and XLM-R are commonly used in combinations of two or more deep learning methods.

2) Combining deep learning with linguistic features: The researchers in [10] combined lexicon and sentiment features with LSTM, obtaining the best accuracy of 87.2% and an F1 of 82.4% on the dataset of AMI EN-EVALITA. They discovered that the optimal feature type identified is Term Frequency-Inverse Document Frequency (TF-IDF) Ngrams. However, lexical features and word2vec embeddings can enhance the outcomes when integrated with TF-IDF. Their model outperforms BERT, XLNET, RoBERTa, and DistilBERT.

In the research of [2], they combined BERT, RoBERTa, CNN, SVM, and LR with the sexism lexicon, respectively. The sexism lexicon is expressed using TF-IDF. The researchers find that RoBERTa, with a sexism lexicon, obtains the best F1 score of 78% on the SWSR dataset. However, it has a lower accuracy value than BERT without a lexicon. It indicates that the sexism lexicon should be combined with the deep learning method with a proper method.

The researchers in [21] integrated LSTM with sentiment analysis, obtaining the best performance detecting sexism with an F1 of 83.01%, which outperforms LSTM-sexism and LSTM-RoBERTa. Thus, pre-trained models are used for word representation.

It can be noted that linguistic features such as sexism and sentiment have been combined into deep learning methods

such as LSTM, BERT, and RoBERTa. However, a proper combining method is required.

3) Sarcasm detection with Chinese sarcasm features. Sarcasm uses irony to mock or express contempt, and there is a discrepancy between the literal meaning and intended meaning of an utterance. It is challenging to detect automatically [4][25]. According to [26], there are text-only methods and extra information methods. The text-only method concentrates on the utterance itself to detect the sarcasm text, such as exploiting linguistic features [27]-[29] and using inconsistent expressions [23][30][31]. In contrast, the extra information method focuses on external knowledge, such as user features and common sense knowledge. In this research, we focus on sarcasm linguistic features.

In the studies in [24][27]-[29][32]-[34], linguistic features of Chinese sarcasm are employed as significant features for the detection of Chinese sarcasm. The chi-square test is employed to pick feature words more pertinent to the target category and obtain Chinese sarcastic feature words. The traits are categorized into more specific classifications, including interjection words, adverbs of degree, internet vocabulary, homophonic words, and punctuation. The Chinese sarcasm feature words presented in the studies referenced as [24][27][28][32][34] are integrated. Table I illustrates some examples of Chinese sarcastic linguistic feature words.

TABLE I. CHINESE SARCASM LINGUISTIC FEATURE WORDS

Word Type	Words
Interjection words	呵呵(hehe), 啊(ah), 哇(wow), 哟(yo), 哈哈(haha), 嘿嘿(heihei), 唉(sigh), 哼(hmm）
Adverbs of degree	真是(really), 非常(very), 极其(extremely), 牛(awesome), 牛逼(awesome), 不愧是(worthy of being）
Internet vocabulary	醉了(drunk), 凡尔赛(Versailles), 躺平(lie flat), 涨姿势(gaining knowledge), 囧(confused), 逗比(funny person),
Homophonic words	河蟹-和谐(harmony), 杯具-悲剧(tragedy), 灰常-非常(very), 餐具-惨剧(tragedy), 表-婊(bitch), 虾米-什么(what),
Punctuation	？，！，“”, ……，。。。

1) Interjection words: The inclusion of interjection words in a sentence imparts a sense of teasing and irony [28]. In sentence 哇？？？？？？？吐槽鬼居然是女性我真的觉得她可爱了哈哈哈哈哈 (Wow??????? The complaining ghost is actually a woman. I really think she is cute, hahahahahaha), both 哇 (Wow) and 哈哈哈哈哈 (hahahahahaha) are interjection words.

2) Adverbs of degree: Adverbs of degree will enhance the semantic intensity of the words in the text, illustrating an exaggerated effect and presenting irony [27]. In sentence 不过我经常碰到异常自信，自以为是的男性，也让人吃惊。(But it's also surprising how often I run into unusually confident, self-righteous males.) 异常(exceptional) expresses exaggeration, and it is ironic to somebody.

3) Internet vocabulary: Internet vocabularies are widely used on the internet, which may be helpful for sarcasm detection, such as 醉了(drunk), which expresses an attitude of finding someone's actions or something unexplainable

4) Homophonic words: Most homophonic words are deformed words with similar pronunciation to the original words or new words created on the internet to incorporate fun and humor or to express a specific mood [28]. Somebody may humorously write 杯具 (cupcake) instead of 悲剧 (tragedy) since they share the same pronunciation.

5) Punctuation: According to [35], some special punctuation can emphasize the effect of irony in the proper context and help readers understand the author's intent. Punctuation marks such as question marks, exclamation points, quotation marks, and apostrophes could emphasize irony [27]. If these symbols are used more than three times consecutively, they convey further emphasis and enhance irony [36].

B. Chinese Sarcasm Datasets

To extract more sarcasm linguistic feature words, a collection of Chinese sarcasm datasets is conducted. Table II illustrates the Chinese sarcasm datasets. All datasets are about the classification of sarcasm except for the first dataset, SMP ECISA, and the last dataset, Weibo Sentiment. The first dataset, SMP ECISA, focuses on implicit sentiment analysis. It contained implicit positive sentiment, implicit negative sentiment and no sentiment. The rationale for including the final dataset is that, although it pertains to sentiment analysis, the author regards sarcastic semantic recognition as a significant aspect of the research.

TABLE II. CHINESE SARCASM DATASET

Name	Size	Used by	Model	Performance
SMP ECISA	Neutral: 10012 Positive: 5925 Negative: 5973	[37]	KG-MPOA	Macro F1 0.777
		[38]	BERT-BiLSTM	Macro F1 0.775
Ciron	Not ironic: 4343 Unlikely ironic: 3391 Insufficient evidence: 64 Weakly ironic: 838 Strongly ironic: 130	[39]	BERT	F1: 0.572 A: 0.603
Ciron + Chinese Sarcasm	Not ironic: 5343 Ironic: 5425	[27]	IDIHR	F1:0.8124 A: 0.8149
ToSarcasm	Not ironic: 2435 Ironic: 2436	[40]	TOSPrompt	F1:0.7320 A: 0.7176
Weibo Sarcasm	Not ironic: 2000 Ironic: 2000	[24]	ISR	F1: 0.7793
		[33]	NB	A: 0.6945
Multimodal Sarcasm	Not ironic: 1179 Ironic: 1009	[41]	FCAM	F1: 0.8778 A: 0.8813

C. Chi-Square Test for Sarcasm Linguistic Feature Extracting

The chi-square test can be used to extract sarcastic linguistic features that are closely related to sarcasm. The chi-square test posited that the characteristics and categories operate independently, subsequently assessing the correlation through deviation analysis. A low chi-square test value indicated that the correlation between the two variables might be coincidental and lacked significance. Conversely, a high chi-square test value suggested a stronger correlation that could be utilized as a categorical feature. Consequently, the chi-

square test stood out as the efficient and suitable approach for feature selection [24].

III. Research Methodology

This section will introduce data description and preprocessing, text representation, text classification, evaluation of sexism detection, and experiment settings.

A. Data Description and Preprocessing

The dataset utilized in this study is a Chinese sexism dataset named Sina Weibo Sexism Review (SWSR). It was gathered from the Sina Weibo platform between June 2015 and June 2020 by [2]. The comments are labeled as sexism or not sexism. Among the 8,969 initial comment texts, there are six instances of duplication in contents but with different user information. In this research, user information is not used. Hence, the duplicated comments are removed, resulting in 8,963 comments.

To eliminate the noise in the text, data preprocessing is conducted. The original dataset was converted from traditional Chinese to simple Chinese by [2]. We conducted the following processes: converting full-corner text content to half-corner, converting punctuation symbols from English to Chinese, and converting uppercase English to lowercase English.

B. Text Representation

Three text representations are conducted in the final model.

1) Pre-trained Nezha tokenizer: Nezha is a transformer-based pre-trained model, like BERT and RoBERTa. It has been trained on large-scale Chinese corpus and has obtained outstanding performance in many Chinese NLP tasks [42]. Then, we expressed the comments with the pre-trained Nezha tokenizer. In this study, the sentence length is set to 150. Sentence lengths exceeding 150 are truncated, and sentences less than 150 are filled with [PAD]. After the Nezha tokenizer, the sentence is expressed with input ids, token type ids, and attention masks.

2) Text representation with sexist lexicon: A sexism lexicon with 3,016 words created by [2] is utilized for comments representation using Bag of Word Vector (BoWV) on the sexism lexicon. The frequency of each word in the sexism lexicon for each sentence phrase is calculated to a vector as a linguistic attribute.

3) Text representation with sarcasm lexicon: Firstly, the sarcasm lexicon is constructed. To construct the sarcasm lexicon, the sarcasm-related feature words are collected. Irony-related feature words are obtained by the chi-square test. The chi-square test is conducted on the SMP ECISA dataset which is for evaluating Chinese implicit sentiment analysis. Non-implicit content and negative implicit content are used, while positive implicit content is excluded. The sarcastic feature words identified using the chi-square test are subsequently picked manually, obtaining 85 words. The same operation is conducted on the Ciron dataset. The content is relabelled before the chi-square test. After relabeling the content, items labeled 1 and 2 are classified as non-sarcasm, while those labeled 4 and 5 are categorized as sarcasm. Meanwhile, entries labelled with 3 are disregarded because of insufficient evidence of sarcasm. Subsequent to the chi-square test and manual selection, 28 words remained as feature words. Accordingly, no prominent sarcasm features are discovered in the ToSarcasm and BSC datasets using the chi-square test. The words selected from SMP ECISA 2021 and Ciron are combined with the words selected from other related research, and 193 sarcastic linguistic feature words are finally obtained.

Then, the sarcastic lexicon with 193 words is utilized for comment representation using BoWV. Following text representation approach with sexism lexicon, the frequency of each word in the sarcasm lexicon for each comment is calculated as a vector to represent sarcasm linguistic features.

C. Text Classification

To enhance the classification accuracy of sexism detection, sarcastic features and sexism features are considered to be combined with the deep learning method. Fig. 1 illustrates the architecture of the Sarcasm-sexism-CNNs-Nezha model. This model has four layers: mCNNs-Nezha layer, sexism lexicon layer, sarcasm lexicon layer, and output layer. The output of the sexism layer is multiplied by the output of the sarcasm lexicon layer, resulting in a combined feature output that contributes 20% to the final output. In contrast, the output from the mCNNs-Nezha layer accounts for 80% of the final output. Upon applying SoftMax to the result, the text is categorized as either sexist or non-sexist.

1) mCNNs-Nezha layer: The mCNN-Nezha layer accounts for 80% of the final result. It is a deep learning-based model consisting of the Nezha layer, transposition layer, mCNNs layer, and linear layer.

a) Nezha layer: Through the Nezha tokenizer, it can obtain input ids, token type ids and masks. As a pre-trained transformer-based model, Nezha contains the Nezha embeddings layer, Nezha encoder layer, and Nezha pooler layer. There are 12 hidden layers in the Nezha encoder layer. The last hidden layer is used as input for the transposition layer. The whole information, both the [CLS] information, which contained the whole information of the sentence and word embedding for characters in the sentence, is used. The data size is [b, l, h]. b is the batch size, l is the length of sentences, and h is the hidden size in the hidden layer. In our research, it is [32, 150, 768].

b) Transposition layer: A transposition is conducted on the dimensions of 2 and 3 in (1) and obtains data with the size of [b, h, l], which is [32, 768, 150] in this research.

$$D' = D^{T_{2,3}} \quad (1)$$

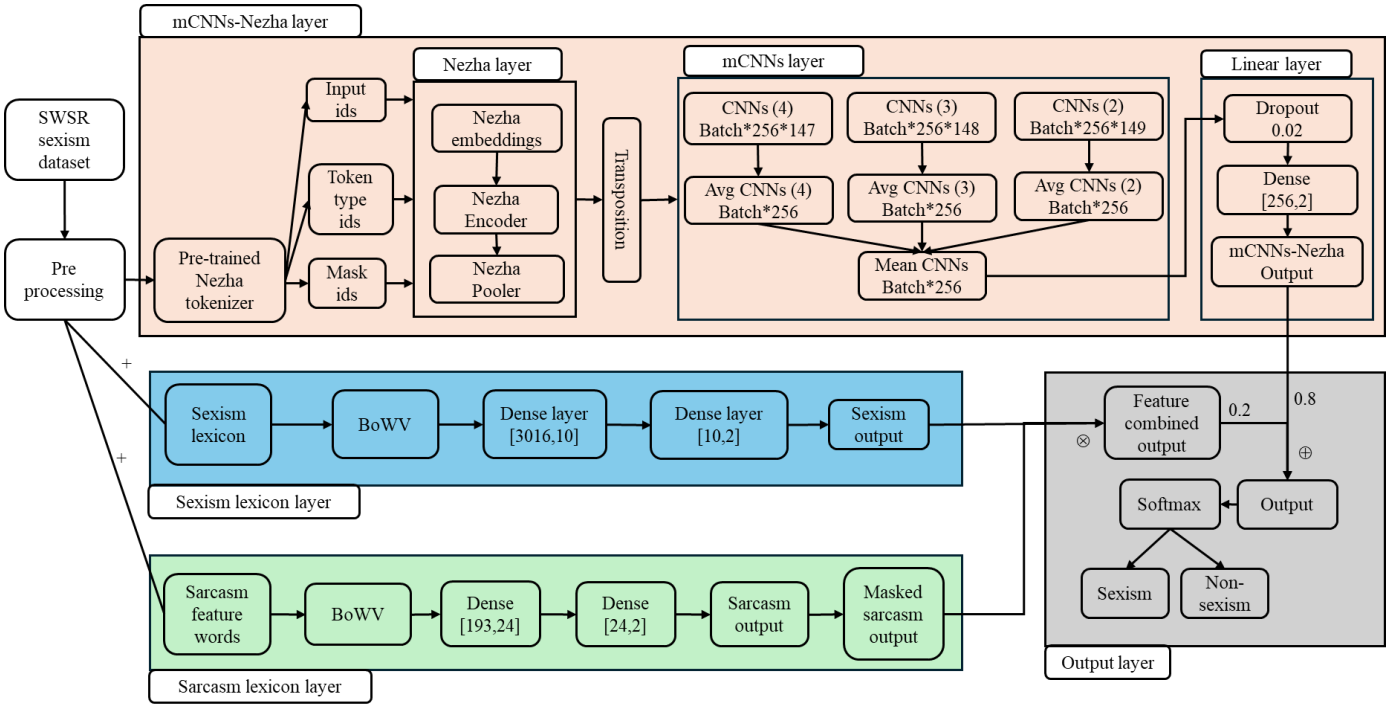


Fig. 1. The architecture of the sarcasm-sexism-mCNNs-Nezha model

c) mCNNs layer: The mCNNs layer contains three CNNs with kernels of 2, 3, and 4. After the CNN in (2), the data size is [b, o, l-k+1]. k is kernel size. o is the output channel with a value of 256. Hence, the feature is reduced.

$$CNN_k = CNN(D', k), \quad k=2,3,4 \tag{2}$$

To align features, an average operation is conducted on the third dimension in (3), and each CNN obtains data with the size of [b,o].

$$mCNN_k = Average(CNN_k^3), \quad k=2,3,4 \tag{3}$$

Then, an average operation is conducted on the three $mCNN_k$ on dimension 2 in (4). The final output of this layer is [b,o]. It can better capture words with two, three, or four characters, which is consistent with the Chinese word expressing a meaningful thing with two, three, or four characters.

$$mCNN = \frac{\sum_2^4 mCNN_k}{3} \tag{4}$$

d) Linear layer: The output from the mCNNs layer is fed into a linear layer. After a dropout of 0.02, a fully connected neural network is conducted and obtains an output with two values. The output of this layer is [b,2].

2) Sexism lexicon layer: The comments have been expressed as vectors using BoWV with the sexism lexicon. Then, the vector is fed into two densely connected neural networks in order in (5) (6) and results in an output with a two-dimensional vector. The first dense layer integrates the characteristics of sexism and executes feature mapping, encapsulating the intricate relationships among the features. The subsequent layer additionally conducts feature mapping and utilizes the extracted features to correlate with the probabilities of the two predictive classifications.

$$z^{(1)} = W^{(1)T} X_{sex} + b^{(1)} \tag{5}$$

$$O_{sex} = z^{(2)} = W^{(2)T} z^{(1)} + b^{(2)} \tag{6}$$

3) Sarcasm lexicon layer: The comments have also been expressed as vectors using BoWV with a sarcastic lexicon. Then, it is fed into two densely connected neural networks in order in (7) (8) and obtains a two-dimensional output. A masked operation is adopted on the output, where the first value is bigger than the second value and masked with -1, otherwise with 1, as indicated in (9). The masked output would multiply with the output of the sexism lexicon layer. This sarcasm masking mechanism has an important impact on the final output.

$$z^{(1)} = W^{(1)T} X_{sar} + b^{(1)} \tag{7}$$

$$O(l) = z^{(2)} = W^{(2)T} z^{(1)} + b^{(2)} \tag{8}$$

$$O_{sar} = \begin{cases} -1 & O(l)_0 > O(l)_1 \\ 1 & O(l)_0 \leq O(l)_1 \end{cases} \tag{9}$$

4) Output layer: The influence of the sexism lexicon layer's output O_{sex} on the final result, whether positive or negative, is dependent upon the sarcasm lexicon layer's output O_{sar}. If the value of O_{sar} is -1, it indicates a negative effect. Conversely, a value of O_{sar} equal to 1 signifies a positive effect, as presented in (10). The combined output, O_{lex}, constitutes 20%, while the output from O_{mn} mCNNs-Nezha layer comprises 80% of the output O_v in (11). A SoftMax is performed on O_v in (12).

$$O_{lex} = O_{sex} \otimes O_{sar} \tag{10}$$

$$O_v=0.8*O_m+0.2*O_{lex} \quad (11)$$

$$Let\ O(v)=(v_0,v_1), P(y=i)=SoftMax(v_i)=\frac{e^{v_i}}{\sum_{j=0}^{1} e^{v_j}},\ i=0,1 \quad (12)$$

The sarcasm masking mechanism algorithm is described as follows:

Algorithm 1: Sarcasm masking mechanism

```
Input: Output from mCNNs-Nezha layer (O1)
       Output from sexism lexicon layer (O2)
Output from sarcasm lexicon layer without masking (O3)
While (data in batch) do
    For (each comment) do
        If (O3[0] > O3[1]) then
            O2[0] = O2[0] * (-1)
            O2[1] = O2[1] * (-1)
        End
        O[0] = O1[0] * 0.8+O2[0] * 0.2
        O[1] = O1[1] * 0.8+O2[1] * 0.2
    End
```

D. Evaluation

As a classification task, confusion metric, accuracy, and F1 score are used for evaluation.

Confusion matrix is an essential instrument for evaluating the performance of classification models in classification tasks. As indicated in Table III, True Positive (TP) refers to the count of TP instances; True Negative (TN) denotes the count of TN instances; False Positive (FP) indicates the count of FP instances; False Negative (FN) represents the count of FN instances.

TABLE III. CONFUSION MATRIX FOR CLASSIFICATION EVALUATION

	Predict Positive	**Predict Negative**
Actual Positive	True Positive (TP)	False Positive (FP)
Actual Negative	True Negative (FN)	True Negative (TN)

Accuracy denotes the proportion of samples accurately identified by the model relative to the total number of samples (13).

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \quad (13)$$

F1 score is the harmonic average of precision and recall, designed to address the trade-off between these two metrics in classification problems, as indicated in (14). Meanwhile, precision denotes the ratio of correctly predicted positive instances to the total number of instances predicted as positive by the model, as presented in (15). At the same time, recall denotes the ratio of accurately predicted positive instances by the model to the total number of actual positive instances, as given in (16).

$$F1\ score=\frac{2*Precision*Recall}{Precision+Recall} \quad (14)$$

$$Precision=\frac{TP}{TP+FP} \quad (15)$$

$$Recall=\frac{TP}{TP+FN} \quad (16)$$

The 5-fold cross-validation method is employed to assess the model's performance and stability by partitioning the dataset into five subsets. By systematically employing several subsets as training and validation sets, the bias in model evaluation can be significantly mitigated.

E. Experimental Settings

Table IV illustrates the experimental settings.

TABLE IV. EXPERIMENTAL SETTINGS

Items	**Value**
GPU	NVIDIA RTX 4090 24G
Memory	64G
Cuda	v12.1.105
PyTorch	v2.2.1
Max Length of Sentence	150
Batch size	32
Epochs	4
Loss	Cross Entropy Loss
Optimizer	AdamW

IV. RESULT

A range of models are conducted to improve the model performance, including BERT-Fc, RoBERTa-Fc, Nezha-Fc, mCNNs-BERT, mCNNs-RoBERTa, mCNNs-Nezha, BERT-BiLSTM, RoBERTa-BiLSTM, and Nezha-BiLSTM, excluding the baseline and the proposed Sarcasm-sexism-mCNNs-Nezha. In the model of BERT/RoBERTa/Nezha-Fc, a fully connected neural network is concatenated after BERT/RoBERTa/Nezha. In the model of mCNNs-BERT/RoBERTa/Nezha, the mCNNs layer is concatenated after BERT/RoBERTa/Nezha. In the model of BERT/RoBERTa/Nezha-BiLSTM, the BiLSTM is concatenated after BERT/RoBERTa/Nezha.

The experiment result is illustrated in Table V. The proposed Sarcasm-sexism-mCNNs-Nezha model achieves the following performance metrics: 82.65% accuracy, 80.49% macro F1 score, 82.49% weighted F1 score, 86.94% non-sexism F1 score, and 74.05% sexism F1 score. It can be reported that the proposed model obtains the best performance on all five metrics except for sexism F1. Although the sexism F1 score in the proposed model is 0.39% lower than that of the RoBERTa-BiLSTM model, its accuracy is 0.55% higher. The mCNNs-Nezha model demonstrates the second-best performance.

As the best-performing model, Sarcasm-sexism-mCNNs-Nezha integrates sexism and sarcasm features, leveraging a deep learning framework with a masking mechanism. It outperforms all other deep learning approaches in Table V that do not incorporate sarcasm detection. It demonstrates significant improvements over the baseline model (BERT [2]) with an increase of 2.05% in accuracy and a 2.89% rise in macro F1. Specifically, the model achieves a 3.65% improvement in sexism F1, outperforming the 1.14% improvement in non-sexism F1. This highlights its superior capability in addressing sexism compared to non-sexism.

TABLE V. EXPERIMENTAL RESULT (%)

Model	Accuracy	Macro F1	Weight F1	Non-sexism F1	Sexism F1
BERT [2]	80.6	77.6	-	85.8	69.4
RoBERTa-lexicon [2]	80.4	78.0	-	85.3	70.7
BERT-Fc	81.86	79.77	81.78	86.25	73.28
RoBERTa-Fc	82.07	80.10	82.03	86.32	73.87
Nezha-Fc	81.84	79.61	81.69	86.31	72.91
mCNNs-BERT	81.55	79.45	81.48	85.99	72.90
mCNNs-RoBERTa	82.45	80.32	82.32	86.78	73.86
mCNNs-Nezha	82.53	80.24	82.32	**86.94**	73.55
BERT-BiLSTM	81.69	79.52	81.58	86.15	72.90
RoBERTa-BiLSTM	82.10	80.33	82.16	86.22	**74.44**
Nezha-BiLSTM	82.06	79.88	81.93	86.48	73.29
Sarcasm-sexism-mCNNs-Nezha	**82.65**	**80.49**	**82.49**	**86.94**	74.05

V. DISCUSSION

A. Confusion Matrix of Proposed Model

The confusion matrix of the proposed model is displayed in Fig. 2. There are 2,229 true predicted as sexisms, 863 false anticipated as sexisms, 5,179 accurately predicted as non-sexisms, and 692 false predicted as non-sexisms.

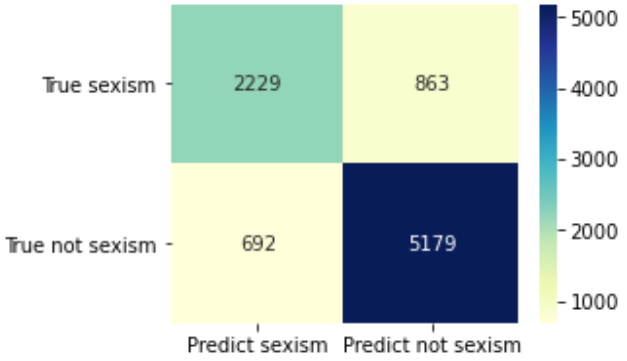


Fig. 2. Confusion matrix of proposed sarcasm-sexism-mCNNs-Nezha model

B. The Stability of Proposed Sarcasm-sexism-mCNNs-Nezha Model

The stability of the proposed model is demonstrated in Fig. 3. Fold 2 to fold 5 exhibits a higher concentration across all metrics. Fold 1 demonstrates a focus on non-sexism F1, suggesting that the capacity to identify non-sexism has not significantly altered. Nonetheless, on the other four metrics, fold 1 exhibits inferior performance compared to other folds, particularly for the sexism F1 score.

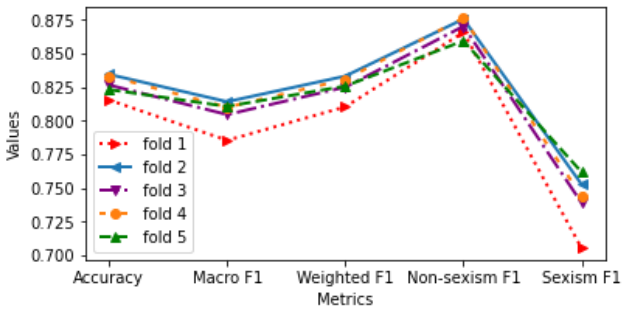


Fig. 3. The stability of the proposed Sarcasm-sexism-mCNNs-Nezha model

To examine the performance of this model on fold 1, a comparison between the mCNNs-Nezha model with the second-highest accuracy and the proposed model is conducted. The result is depicted in Fig. 4. It illustrates that the non-sexism F1 score is improved in this model, while the sexism F1 score is decreased. However, the accuracy is increased. This indicates that even though the performance of this model is worse on fold 1, it also improved the accuracy compared with the mCNNs-Nezha model. It illustrates the effectiveness of the proposed model.

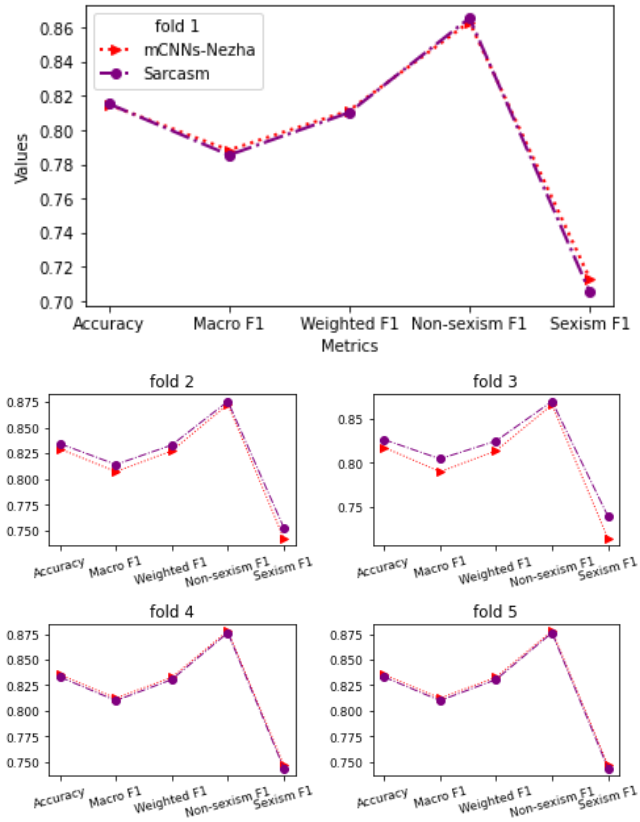


Fig. 4. Compared with the mCNNs-Nezha model on five folds

Nevertheless, the comparison of fold 2 to fold 5 demonstrates similar results in fold 4 and fold 5, while exhibiting a relatively greater enhancement in fold 2 and fold

3, as displayed in Fig. 4. It indicated that the proposed model is beneficial in enhancing performance in most cases.

C. Manually Analysis

A manual evaluation of sexist texts incorrectly identified by the mCNNs-Nezha model yet accurately by the proposed model is conducted. Table VI presents identical examples of sexist remarks that are misclassified by the mCNNs-Nezha model but accurately classified by the proposed approach. Sentences 2 and 3 contain metaphors, such as 田园狗 (a rural dog), which means pastoral feminist, and 鸡 (chicken), which means prostitute. Sentence 3 contains irony, which means, 'I do not admire you at all.'

TABLE VI. EXAMPLES OF SEXISM TEXT WRONGLY CLASSIFIED BY BASIC MODEL BUT RIGHTLY BY SARCASM-SEXISM-MCNNS-NEZHA MODEL

Id	Comments
1	明显田园狗做法。哈哈哈。这婚结的都不如去找鸡 This is obviously a rural dog behavior. Hahaha. This kind of marriage is worse than finding a prostitute.
2	说人家坦克真的好笑吗？那你们金针菇我们也可以笑吧？不会这么开不起玩笑吧，不会吧不会吧？ Is it really funny to say that others are tanks? Then can we, the Enoki mushrooms, also laugh? You can't be so incapable of taking a joke, right? No way, no way?
3	你真贴心，身为女人这么照顾男人的各种情绪，佩服你呢 You are so considerate. As a woman, you take care of a man's emotions. I admire you.

D. Ablation Experiment

To evaluate the significance of each component, an ablation experiment is conducted. There are three components except for Nezha. Table VII illustrates the result of the ablation experiment.

TABLE VII. ABLATION EXPERIMENT RESULT (%)

Id	Sarcasm	Sexism	mCNNs	accuracy	Macro F1	Weight F1	Non-sexism F1	Sexism F1
1	√	√	√	**82.65**	**80.49**	**82.49**	86.94	**74.05**
2		√	√	82.56	80.32	82.37	86.93	73.71
3	√		√	82.55	80.23	82.32	**86.98**	73.48
4	√	√		82.26	80.05	82.10	86.67	73.43
5	√			82.07	79.65	81.81	86.63	72.67
6		√		82.14	79.81	81.93	86.63	73.00
7			√	82.53	80.24	82.32	86.94	73.55
8				81.84	79.61	81.69	86.31	72.91

1) All three components have a positive impact: As summarized in Table VII, no matter eliminating one, two, or three components like sarcasm layer, sexism layer, and mCNNs layer from this model, the performance of the model on each metric such as accuracy, macro F1 score, weighted F1 score, sexism F1, non-sexism F1, is decreased except that the non-sexism F1 is improved 0.04% while eliminating sexism component. This indicates that every component has a positive effect on the proposed model.

2) mCNNs contributed more than sarcasm or sexism: Fig. 5 illustrates the model performance degradation while eliminating one component from the proposed model. It is evident that the model's performance declines when one component is removed and that mCNNs are the most crucial component of the proposed model, followed by sarcasm and sexism. It is noteworthy that the non-sexism F1 improves when the sexism component is eliminated, even though overall performance declines. The reason may attributed to the fact that the sexism lexicon layer tends to classify text containing sexist words as sexism. When the sexism component is deleted, some text containing sexist words may be classified as non-sexism, leading to improvement in non-sexism F1 score.

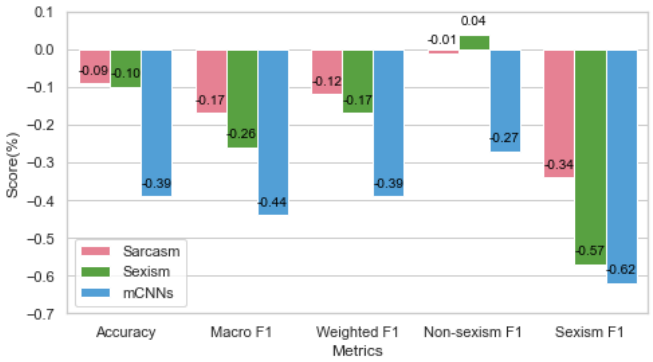


Fig. 5. Eliminating one component from Sarcasm-sexism-mCNNsNezha

3) Impact of sarcasm component: Fig. 6 illustrates the degradation of eliminating the sarcasm component and eliminating both the sarcasm component and the sexism component. If the sarcasm component is removed from the model, there will be a decrease of 0.09%, 0.17%, 0.12%, 0.01%, and 0.34% on the accuracy, macro F1, weighted F1, non-sexism F1, and sexism F1, respectively. It indicates that the sarcasm component has a positive effect on sexism detection, and it has affected more the sexism content than the non-sexism content.

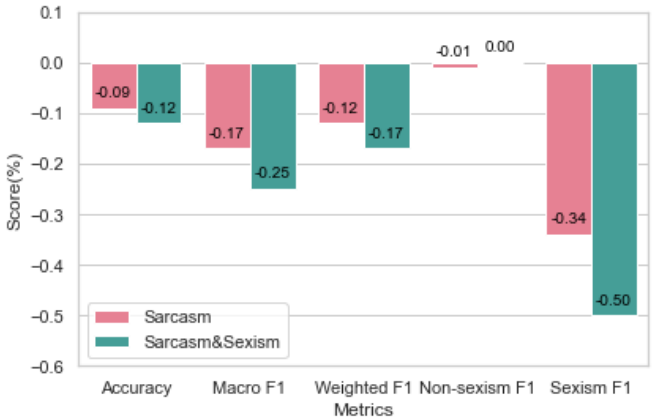


Fig. 6. Impact of sarcasm

It also illustrates that it decreases more by deleting both sarcasm and sexism from this model. It indicates that the combination of sarcasm and sexism has more influence than sarcasm. The effectiveness of sarcasm masking mechanisms is proven.

Both strategies demonstrate more impact on sexist content than non-sexism content. This indicates that they can improve the detection of sexism more than non-sexism.

4) Impact of sexism component: Fig. 7 illustrates the degradation of eliminating the sexism component and eliminating both sarcasm and sexism. It suggests that no matter whether the sexism *component* or a combination of sarcasm and sexism has a positive impact on the model. From the perspective of accuracy, the combination of sarcasm and sexism has a higher impact. Moreover, they have the same weighted F1 score.

Nevertheless, from the perspective of the macro F1 score, the combination of sarcasm and sexism has more influence than sarcasm (Fig. 6) but less influence than sexism. In addition, it is evident that the sexism *component* has a positive impact on detecting sexism while a negative impact on detecting non-sexism. Meanwhile, the combination of sexism and sarcasm has a positive impact on sexism but has no impact on non-sexism. It indicates a complex interaction between sexism and sarcasm.

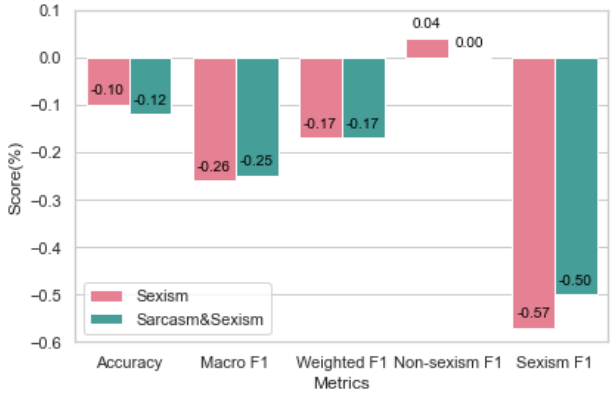


Fig. 7. Impact of sexism

5) Decision for coefficients: The output of the combination of sarcasm and sexism accounts for 20% of the final output, and the output of the mCNNs accounts for 80% of the final output. Different coefficients are adopted.

TABLE VIII. MODEL WITH DIFFERENT COEFFICIENTS (%)

Coefficient For mCNNs-Nezha	Coefficient for Combination of Sexism and Sarcasm	Accuracy	Macro F1
0.75	0.25	82.56	80.40
0.8	0.2	82.65	80.49
0.85	0.15	82.60	80.44
0.9	0.1	82.60	80.55

Table VIII illustrates the performance of the model with different coefficients. It can be discovered that the coefficient of (0.8, 0.2) contributes the best performance on accuracy. However, the coefficient of (0.9, 0.1) contributes the best performance on macro F1. Since (0.8, 0.2) has the highest accuracy, it is selected for the final model.

VI. CONCLUSION

An effective way to detect Chinese sexism text is proposed. The model combines a sarcasm lexicon, sexism lexicon, transformer-based Nezha, and three CNNs. The accuracy and macro F1 score reached 82.65% and 80.49%, outperforming the baseline with 2.05% and 2.89%, respectively. A sarcasm lexicon is constructed, and a sarcasm masking mechanic is proposed. In the ablation experiment, all evaluation metrics are decreased when sarcasm is removed, proving the effectiveness of the sarcasm masking mechanic. The sarcasm masking mechanic has the potential to generate the detection of sexism in other languages. However, this research only considers sarcasm feature words and does not account for inconsistent expressions in text or other sarcasm detection techniques that may be useful for detecting sexism, which contains sarcasm. Additionally, data augmentation methods, particularly back translation, can be considered, as some implicit sexist comments become explicit sexist content after being translated into English.

REFERENCES

[1] H. Mulki and B. Ghanem, “Let-Mi: an Arabic levantine twitter dataset for misogynistic language,” In Proceedings of the Sixth Arabic Natural Language Processing Workshop, pp. 154–163, Apr 2021.

[2] A. Jiang, X. Yang, L. Yang, and A. Zubiaga, “SWSR: a Chinese dataset and lexicon for online sexism detection,” Online Social Networks and Media, vol. 27, pp. 100182, Jan 2022.

[3] F. Rodriguez-Sanchez, J. Carrillo-de-Albornoz, and L. Plaza, “Automatic classification of sexism in social networks: an empirical study on twitter data,” IEEE Access, vol. 8, pp. 219563–219576, Dec 2020.

[4] S. Khotijah, J. Tirtawangsa, and A. A. Suryani, “Using LSTM for context based approach of sarcasm detection in twitter,” ACM Int. Conf. Proceeding Ser., no. 19, pp. 1–7, Jul 2020.

[5] T. Abdullah, A. Ahmet, and U. Kingdom, “Deep learning in sentiment analysis : a survey of recent architectures,” ACM Computing Surveys, vol. 55, no. 159, pp. 1–37, Dec 2022.

[6] F. M. Plaza-Del-Arco, M.-D. Molina-González, L. A. Ureña-López, and M.-T. Martín-Valdivia, “Exploring the use of different linguistic phenomena for sexism identification in social networks,” in CEUR Workshop Proceedings, vol. 3202, Sept 2022.

[7] A. Y. Muaad et al., “Artificial intelligence-based approach for misogyny and sarcasm detection from Arabic texts,” Comput. Intell. Neurosci., vol. 2022, pp. 7937667, March 2022.

[8] S. Sharifirad and A. Jacovi, “Learning and understanding different categories of sexism using convolutional neural network’s filters,” in Proceedings of the 2019 Workshop on Widening NLP, pp. 21–23, Aug 2019.

[9] M. A. Bashar, R. Nayak, and N. Suzor, “Regularising LSTM classifier by transfer learning for detecting misogynistic tweets with small training set,” Knowl. Inf. Syst., vol. 62, no. 10, pp. 4029–4054, Jun 2020.

[10] A. Rahali, M. A. Akhloufi, A.-M. Therien-Daniel, and E. Brassard-Gourdeau, “Automatic misogyny detection in social media platforms using attention-based Bidirectional-LSTM,” in 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2706–2711, 2021.

[11] T. Lynn, P. T. Endo, P. Rosati, I. Silva, G. L. Santos, and D. Ging, "A comparison of machine learning approaches for detecting misogynistic speech in urban dictionary," 2019 International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA), pp. 1-8, Jun 2019.

[12] P. Chiril, V. Moriceau, F. Benamara, A. Mari, G. Origgi, and M. Coulomb-Gully, "An annotated corpus for sexism detection in French tweets," Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc., pp. 1397–1403, May 2020.

[13] N. Safi Samghabadi, P. Patwa, S. Pykl, P. Mukherjee, A. Das, and T. Solorio, "Aggression and misogyny detection using BERT: a multi-task approach," Proc. Second Work. Trolling, Aggress. Cyberbullying, pp. 126–131, May 2020.

[14] R. Calderón-Suarez, R. M. Ortega-Mendoza, M. Montes-Y-Gómez, C. Toxqui-Quitl, and M. A. Márquez-Vera, "Enhancing the detection of misogynistic content in social media by transferring knowledge from song phrases," IEEE Access, vol. 11, pp. 13179–13190, Feb 2023.

[15] A. Singh, D. Sharma, and V. K. Singh, "Misogynistic attitude detection in YouTube comments and replies: A high-quality dataset and algorithmic models," Comput. Speech Lang., vol. 89, pp. 101682, Jan 2025.

[16] A. Y. Muaad, H. J. Davanagere, M. A. Al-antari, J. V. B. Benifa, and C. Chola, "AI-based misogyny detection from Arabic levantine twitter tweets," vol. 2. no. 1, pp. 15, Sept 2021.

[17] E. W. Pamungkas, V. Basile, and V. Patti, "Misogyny detection in twitter: a multilingual and cross-domain study," Inf. Process. Manag., vol. 57, no. 6, pp. 102360, Nov 2020.

[18] T. Jain et al., "Detection of sexually harassing tweets in Hindi using deep learning methods," Int. J. Softw. Innov., vol. 10, no. 1, pp. 1-15, 2022.

[19] N. A. Hamzah and B. N. Dhannoon, "Detecting Arabic sexual harassment using bidirectional long-short-term memory and a temporal convolutional network," Egypt. Informatics J., vol. 24, no. 2, pp. 365–373, Jul 2023.

[20] H. Mohammadi, A. Giachanou, and A. Bagheri, "A transparent pipeline for identifying sexism in social media: combining explainability with model prediction," Appl. Sci., vol. 14, no. 19, 2024, doi: 10.3390/app14198620.

[21] F. Belbachir, T. Roustan, and A. Soukane, "Detecting online sexism: integrating sentiment analysis with contextual language models," AI, vol. 5, no. 4, pp. 2852 – 2863, Dec 2024.

[22] F. Rodríguez-Sánchez, J. Carrillo-de-Albornoz, and L. Plaza, "Leveraging unsupervised task adaptation and semi-supervised learning with semantic-enriched representations for online sexism detection," Expert Syst., vol. 42:e13763, no. 2, 2025.

[23] R. Wang, Q. Wang, B. Liang, Y. Chen, and B. Qin, "Masking and generation : an unsupervised method for sarcasm detection," In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, vol. 1, no. 1, pp. 2172-2177, Jul 2022.

[24] S. Wei, G. Zhu, G. Tan, and S. Zhang, "Ironic sentence recognition model integrating ironic language features," CAAI Trans. Intell. Syst., vol. 19, pp. 689–696, May 2024. http://tis.hrbeu.edu.cn/Upload/PaperUpLoad/f612a74d-dd2c-49b3-81b1-28422add80a4.pdf.

[25] E. Lunando and A. Purwarianti, "Indonesian social media sentiment analysis with sarcasm detection," 2013 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2013, pp. 195–198, 2013.

[26] A. Joshi, P. Bhattacharyya, and M. J. Carman, "Automatic sarcasm detection: a survey," ACM Comput. Surv., vol. 50, no. 5, pp. 1-22, Sept 2017.

[27] S. Li, G. Zhu, and J. Li, "A method of Chinese ironic detection integrated with hyperbolic representation," Data Anal. Knowl. Discov., no. 3, pp. 1–10, 2024. https://www.cnki.net/KCMS/detail/detail.aspx?dbcode=CAPJ&filename=ZZDZ20241030001.

[28] W. Hu, L. Chen, T. Han, Z. Zhong, and C. Ma, "A multimodal Chinese sarcasm detection model integrated with linguistic features," J. Zhengzhou Univ. Sci. Ed., pp. 1–8, 2024. https://www.cnki.net/KCMS/detail/detail.aspx?dbcode=CAPJ&filename=ZZDZ20241030001.

[29] J. Zhang and L. Jiang, "Research on irony recognition of travel reviews based on multi-modal deep learning," Inf. Stud. Theory Appl., vol. 45, no. 7, pp. 158–164, Jul 2022. https://www.cnki.net/KCMS/detail/detail.aspx?dbcode=CJFD&filename=QBLL202207022.

[30] A. Joshi, S. Agrawal, P. Bhattacharyya, and M. J. Carman, "Expect the unexpected: harnessing sentence completion for sarcasm detection," Commun. Comput. Inf. Sci., vol. 781, pp. 275–287, Mar 2018.

[31] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, "Sarcasm as contrast between a positive sentiment and negative situation," Conference on Empirical Methods in Natural Language Processing, pp. 704–714, Oct 2013.

[32] X. Bai and R. Huo, "Research on recognition model of ironic text integrating language characteristics," Commun. Technol., vol. 54, no. 5, pp. 1126–1130, May 2021 https://cstj.cqvip.com/Qikan/Article/Detail?id=7104582899.

[33] X. Bai and R. Huo, "Ironic text recognition model incorporating language characteristics," Commun. Technol., vol. 54, no. 1, pp. 73–76, Jan 2021. https://www.cnki.net/KCMS/detail/detail.aspx?dbcode=CJFD&filename=TXJS202101012.

[34] X. Lu, Y. Li, and Wang , Suge, " Linguistic features enhanced convolutional neural networks for irony recognition," J. Chinese Inf. Process., vol. 33, no. 5, pp. 31 – 38, May 2019. https://www.cnki.net/KCMS/detail/detail.aspx?dbcode=CJFD&filename=MESS201905004.

[35] V. Govindan and V. Balakrishnan, "A machine learning approach in analysing the effect of hyperboles using negative sentiment tweets for sarcasm detection," J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 8, pp. 5110–5120, Sept 2022.

[36] X. Jia, Z. Deng, F. Min, and D. Liu, "Three-way decisions based feature fusion for Chinese irony detection," Int. J. Approx. Reason., vol. 113, pp. 324–335, Oct 2019.

[37] J. Liao, M. Wang, X. Chen, S. Wang, and K. Zhang, "Dynamic commonsense knowledge fused method for Chinese implicit sentiment analysis," Inf. Process. Manag., vol. 59, no. 3, pp. 102934, May 2022.

[38] J. Wei, J. Liao, Z. Yang, S. Wang, and Q. Zhao, "BiLSTM with multi-polarity orthogonal attention for implicit sentiment analysis," Neurocomputing, vol. 383, pp. 165–173, 2020.

[39] R. Xiang et al., "Ciron: A new benchmark dataset for Chinese irony detection," Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc., no. May, pp. 5714–5720, 2020.

[40] B. Liang, Z. Lin, R. Xu, and B. Qin, "Topic-oriented sarcasm detection: new task, new dataset and new method," J. Chinese Inf. Process., vol. 37, no. 2, pp. 138–157, Oct 2023.

[41] W. Hu, L. Chen, X. Huang, C. Chen, and Z. Zhong, "A multimodal Chinese sarcasm detection model for emergencies based on cross attention," CAAI Trans. Intell. Syst., vol. 19, no. 2, pp. 392–400, 2024.

[42] J. Wei, X. Ren, X. Li, W. Huang, Y. Liao, and Y. Wang, et al. "NEZHA: neural contextualized representation for Chinese language understanding," arXiv, pp. 1–9, Nov 2021. https://doi.org/10.48550/arXiv.1909.00204%0A.

Machine Learning-Enabled Personalization of Programming Learning Feedback

Mohammad T. Alshammari
College of Computer Science and Engineering, University of Ha'il, Ha'il, Saudi Arabia

***Abstract*—Acquiring programming skills is daunting for most learners and is even more challenging in heavily attended courses. This complexity also makes it difficult to offer personalized feedback within the time constraints of instructors. This study offers an approach to predict programming weaknesses in each learner to provide appropriate learning resources based on machine learning. The machine learning models selected for training and testing and then compared are Random Forest, Logistic Regression, Support Vector Machine, and Decision Trees. During the comparison based on the features of prior knowledge, time spent, and GPA, Logistic Regression was found to be the most accurate. Using this model, the programming weaknesses of each learner are identified so that personalized feedback can be given. The paper further describes a controlled experiment to evaluate the effectiveness of the personalized programming feedback generated based on the model. The findings indicate that learners receiving personalized programming feedback achieve superior learning outcomes than those receiving traditional feedback. The implications of these findings are explored further, and a direction for future research is suggested.**

***Keywords*—Machine learning; programming; learning outcome; feedback; personalization**

I. Introduction

The advances in Artificial Intelligence (AI) and its subfield, Machine Learning (ML), are rapidly transforming the education landscape, offering innovative approaches to improve teaching methods, accelerate learning, and deliver personalized educational experiences. AI-powered tools, including adaptive learning systems, intelligent tutoring platforms, and predictive analytics, are revolutionizing the traditional pedagogical landscape by generating personalized feedback, suggesting individualized learning pathways, and forecasting academic achievement [1], [2]. These technologies serve the diverse interests and needs of learners and support data-driven interventions, thereby enhancing the levels of engagement and retention both in traditional and online learning environments [3].

An example of this can be seen in adaptive learning systems that use AI to modify the material and pace of learning based on how each learner performs, resulting in a highly personalized educational experience [4]. Another AI application is Intelligent Tutoring Systems (ITSs), which can improve personalization by simulating one-on-one teaching and using advanced algorithms to analyze how a learner behaves and provide personalized guidance based on those metrics. Learner interaction data can also be employed as predictive analytics, which fully utilizes advanced ML models to analyze historical and real-time data, predict learners' outcomes, identify at-risk learners, and allow for considerable strategic resource provision. ML can also guide evidence-based decision-making, enabling educators to adjust curricula and instructional strategies by appropriately identifying learner behavior and patterns during learner-content interaction [5].

Programming is one of the most common subjects taught in a computer science curriculum. Still, it introduces several challenges, including mapping abstract concepts with practical activities and high dropout rates in introductory courses [6]. AI and ML in education can help mitigate these challenges by creating dynamic and tailored learning experiences. AI-based educational systems, for example, may analyze learner-system interaction data to develop personalized learning pathways, adjusting how content is delivered to suit learners' needs [7], [8].

Many ML models have been developed. For example, neural networks can be utilized to create learning recommendations that recommend personalized learning activities to keep learners engaged and retain knowledge [9]. Other ML models, like fuzzy logic, have also been used to offer personalized feedback and modify instructional approaches according to learners' characteristics, such as knowledge level and task accomplishments [10]. Recent approaches view chatbot-assisted programming systems powered by ML techniques as powerful learning tools that aid debugging, explain code errors, and suggest possible solutions, thereby serving as a proxy for a human tutor [11].

A vast majority of ML algorithms have been employed to predict programming learning performance; some of these are Decision trees (DT), Random Forests (RF), Support Vector Machines (SVM), Logistic Regression (LR), k-nearest Neighbors (kNN), and Artificial Neural Networks (ANN) [3], [12], [13], [14]. These models leverage learner engagement patterns, prior learning activities, and coding behavior to provide learners with timely intervention [15].

ML-enabled behavioral analytics tools can track and analyze learner engagement with resources such as coding environments, learning management systems, online coding platforms, and discussion forums. The analyses can reveal meaningful relationships between activities like active coding and passive lecture reviewing that may impact overall learning. Thus, actionable insights drive the curriculum design, define effective intervention strategies that address individual learning needs, and optimize the effectiveness of programming educational outcomes.

A key limitation in existing ML model-based investigations for programming education is the lack of a comprehensive approach that spans from data collection to evaluation in authentic learning environments [8], [16], [17]. While prior research has explored ML applications in programming learning, many studies focus on isolated aspects, such as predictive modeling or feedback generation, without fully integrating these components into a structured learning framework. This research gap highlights the need for further investigation into how ML models can be systematically leveraged to enhance personalized learning experiences. To address this, the present study conducts a comparative analysis of ML models to identify programming learning weaknesses based on learners' prior knowledge, time spent, and Grade Point Average (GPA). Identifying the most effective ML model can facilitate the delivery of personalized programming feedback tailored to individual learners' needs. Additionally, an experimental evaluation is implemented to explore the impact of personalized feedback on learning performance. This dual approach enhances the theoretical understanding of ML-driven feedback mechanisms and provides empirical evidence to guide future advancements in AI-supported programming education.

The key research questions are as follows:

RQ1. How can machine learning be used to predict specific weaknesses in learners' programming skills, enabling feedback personalization?

RQ2. Can personalized feedback derived from machine learning predictions enhance learning outcomes?

This paper is structured as follows: Section II presents related work. Section III outlines the methodology used in this study. Section IV offers the results. Section V discusses the findings, and Section VI concludes the paper.

II. Related Work

The recent reviews on integrating AI and ML in online learning platforms highlight their essential role in personalizing learning through customized content delivery for individual learners [1], [12], [13], [14]. The dynamic adaptation and targeted interventions through ML techniques (e.g., clustering, reinforcement learning, deep learning) are promising to enhance learner engagement, retention, and academic performance [18]. Yet, data privacy concerns, high resource requirements, and the risk of diminishing human interaction limit the widespread adoption. Careful implementation and ongoing refinement would be critical to mitigate these challenges and realize the potential benefits of more mainstream AI-based online learning platforms.

A predictive analysis of learner performance in programming tutoring comparing different ML models with ANN on dataset input finds that ANN outperforms other methodologies such as SVM, DT, and kNN [19]. However, the study's limitations include its targeting of a specific cohort of Norwegian learners, domain-specific instructions, and potential biases integrating into the suggested framework that may undermine the present findings' utility to broader domains.

Another approach investigates how ML algorithms are implemented to predict learner programming performance as high or low based on different variables, including computational identity, computational thinking, programming empowerment, gender, and programming anxiety [16]. DT algorithm had the best accuracy of 96.6%, while the best cross-accuracy was obtained with RF. On the contrary, kNN performed the least favorably. Their study shows that male and university learners scored higher than female and high school learners. However, the groups were not significantly different as far as programming anxiety is concerned. The limitations of the research are the focus on the demographic of Turkish learners, the specificity of the constructs relating to specific tasks, and the need for generalisability across datasets of different populations.

The study in [3] offers a prediction model powered by AI with Genetic Programming (GP) modeling to analyze and predict learner academic performance in an online engineering education system. The performance of the GP model was better than that of traditional AI methods (e.g., ANN and SVM), demonstrating high accuracy and efficient predictions regarding learning effectiveness. Key determinants of performance are knowledge transfer, participation in class, and summative assessment, with prior knowledge having a limited effect. However, though it does advance these areas, the limited sample size, lack of real-time adaptability of the model, and locally specific findings limit its wider application and require additional research.

An extensive review of ML techniques applied to predict the learner's performance in the context of programming courses emphasizes the strength of different algorithms, the nature of the organization of datasets, and the significance of the evaluation metrics [20]. According to the review results, SVM had the best accuracy of 93.97%, whereas deep learning methods like Deep Neural Networks (DNN) have achieved significant success in detecting complex data patterns. Most studies used academic records as a dataset type; however, a meta-analysis of multiple data sources improved prediction accuracy. The review highlights the importance of using different metrics to evaluate model performance, including accuracy, F1-Score, precision, and recall, especially in unbalanced datasets. Noted limitations include using small or unbalanced datasets, a narrow range of algorithms, and overemphasizing accuracy. The results emphasize the need for larger datasets, more varied evaluation methods, and more examination of deep learning approaches to improve models of future predictive capabilities.

PerFuSIT is an example of a fuzzy logic-based module designed to personalize pedagogical approaches in ITSs for programming education [10]. This adaptive system updates different tutoring approaches dynamically, according to parameters such as performance measures, coding error types, help requests, and the time taken to solve a task. Therefore, through such an application, a fuzzy logic system can offer flexibility in the learning process, where the learners have different levels of interactivity, improving learners' performance and involvement. However, the dependency on expert regulations and the constrained testing environments may generate scaling challenges.

The study in [21] studied the relationship between the learning behaviors of learners and their grades in an online programming course employing the Random Matrix Theory to filter noisy data and discover possible patterns. It was observed that learners with superior academic performance regularly participated in practical tasks such as lab activities and exercises. The lower-achieving learners relied more on lecture notes and fell behind on engagement in practical tasks over time. This study demonstrated that data obtained from early-stage learning behaviors can be used as powerful binary predictors (pass or fail) of learning outcomes, and the use of cleaned datasets enabled significantly improved accuracy using ML algorithms, specifically SVM and XGBoost. However, this research also granted limitations; it had one institutional focus, short-term job performance outcomes, and methodological technical complexity, limiting the additional dissemination and uptake by broader generalization.

Another ML study uses data from a chatbot-assisted programming platform to explore the relationship between learning behaviors and programming performance [17]. The key performance features are solution verification, frequency of code review, code error correction using quizzes, programming practice logs, and engagement patterns. ANN produced better model predictions than other ML models like RF and SVM. While the study demonstrates the potential of ML for predicting outcomes, it is limited by its small sample size, single-institution focus, and short-term analysis of learning performance.

As an ML model, Recurrent Neural Networks (RNN) are also used as a basis for a personalized learning path recommendation system in the programming education domain [22]. The system can recommend appropriate problems to help the learner's speed and engagement by analyzing learners' ability charts and clustering users with similar skill levels. However, the system's reliance on advanced algorithms and lack of scalability testing may restrict its applicability and implementation in real-world scenarios, as they focus on data generated from a single online judge.

The study in [15] proposed personalized interventions based on the skills of at-risk learners as part of a Python programming course. Their study adopted advanced AI technologies (e.g., BERT and GPT-2) to generate personalized remedial content. Its personalized feedback significantly enhanced learners' coding competencies and learning strategies, demonstrating the power of AI-based interventions. However, the study also has limitations, including a relatively small cohort size, a focus on short-term outcomes, and reliance on resource-intensive AI models, which may limit applicability and scalability.

The study detailed in this manuscript differs from the above efforts to apply ML to programming education in three key aspects. First, it can be based on data retrievable from online learning platforms during a preliminary, intermediate, or final learning process. Second, different ML models can be evaluated using this data to find the best model for providing personalized programming feedback. Third, the empirical efficacy of such feedback in enhancing learning outcomes is assessed using controlled experimental evaluation.

III. METHOD

The experiment aims to use ML models to identify learners' weaknesses in Java programming learning and provide personalized learning feedback. Furthermore, it seeks to validate the effectiveness of this feedback in improving programming skills.

A. Hypotheses

There are two main hypotheses in this study as follows:

H1. Machine learning can be utilized to accurately predict specific weaknesses in learners' Java programming learning.

H2. Providing personalized feedback based on machine learning predictions will improve learners' performance in Java programming learning.

B. Data Collection

The experiment administers a comprehensive pre-test covering the essential concepts of basic Java programming. These concepts include syntax, variables, control flows (if, if-else, and switch), loops (while, do-while, and for loops), arrays, methods, classes, and objects. The pre-test contains 60 questions to cover 12 Java topics. Learners' responses to each concept are recorded to form the input dataset for training and testing the ML models. In addition, prior knowledge performance, time spent, and GPA are recorded.

C. Machine Learning Models

From data pre-processing, model selection, and training to evaluation, this phase lays the foundation for an ML model specifically designed to predict each learner's weaknesses from their pre-test scores, time spent on the exam, and GPA. The primary step involves validating the data by cleansing missing data, normalizing scores, and converting the categorical data (e.g., concept titles or feedback categories) to numerical values when required. The dataset will subsequently be randomly partitioned, allocating 70% for training and 30% for testing. The training dataset is utilized to develop the ML model, whereas the testing dataset evaluates its performance. So, several ML models such as RF, LR, SVM, and DT are compared to identify which model could best predict learners' weaknesses based on their pre-test, time spent, and GPA data. After finding a well-performing model, personalized guidance feedback will accordingly be delivered.

D. Personalized Feedback Generation

Based on the model's predictions, learners should be provided relevant feedback on the Java programming concepts they struggled with. For instance, personalized feedback will be prioritized and released when the model predicts that a learner struggles to grasp the programming concept while loops. The feedback input is pre-created per learning concept, with a theoretical overview, an example, an exercise, and a link to a video lesson suited to reflect the various learning types of learners.

E. Evaluation Metrics

In addition to the confusion matrix and Receiver Operating Characteristic (ROC) curves, alongside Area Under the Curve (AUC) scores, the predictive performance of the ML models is evaluated using metrics such as accuracy, precision, recall, and

F1-Score (displayed in formulas 1 through 4). These metrics are typically used in relevant work and are suitable for this study [23], [24]. A post-test is also used to investigate the effect and provide evidence for the advantages of using personalized feedback regarding improved learning outcomes.

$$Accuracy = \frac{TP+TN}{Total\ Predictions} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1-Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (4)$$

where,

- TP: True positive
- TN: True negative
- FP: False positive
- FN: False negative

F. Procedure

The experimental procedure is simplified and depicted in Fig. 1. After obtaining consent to participate in the experiment, a pre-test was conducted. The data collected was then used to train and test the ML models to select the best model for providing personalized feedback. Participants were randomly divided into two independent groups: experimental and control groups. The experimental group receives personalized feedback based on the selected ML model predictions, while the control group receives conventional feedback. Both groups were asked to follow the feedback for their learning.

This experimental setting was conducted in computer labs to ensure the experiment's control and to guarantee that participants completed the learning process based on the offered feedback. This process took five sessions, with each session lasting about 120 minutes. Once all learning sessions were finished, all participants completed the post-test.

IV. Results

This section presents the results pertaining to the prediction outcomes of the ML models. These predictions can be used of identifying the programming learning weaknesses of each learner. Furthermore, this section provides the results of the controlled experiment that aimed to investigate the effectiveness of learning Java programming through personalized feedback.

A. Dataset

This experiment involved 205 first-year undergraduate learners majoring in various computing disciplines. After collecting and pre-processing the pre-test results, 180 valid cases were identified for inclusion in the analysis of the ML models. The performance of these models was evaluated using standardized scores for each Java concept, categorizing learners as “weak” or “strong” according to a 50% threshold, time spent on the exam and GPA. Based on the optimal model, personalized feedback can be provided.

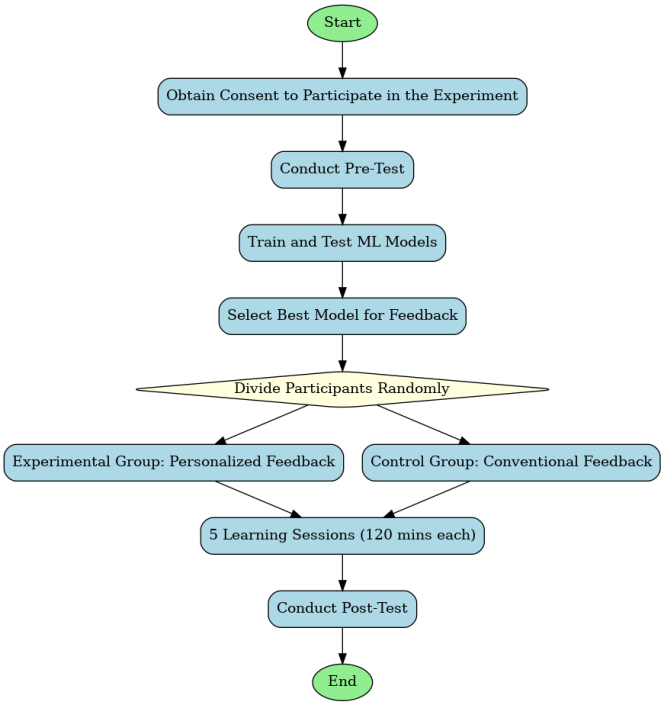


Fig. 1. Experimental procedure.

B. Analysis of the Machine Learning Models Performance

The dataset was utilized to analyze the four ML models: RF, LR, SVM, and DT. Table I summarizes the results of these models: accuracy, precision, recall, and F1-Score. Fig. 2 and Fig. 3 also present the confusion matrix and ROC curves with AUC scores, respectively. These findings will assist in identifying the appropriate ML model for the dataset.

Based on metrics analysis results, LR performs best at 98.33% accuracy meaning less errors, followed by SVM at 95% accuracy and finally RF at 81.67% accuracy. Looking into precision, SVM and LR is ahead in this perspective, with near-perfect precision that suggests low number of false positives. DT and RF are less effective since they have more false positives. Looking at the recall results, LR is outstanding (100%) at detecting all of the actual positive cases. SVM (97%) and RF (58%) display varying performance levels, while DT had the lowest recall results. For the F1-Scores, SVM (98.59%) and LR (97.96%) achieve a strong ratio between precision and recall. However, DT and RF can be less effective based on the results.

TABLE I. Performance Analysis of ML Models

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	81.67%	93%	58%	72.00%
Logistic Regression	98.33%	96%	100%	97.96%
SVM	95.00%	100%	97%	98.59%
Decision Tree	75.00%	83%	68%	75.00%

In the confusion matrix results, LR has shown an exceptional level of accuracy and recall with a near 100% score in both metrics. LR can also predict positive and negative

outcomes with remarkable precision. SVM also had a high precision at 100% and a high recall at 97%, making it a strong candidate due to the balance between those two evaluation metrics. Although its recall is slightly lower than that of LR, this results in a much more competitive score. The RF model also performed well, with many true positives (35) and fewer false negatives (1). However, it was slightly decreased due to several false positives (10), leading to 93% precision. Given the model's simplicity, DT showed satisfactory recall (68%) and accuracy (75%). However, DT has also demonstrated a relatively high number of false negatives (11), signifying a worse performance than the other ML models.

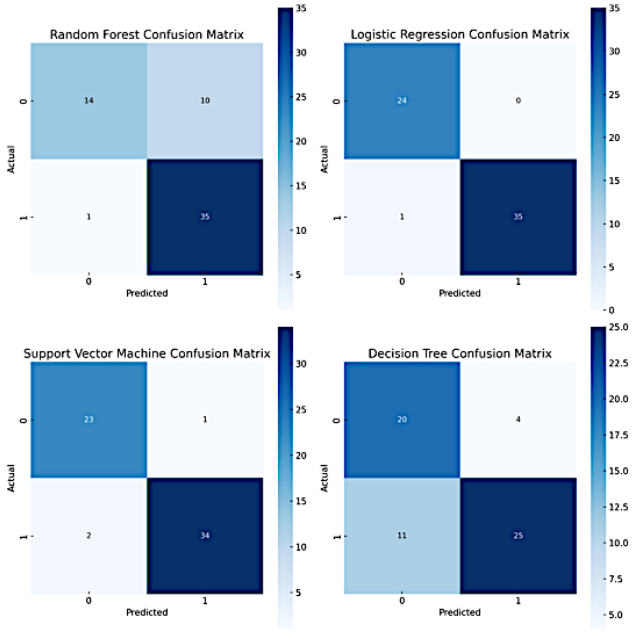


Fig. 2. Confusion matrix for the machine learning models.

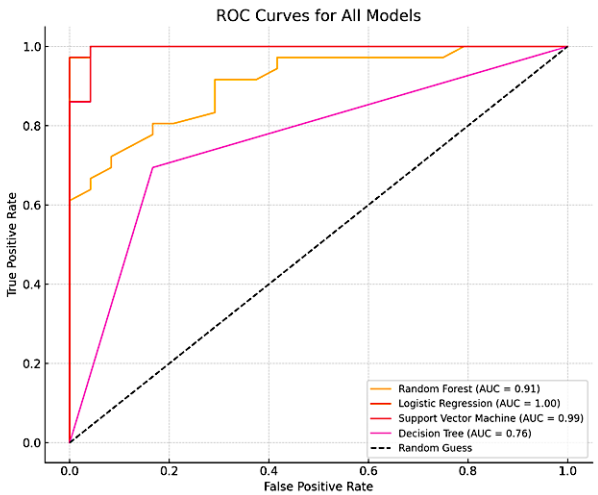


Fig. 3. ROC curves for the machine learning models.

When considering the ROC curves and AUS scores, LR stands out with an AUC of 1.00, implying perfect discrimination. SVM is also very close, with an AUC of 0.99, suggesting excellent performance. RF has a solid AUC of 0.91, indicating strong classification abilities but slightly worse than SVM and LR. DT falls behind with an AUC of 0.76, reflecting its lower performance in separating the weaknesses classes.

LR had the best overall outcomes considering all these metrics, with perfect recall, high precision, and balanced metrics. SVM performed competitively, with only slightly lower recall. RF and DT can be supplementary models for exploratory analysis or explainability. Therefore, LR is the candidate model for classifying and predicting each learner's weaknesses in Java programming learning.

C. Identifying Weaknesses of Programming Learning

LR has been identified as the most effective model among the ML approaches considered. To illustrate its capability to discern the weaknesses of each learner, two specific scenarios have been selected and presented, as depicted in Fig. 4.

Firstly, the identification of the highest-performing learner within the dataset is depicted in Fig. 4 (A). This learner exhibited deficiencies in knowledge and skills across three topics out of the twelve evaluated. These topics pertain to Java loops (including do-while and for loops) and arrays. Secondly, the identification of the lowest-performing learner in the dataset is illustrated in Fig. 4 (B). This learner demonstrated weaknesses in knowledge and skills across ten topics. These identified areas of weakness can be prioritized for each learner to facilitate the provision of personalized feedback.

Based on these findings, H1 is confirmed. It can be stated that ML can be utilized to accurately predict specific weaknesses in learners' Java programming learning.

Fig. 4. Identification of weak topics for the best-performing learner (case A) and the poorest-performing learner (case B)

D. Learning Effectiveness of Personalized feedback

After completing the phase of selecting the best-performing ML model (i.e., LR), each learner can receive personalized learning feedback. Table II gives a sample of the personalized recommendations generated for three random learners from the dataset.

TABLE II. TOPIC PERSONALIZED RECOMMENDATIONS FOR LEARNERS

Learner ID	Weak topics with scores
20161	[('Arrays', 20), ('Do-While Loops', 40), ('Classes', 40), ('Syntax', 60), ('Variables', 60), ('If Statements', 60), ('Switch Statements', 60), ('While Loops', 60), ('For Loops', 60), ('Objects', 60), ('If-Else Statements', 80), ('Methods', 100)]
20125	[('Methods', 20), ('Variables', 40), ('Do-While Loops', 40), ('For Loops', 40), ('Objects', 40), ('If Statements', 60), ('Switch Statements', 60), ('While Loops', 60), ('Syntax', 80), ('If-Else Statements', 80), ('Arrays', 100), ('Classes', 100)]
20310	[('If-Else Statements', 0), ('Do-While Loops', 0), ('For Loops', 0), ('Methods', 0), ('Classes', 20), ('Syntax', 40), ('Variables', 40), ('Arrays', 40), ('Switch Statements', 60), ('While Loops', 60), ('Objects', 60), ('If Statements', 80)]

LR can be used to adeptly identify the specific weaknesses of each learner to deliver personalized feedback. Nonetheless, a pertinent question arises: are these recommendations practical for learning? To address this issue, a controlled experimental evaluation was undertaken, wherein learners completed the learning process based on either personalized or conventional feedback as offered according to their group. The outcomes of this experiment are detailed in Table III. The sample consisted of 40 participants (out of 180 participants) who completed the experiment. Each group had 20 participants randomly assigned to either the control or experimental groups. The pre-test results indicate that they had almost similar scores. In contrast, the experimental group had better scores than the control group regarding the post-test and overall learning outcome (i.e., post-test – pre-test).

TABLE III. SUMMARY RESULTS OF THE PRE-TEST, POST-TEST AND LEARNING OUTCOME

Variable	Group	N	Mean	SD
Pre-test	Control	20	39.92	3.52
	Exp.	20	41.17	3.83
Post-test	Control	20	59.45	10.68
	Exp.	20	73.40	13.00
Learning outcome	Control	20	19.53	10.47
	Exp.	20	32.23	13.29

An independent sample t-test was run for all these variables. It was found that there was no statistically significant difference between the pre-test of the experimental group compared to the control group, $t(38) = -1.074$, $p = .289$. Regarding the post-test results, there was a statistically significant difference between the experimental group and the control group, $t(38) = -3.708$, $p<.001$. For the overall learning outcome, there was also a statistically significant difference between the experimental group and the control group, $t(38) = -3.358$, $p = .002$. According to the findings, H2 can be confirmed. It can be stated that providing personalized feedback based on ML predictions will improve learners' performance in Java programming learning.

V. DISCUSSION

The research presented in this study aimed to explore the use of ML to predict the programming learning weaknesses of learners. It emphasized the importance of learners' key data as input to ML models to identify learning patterns and gaps of programming concepts to provide adaptive and personalized interventions with relevant learning resources and feedback. The data features considered were prior knowledge obtained from a pre-test, time spent on the test, and GPA.

The study findings found that LR was the most effective model for identifying the programming learning weaknesses of learners compared to other ML models. It achieved the highest accuracy among other models, confirming its suitability for educational data analytics, particularly on the identified features for the programming domain. SVM also had a competitive performance but with lower recall compared to LR. This finding confirms the importance of precision and sensitivity in programming learning platforms when using ML. However, the limited utility of other models (i.e., RF and DT) is due to the higher false positives, though they can be helpful to exploratory analyses. These results align with previous research highlighting the robustness of LR and SVM in educational contexts [6], [7].

The presented study also took a step further by conducting a controlled experimental evaluation to investigate the effectiveness of personalized feedback based on the results of the selected ML model (i.e., LR). The findings revealed that providing personalized feedback based on LR predictions improves learners' outcomes in Java programming. By addressing these learning weaknesses, instructors and online learning platforms can implement targeted interventions to bridge learning gaps, thereby enhancing overall comprehension and engagement [10], [15]. These findings highlight the importance of integrating ML tools into programming curricula to meet individual learning needs. This is consistent with existing literature highlighting individualized learning pathways as a form of personalization for deeper understanding and long-term knowledge retention [2], [22].

The presented study offers significant implications for the design and deployment of ML in programming education interventions. This study demonstrated how ML can be used to provide precise and actionable insights that enhance learning performance. The findings also underscore the potential of integrating ML tools into computer science curricula to meet the diverse needs of learners effectively. However, some limitations need to be considered. Generalization can be limited since the results were based on a dataset from a single institute with restricted data features. Thus, more experiments and diverse data features are needed in future research. Nevertheless, this study provided initial findings that can serve as a foundation for further explorations highlighting the importance of ML-enabled personalization in education.

VI. CONCLUSION

This study addressed the challenge of identifying and resolving weaknesses in programming education using ML.

Three key contributions were made. First, multiple ML models were evaluated to predict learners' programming difficulties based on prior knowledge, time spent, and GPA. Second, a comparative analysis determined that LR was the most effective model for generating personalized feedback. Third, a controlled experimental evaluation provided empirical evidence that personalized feedback significantly enhances programming learning outcomes compared to conventional methods.

These findings highlight the potential of ML-driven personalized learning in programming education. Future research will enhance this approach by incorporating additional learner-system interaction features, such as time spent on specific concepts, quiz scores, quiz attempts, and lesson visits, to build more dynamic learner profiles. By leveraging these profiles, intelligent online learning platforms can be developed to generate adaptive learning pathways, provide targeted feedback, and integrate gamification elements to boost engagement and motivation. Furthermore, a more extensive experimental evaluation will be conducted to assess the long-term impact of personalized feedback on learning effectiveness.

REFERENCES

[1] Gligorea, M. Cioca, R. Oancea, A.-T. Gorski, H. Gorski, and P. Tudorache, "Adaptive learning using artificial intelligence in e-learning: a literature review," Educ Sci (Basel), vol. 13, no. 12, p. 1216, 2023.

[2] M. Tedre et al., "Teaching machine learning in K–12 classroom: Pedagogical and technological trajectories for artificial intelligence education," IEEE Access, vol. 9, pp. 110558–110572, 2021.

[3] P. Jiao, F. Ouyang, Q. Zhang, and A. H. Alavi, "Artificial intelligence-enabled prediction model of student academic performance in online engineering education," Artif Intell Rev, vol. 55, no. 8, pp. 6321–6344, 2022, doi: 10.1007/s10462-022-10155-y.

[4] M. T. Alshammari and A. Qtaish, "Effective Adaptive E-Learning Systems According to Learning Style and Knowledge Level.," Journal of Information Technology Education: Research, vol. 18, pp. 529–547, 2019, doi: https://doi.org/10.28945/4459.

[5] R. Mustapha, G. Soukaina, Q. Mohammed, and A. Es-Saadia, "Towards an Adaptive e-Learning System Based on Deep Learner Profile, Machine Learning Approach, and Reinforcement Learning," International Journal of Advanced Computer Science and Applications, vol. 14, no. 5, pp. 265–274, May 2023.

[6] S. Marwan, G. Gao, S. Fisk, T. W. Price, and T. Barnes, "Adaptive Immediate Feedback Can Improve Novice Programming Engagement and Intention to Persist in Computer Science," in Proceedings of the 2020 ACM Conference on International Computing Education Research, in ICER '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 194–203. doi: 10.1145/3372782.3406264.

[7] M. Murtaza, Y. Ahmed, J. A. Shamsi, F. Sherwani, and M. Usman, "AI-based personalized e-learning systems: Issues, challenges, and solutions," IEEE Access, vol. 10, pp. 81323–81342, 2022.

[8] W. Xu and F. Ouyang, "A systematic review of AI role in the educational system based on a proposed conceptual framework," Educ Inf Technol (Dordr), vol. 27, no. 3, pp. 4195–4223, 2022.

[9] T. Saito and Y. Watanobe, "Learning path recommendation system for programming education based on neural networks," International Journal of Distance Education Technologies (IJDET), vol. 18, no. 1, pp. 36–64, 2020.

[10] K. Chrysafiadi and M. Virvou, "PerFuSIT: Personalized Fuzzy Logic Strategies for Intelligent Tutoring of Programming," Electronics (Basel), vol. 13, no. 23, p. 4827, 2024.

[11] M. Abolnejadian, S. Alipour, and K. Taeb, "Leveraging ChatGPT for Adaptive Learning through Personalized Prompt-based Instruction: A CS1 Education Case Study," in Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems, in CHI EA '24. New York, NY, USA: Association for Computing Machinery, 2024. doi: 10.1145/3613905.3637148.

[12] P. L. S. Barbosa, R. A. F. do Carmo, J. P. P. Gomes, and W. Viana, "Adaptive learning in computer science education: A scoping review," Educ Inf Technol (Dordr), 2023, doi: 10.1007/s10639-023-12066-z.

[13] A. T. Bimba, N. Idris, A. Al-Hunaiyyan, R. B. Mahmud, and N. L. B. M. Shuib, "Adaptive feedback in computer-based learning environments: a review," Adaptive Behavior, vol. 25, no. 5, pp. 217–234, 2017, doi: 10.1177/1059712317727590.

[14] F. Okubo, T. Shiino, T. Minematsu, Y. Taniguchi, and A. Shimada, "Adaptive Learning Support System Based on Automatic Recommendation of Personalized Review Materials," IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, vol. 16, no. 1, pp. 92–105, Feb. 2023, doi: 10.1109/TLT.2022.3225206.

[15] A. Y. Q. Huang et al., "Personalized Intervention based on the Early Prediction of At-risk Students to Improve Their Learning Performance," Educational Technology & Society, vol. 26, no. 4, pp. 69–89, 2023, [Online]. Available: https://www.jstor.org/stable/48747521

[16] A. Durak and V. Bulut, "Classification and prediction-based machine learning algorithms to predict students' low and high programming performance," Computer Applications in Engineering Education, vol. 32, no. 1, p. e22679, Jan. 2024, doi: https://doi.org/10.1002/cae.22679.

[17] Y.-S. Su, Y.-D. Lin, and T.-Q. Liu, "Applying machine learning technologies to explore students' learning features and performance prediction," Front Neurosci, vol. 16, 2022, [Online]. Available: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2022.1018005

[18] Y. Jing, L. Zhao, K. Zhu, H. Wang, C. Wang, and Q. Xia, "Research Landscape of Adaptive Learning in Education: A Bibliometric Study on Research Publications from 2000 to 2022," Sustainability, vol. 15, no. 4, 2023, doi: 10.3390/su15043115.

[19] M. Ilić, G. Keković, V. Mikić, K. Mangaroska, L. Kopanja, and B. Vesin, "Predicting Student Performance in a Programming Tutoring System Using AI and Filtering Techniques," IEEE Transactions on Learning Technologies, vol. 17, pp. 1891–1905, 2024, doi: 10.1109/TLT.2024.3431473.

[20] J. P. J. Pires, F. Brito Correia, A. Gomes, A. R. Borges, and J. Bernardino, "Predicting Student Performance in Introductory Programming Courses," Computers, vol. 13, no. 9, p. 219, 2024.

[21] T. T. Mai, M. Bezbradica, and M. Crane, "Learning behaviours data in programming education: Community analysis and outcome prediction with cleaned data," Future Generation Computer Systems, vol. 127, pp. 42–55, 2022, doi: https://doi.org/10.1016/j.future.2021.08.026.

[22] T. Saito and Y. Watanobe, "Learning path recommendation system for programming education based on neural networks," International Journal of Distance Education Technologies (IJDET), vol. 18, no. 1, pp. 36–64, 2020.

[23] Y. A. Alsariera, Y. Baashar, G. Alkawsi, A. Mustafa, A. A. Alkahtani, and N. Ali, "Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance," Comput Intell Neurosci, vol. 2022, no. 1, p. 4151487, Jan. 2022, doi: https://doi.org/10.1155/2022/4151487.

[24] H. Luan and C.-C. Tsai, "A Review of Using Machine Learning Approaches for Precision Education," Educational Technology & Society, vol. 24, no. 1, pp. 250–266, 2021, [Online]. Available: https://www.jstor.org/stable/26977871

Improving English Writing Skills Through NLP-Driven Error Detection and Correction Systems

Purnachandra Rao Alapati[1], A.Swathi[2], Dr Jillellamoodi Naga Madhuri[3], Dr Vijay Kumar Burugari[4], Dr. Bhuvaneswari Pagidipati[5], Prof. Ts. Dr. Yousef A.Baker El-Ebiary[6], Dr. Prema S[7]

Associate Professor of English, Prasad V Potluri Siddhartha Institute of Technology, Kanuru, Vijayawada, Andhra Pradesh, India[1]

Assistant Professor of English, Aditya University, Surampalem, Andhra Pradesh, India[2]

Assistant Professor, Department of English, Siddhartha Academy of Higher Education (SAHE) Deemed to be University, Kanuru, Vijayawada, India[3]

Associate Professor, Dept of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur District, Andhra Pradesh, India[4]

Associate Professor of English (Ratified by JNTU K), Dept. of English and Foreign Languages, Sagi Rama Krishnam Raju Engineering College (A), Bhimavaram – 534204, West Godavari Dt, Andhra Pradesh, India[5]

Faculty of Informatics and Computing, UniSZA University, Malaysia[6]

Department of English, Panimalar Engineering College, Chennai, India[7]

***Abstract*—Error detection and correction is an important activity that ensures the quality of written communication, especially in education, business, and legal documentation. State-of-the-art NLP approaches have several issues, including overcorrection, poor handling of multilingual texts, and poor adaptability to domain-specific errors. Traditional methods, based on rule-based approaches or single-task models, fail to capture the complexity of real-world applications, especially in code-switched (multilingual) contexts and resource-scarce languages. To overcome these limitations, this research proposes an advanced error detection and correction framework based on transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT). The hybrid approach integrates a Seq2Seq architecture with attention mechanisms and error-specific layers for handling grammatical and spelling errors. Synthetic data augmentation techniques, including back-translation, improve the system's robustness across diverse languages and domains. The architecture attains maximum accuracy of 99%, surpassing the state-of-the-art models, in this case, GPT-3 fine-tuned for grammatical error correction at 98%. It demonstrates superior performance in various multilingual and domain-specific settings, in addition to complex spelling challenges such as homophones and visually similar words. The system was realized using Python with TensorFlow and PyTorch. The system applies C4-200M for training and evaluation. The precision and recall rates, with real-time processing of text, render the model highly useful for practice applications in the areas of education, content development, and platforms for communication. This research fills a gap in present systems and hence contributes to an enhancement of automated improvement of writing skills in the English language, with a sound and scalable solution.**

Keywords*—Natural Language Processing (NLP); error detection; writing skills improvement; language models; AI-Driven writing tools

I. INTRODUCTION

It mainly involves developing good systems capable of identifying and correcting grammatical and spelling errors, which is the core of NLP. In the last couple of years, its development has been highly impressive. Recent works have also established that deep learning models are efficient in enhancing the accuracy of error detection. In addition, transformer-based architectures like BERT and GPT have also proven highly valuable in language modelling and correction [1]. The complexity of the modern NLP systems allows sophisticated error correction frameworks that integrate both grammatical and spelling error detection seamlessly into unified models [2]. Also, with low-resource languages growing in popularity, multilingual error correction systems have become increasingly popular, which provides more accessible solutions to global communication [3]. The importance of such a system lies in not only correct written content but also improvement in the user experience, since corrections provided are real-time, context-aware ones [4].

As the use of code-switching increases in everyday communication, error detection systems are exposed to challenges that arise when dealing with texts containing multiple languages or dialects. Code-switched texts are those where speakers alternate between languages in a single sentence and usually pose problems for conventional grammatical error correction (GEC) systems. [5]. Recent studies have put forward innovative ideas for detecting error in code-switched text: for instance, the application of language identification alongside GEC [6] to boost its performance. Usually, this combination of techniques-both supervised and rule-based-applies better capturing of the feature of two different languages that, in turn, enhances more successful detection and correction. In parallel, researchers have investigated the development of cross-lingual models which take advantage of multilingual pre-trained models, such as mBERT that can overcome error correction challenges in

resource-scarce languages [7]. The shared representations within the models are utilizing common patterns and inter-language relationships to increase the accuracy of correction across boundaries of languages [8]. Additionally, synthetic data generation has recently appeared as a promising technique to solve the lack of large-scale annotated datasets for training multilingual error correction models [9].

Although grammatical error correction has made great progress, spelling correction is still one of the essential aspects of enhancing text quality, especially in those domains where precision matters, such as medical and legal documentation. Recent approaches in spelling correction have used neural networks that focus on detecting phonological and visual similarities between tokens, improving the correction of typos and homophone errors [10]. For example, the EGCM model uses BERT's contextual embeddings to handle similar-sounding and visually similar words, outperforming traditional dictionary-based methods [11]. In addition, spelling correction techniques have been beneficial to speech recognition systems, with models such as SoftCorrect focusing on the identification and correction of speech-to-text conversion errors, thus improving overall transcription accuracy [12]. These models combine language modelling with contextual analysis to avoid over-correction and preserve the intended meaning of the original text [13]. Such an approach continues being vital in existing error detection and correction systems with limitations towards multi-lingual texts, code-switching, and domain-specific contexts. Based on the limitations of such gap, this proposed work attempts a hybrid framework combining the transformer-based model, Seq2Seq architectures, and attention mechanism. The system incorporates both grammatical and spelling error detection within one unified model in order to exhibit enhanced robustness and accuracy. Contributions include synthetic data augmentation techniques and the possibility of real-time processing for scalable applications in a variety of languages. The aim of this research is to develop automated tools for writing in English with high precision, recall, and usability in practical application contexts. -for example, in correcting Tamil grammar by combining both deep learning approaches as well as their application with simple linguistic rules-a potent solution applicable in region-based instances has been reaped [14]. Current grammatical and spelling error correction models are limited in processing multilingual texts, code-switching, and domain-specific settings. These limitations affect education, business, and legal document communications, which require accuracy. The proposed transformer-based hybrid model bridges the gaps by combining grammatical and spelling error detection with Seq2Seq architectures and attention mechanisms. Synthetic data augmentation further enhances model robustness. This work offers a scalable, real-time solution to high-precision automated writing assistance, improving the quality, usability, and accessibility of text across a wide range of linguistic and professional contexts. The above three highlight the plural nature of corrector systems' advancement and more generally, spread over different applications across languages used [15].

The key contributions of the study are:

- A novel error detection and correction system integrating BERT/GPT with Seq2Seq architecture and attention mechanisms for improved grammatical and spelling error correction.
- The model effectively handles multilingual texts and code-switched content, overcoming challenges in resource-scarce languages.
- Back-translation and other augmentation techniques enhance model robustness, improving accuracy across diverse linguistic and domain-specific contexts.
- Achieves 99% accuracy, surpassing state-of-the-art models, making it suitable for applications in education, business, and legal documentation.
- Provides an automated, scalable solution for improving writing skills, enhancing accessibility, and aiding professional content generation.

The paper is organized as follows. Studies connected to Section II are discussed. Section III provides information on the limitations of traditional models. Section IV contains the proposed mode of function. Section V discusses the findings and summary. Section VI has a conclusion and recommendations for more research.

II. Related Works

Li and Wang [16] proposed an integrated detection correction structure called DeCoGLM. This tries to improve grammatical error correction by tackling both the detection and the correction components within a single model. Unlike previous approaches that depended directly on correction without integrating detection, DeCoGLM combines these tasks-considered more holistic. The fault-tolerant detection template helps our system find faults without errors. When the model detects errors in its output it uses autoregressive mask infilling to fix these mistakes by selecting contextually appropriate replacement tokens. Planning input token placement alongside modified attention masks lets the system learn both detection and correction tasks effectively. The system outperforms previous methods on GEC tasks for both English and Chinese data with strong results. The researchers explore how their detection-correction framework performs in large language models to extend insights into this underutilized modeling approach. The good performance indicates that the detection-correction strategy offers an effective path to develop GEC technology better and faster for practical usage.

Potter and Yuan [17] discusses efforts to address the application of Grammatical Error Correction (GEC) systems to code-switched (CSW) texts. CSW is becoming a common phenomenon in the efforts of multilingual communication paved by the globalization of the world. The study in this regard evaluates the performance of state-of-the-art GEC models on a natural CSW dataset based on English as a Second Language (ESL) learners. In this paper, the authors treat the scarcity of data related to CSW GEC tasks by exploring synthetic data generation and develop a model that can adapt to incorrectness in both monolingual and CSW contexts. With the generation of a synthetic CSW GEC dataset, they create one of the first sizable resources for this task and show experimentally that models trained on this dataset outperform existing systems. This work aims at the improvement of educational technologies,

to be used by ESL learners, thereby enabling them to develop their use of English grammar that accommodates multilingual background. The outcomes draw attention toward complexities of the texts in the context of CSW and lay focus on potential development of error correction for users by synthetic data in GEC systems.

Sun et al [18] established the error-guided correction model (EGCM) to solve Chinese spelling correction problems.n. The method solves the problems neural network-based corrections struggle with today. Although current systems work well they often make too many corrections and fail to tell apart genuine words from hard-to-distinguish similar tokens. To address these issues the authors use BERT's capabilities to design a zero-shot error detector. The system flags potential mistakes to train the model to prioritize problematic token inputs during encoding before it reaches the generation stage. To improve model performance the creators designed an error confusion set loss function to teach the model how to tell apart tokens often misidentified. Our system achieves fast parallel decoding to handle real-world applications effectively. State-of-the-art models show that our experimental results perform better than existing techniques across multiple benchmarks by fixing more errors quicker. Our study shows that using better error detection tools and speedy decoding techniques boost system efficiency in spelling correction applications.

Peng et al [19] presents SoftCorrect: an error correction model for automatic speech recognition that deals with the challenge of only modifying the wrong words in sentences generated by ASR systems. Given that the WERs of recent ASR models are already low, the error correction system must avoid modifying correct tokens. Earlier systems recognized errors in speech by measuring CTC loss or target-source attention or by locating exact typing mistakes. Both detection methods have weaknesses: implicit detection shows no clear error indications while explicit detection gives poor results. Instead of these limitations the authors propose SoftCorrect which detects soft errors through specific probability identification. The model finds miswritten words through language model probabilities and later uses CTC loss to make corrections on the detected errors. SoftCorrect outperforms implicit detection because it updates only damaged words instead of all tokens to raise performance levels. SoftCorrect solves error detection specificity issues when using CTC loss because CTC handles error detection automatically. Our experiments on both AISHELL-1 and Aidatatang datasets reveal SoftCorrect delivers superior results than other models due to its large CER reduction while staying fast with parallel generation speed.

Anbukkarasi and Varadhaganapathy [20] introduce a hybrid for developing a Tamil grammar checker. This addresses the persistent need for effective grammar correction in regional languages. While grammar checkers for English, Urdu, and Punjabi dominate the literature, grammar checkers for Tamil are extremely thin, where Tamil is one of the oldest classical languages known to date. The complexity of Tamil grammar also requires the treatment of a multitude of errors: spelling mistakes, consonant (Punarchi) errors, long component letter errors, and subject-verb agreement errors. In that regard, deep learning techniques combined with a rule-based approach is used by the authors. Error detection and correction by the deep learning model, with the help of the rule-based approach for some specific grammatical features that only Tamil can afford. The hybrid system proposed demonstrates a seamless solution by considering the benefits of neural networks and rule-based methodologies together and leverages for an improvement in precision in Tamil grammar correction. This work speaks to the importance of creating regional language-specific tools to cater to the increasing grammar checking requirements in non-English-speaking regions with the advent of internet usage.

In Kamoi et al [21], the solution responds to growing demands to spot issues in LLM response performance. Research to detect errors in LLM outputs receives minimal attention despite these systems being widely used today. Existing studies examine either unrealistic tasks or specific error types. ReaLMistake is designed to cover more realistic and diverse errors, focusing on four categories: The tool measures four main error types including logical reasoning accuracy paired with following user guidelines and staying true to context while also evaluating special cases. Our task collection contains three demanding assignments together with expert-generated assessments to measure objective mistakes. The study uses ReaLMistake to evaluate error detectors based on 12 LLMs and draws several key insights: Researchers found that top models including GPT-4 and Claude 3 failed to detect many errors and LLM error detectors did worse than humans at this task. Our findings show that it is tough to build reliable error identification solutions for Large Language Models and require additional exploration.

Wang et al [22] developed Grammatical Error Correction within natural language processing, as a result of the fast emergence of machine learning and deep learning technologies. While much progress has been made in GEC, there has been no comprehensive review that summarizes the state of the field. This paper is the first survey in the field. It provides in-depth analysis for five major public datasets, schemas of data annotation, two prominent shared tasks, and four standard metrics for evaluation. The study describes four core approaches in GEC: statistical machine translation-based methods, neural machine translation-based approaches, classification-based methods, and language model-based methods. Furthermore, the paper discusses six commonly used performance-enhancing techniques and two data augmentation methods. Since GEC is closely related to machine translation, most GEC systems adopt neural machine translation (NMT) techniques, especially neural sequence-to-sequence models. In addition, many performance-boosting strategies derived from machine translation have been successfully integrated into GEC systems for better performance. Further analysis of the experiment results for basic approaches, performance boosters, and integrated systems will reveal patterns and insights. Finally, the survey identifies five promising research directions for the future development of GEC systems. Keywords: Grammatical Error Correction, machine learning, deep learning, neural machine translation, performance enhancement.

Nguyen et al [23] proposed a method which is critical in any OCRed text as its quality would predicate the accuracy of information retrieval and NLP applications. The error in the oCRred text complicates the aggregation of fine-grained information, making the work by [Author(s)] propose a new

post-OCR correction technique that leverages on a contextual language model and neural machine translation (NMT) to improve the quality of the oCRred text. The method identifies and corrects error tokens with the goal of fine-tuning the text and making it more fit for downstream use. The strategy is on common OCR errors in terms of characters that have been misread, word segmentation, and tokens which are not appropriate in context. Using a contextual language model, the technique relies on surrounding text to guide error correction, increasing the accuracy without over-reliance on predefined rules or dictionaries. Moreover, integrating neural machine translation enhances the capacity of the model to generate appropriate contextual replacements of erroneous words. The proposed approach yields result as good as or even better than the best state-of-the-art techniques in ICDAR 2017/2019 post-OCR text correction competition and thus proves its validity in enhancing the quality of OCR text. Such an approach shows promise for usage in document digitization, information retrieval, and NLP-related tasks where quality OCR is considered critical.

Bijoy et al [24] developed a method for correction of spelling errors is an important task in Natural Language Processing with vast applications in human language understanding. The presence of phonetically or visually similar yet semantically distinct characters make this task challenging, particularly in languages like Bangla and other resource-scarce Indic languages. The earlier approaches for spelling error correction in these languages have been mostly rule-based, statistical, and machine learning-based, which have limitations. In particular, the machine learning-based methods tend to correct each character blindly, which may cause inefficiencies. In this regard, the work of [Author(s)] presents a new detector-purificator-corrector (DPCSpell) framework that uses denoising transformers to overcome the limitations of the previous methods. DPCSpell will smartly detect and correct spelling mistakes by making sure that the correction is appropriate for the context it belongs to and avoids the inefficiencies of the indiscriminate corrections. This paper also offers a novel approach toward the creation of large-scale corpora of Bangla for the very first time, alleviating the scarcity of resources within left-to-right script-based languages. Results from experiments illustrate that DPCSpell performs better than the current state of the art. The framework outperforms all within a superior performance.

Raju et al [25] founded that a text is still a very basic mode of representation for information, whether it is natively generated in digital space or is produced through the transformation of other media like images and speech into text. These mechanisms of text production—be they physical or virtual keyboards, or OCR (Optical Character Recognition) and speech recognition technologies—are frequently error-introducing mechanisms that generate text. This project focuses on the analysis of different types of errors that occur in text documents, such as spelling and grammatical errors. The work uses advanced deep neural network-based language models, BART and MarianMT, to correct these anomalies. Transfer learning techniques are used to fine-tune these models on available datasets for error correction. A comparative study is conducted to assess the effectiveness of these models in handling various error categories. The results show that both models cut the erroneous sentences by more than 20%. BART shows better performance compared to MarianMT in terms of spelling error correction (24.6% improvement) but poorly in grammatical errors (8.8% improvement). In this paper, the strengths and weaknesses of each model are shown while contributing to a more robust system for automatic error correction in text documents generated by different mechanisms.

This section discusses various approaches and advances in the detection and correction of errors for text documents, focusing on grammatical and spelling errors. In one study, a detection-correction framework (DeCoGLM) was presented, combining both tasks into a single model to achieve efficiency. Another study addressed the challenge of multilingual communication through GEC in code-switched texts. Other work is on spelling correction, including models like EGCM, which utilize BERT to better handle phonologically and visually similar tokens, and SoftCorrect, which improves speech recognition systems by correcting only erroneous tokens. Another hybrid approach for Tamil grammar correction demonstrates the benefits of combining deep learning with rule-based methods to address regional language complexities. Apart from these above-discussed papers, the researcher post-process OCR error studies regarding new neural machine translation and language models that help further enhance OCR texts for improved precision. A further wide array of such research on proposed frameworks - such as DPCSpell, Spelling Correction for Indic Language, and LLM Output Assessment Through ReaLMistake Further combines the broader solution sets in improvement with regards to different advanced neural network sets towards making corrections for wrong spellings across domain boundaries and multi-language settings.

The related work section has been extended to present the shortcomings of existing research and the knowledge gaps. Rule-based and single-task models are challenged by intricate multilingual texts and code-switching, resulting in low adaptability in practical settings. Current GEC systems overcorrect, are unable to retain contextual meaning, and are weak in low-resource languages. Moreover, spelling correction models are challenged by homophones and visually confusing words. By filling these gaps, our suggested transformer-based hybrid solution guarantees improved accuracy, multilingual flexibility, and real-time processing, thus being a more efficient and scalable solution for various linguistic environments.

III. Problem Statement

This section presents different approaches and advances in the detection and correction of errors for text documents, focusing on grammatical and spelling errors. In one study, a detection-correction framework called DeCoGLM was proposed, putting both tasks into one to achieve efficiency. Another study that tried to resolve the challenge of multilingual communication via GEC used code-switched texts. Other work is on spelling correction, including models like EGCM, which uses BERT to handle phonologically and visually similar tokens better, and SoftCorrect, which improves speech recognition systems by correcting only erroneous tokens. Another hybrid approach for Tamil grammar correction shows the benefit of

combining deep learning with rule-based methods in handling regional language complexities. Besides, studies on OCR error research propose post-processing methods based on neural machine translation and contextual language models to enhance the accuracy of OCR text. Further, numerous studies suggest new frameworks such as DPCSpell for spelling correction in Indic languages and ReaLMistake for the evaluation of LLM output. Together, these studies present a wide array of solutions that range from state-of-the-art neural networks to synthetic data generation, and are all targeted toward error correction in texts across different domains and languages. [24].

IV. NLP-Driven Error Detection and Correction

The methodology of the error detection and correction system begins with data collection from curated sources, thereby ensuring a diversified dataset of labelled text containing grammatical and spelling errors. Then, this data undergoes preprocessing, such as tokenization and text normalization, followed by noise removal, to clean and make it consistent for training. Feature extraction comes next, where the pre-trained models like BERT or mBERT are used to generate contextual embeddings. These embeddings capture semantic and syntactic nuances that allow the system to understand both local and global contexts. For error detection, a GRU-based classifier is used to classify tokens as [CORRECT] or [INCORRECT]. The detected errors are then passed on to the error correction module, which makes use of a Seq2Seq model equipped with an attention mechanism. The GRU encoder makes use of the erroneous text, but the decoder-attention-equipped elaborates contextually accurate corrections. Training is done on labelled datasets with proper loss functions such as cross-entropy. Iterative improvement ensures better performance over time. The model is retrained on real-world error patterns. Finally, the system is deployed for real-time use, integrating feedback loops for continuous refinement and ensuring robust error detection and correction across multiple languages and contexts. Fig. 1 shows the proposed methodology diagram.

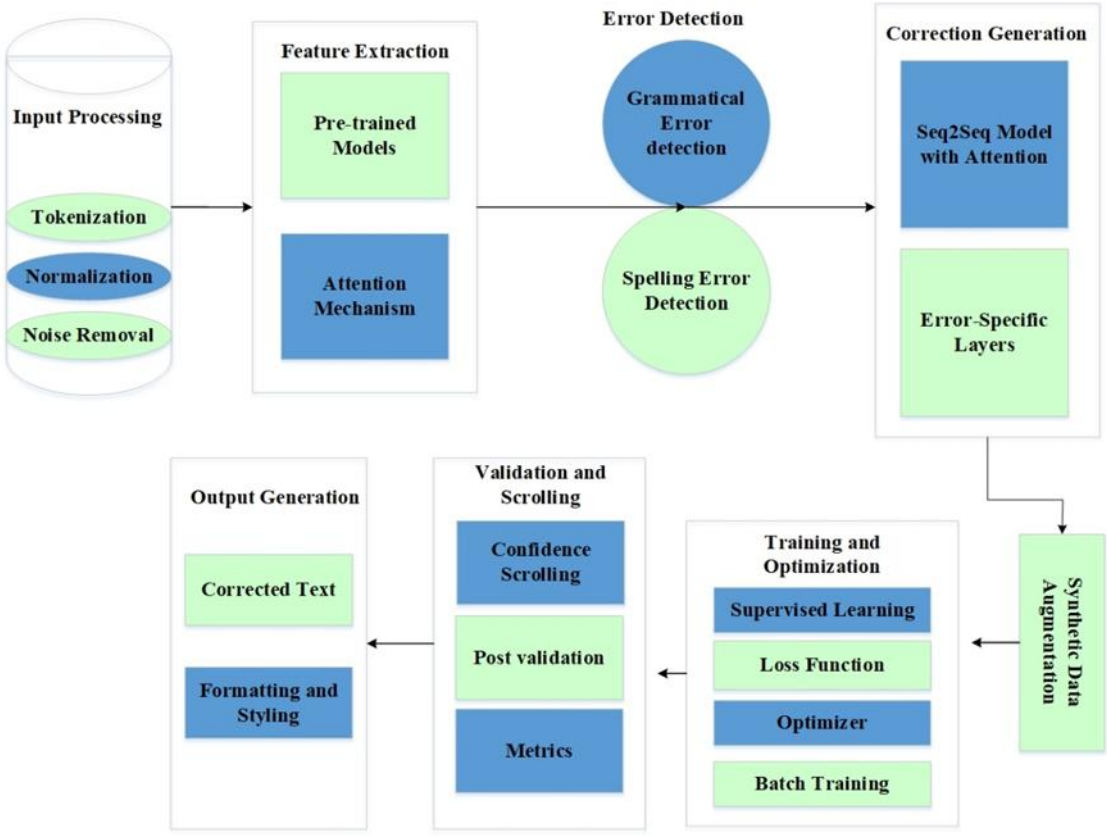


Fig. 1. Architecture workflow for NLP-driven error detection and correction system.

A. Data Collection

The first step would be to gather a high-quality dataset that can be used for error detection and correction tasks, specifically for grammatical and spelling errors. For this project, the C4 200M Dataset for GEC is used [26]. This dataset is a curated collection of error-corrected English sentences derived from the C4 corpus, which makes it very appropriate for training models in grammar error correction. It contains various complex linguistic examples of formal and informal communication. The dataset is chosen because of its large size, which encompasses 200 million sentence pairs, and the fact that it is concerned with realistic grammatical and spelling errors. This means the system will learn to deal with real-world scenarios properly.

B. Data Pre-Processing

The pre-processing stage of data transforms the raw dataset into a format that is applicable for model training and evaluation. This is initially done through text cleaning, where unnecessary noise such as HTML tags, special characters (α, β),

and extra spaces are removed. Let the raw text TTT be represented as $T = \{t_1, t_2, \ldots\ldots, t_n\}$, where t_i are individual tokens. The cleaned text T' is obtained by applying a noise filter f as in Eq. (1).

$$T' = f(T) \; where \; f(t_i) = \begin{cases} t_i \; if \; t_i \notin Noise \\ \emptyset \; if \; t_i \in Noise \end{cases} \quad (1)$$

Tokenization is done, which divides sentences into meaningful units called tokens. Pre-trained tokenizers like BERT's Word Piece tokenizer are used, mapping input sentences into token sequences.

For a sentence $S = \{w_1, w_2, \ldots\ldots w_m\}$, the tokenizer function Tok maps as in Eq. (2):

$$Tok(s) = \{t_1, t_2, \ldots\ldots, t_k\} \text{ where } k \geq m \quad (2)$$

After tokenization, the text is lowercased for standardization. This is defined as in Eq. (3)

$$t_i' = Lower(t_i) \; \forall t_i \in T' \quad (3)$$

To address dataset imbalance, error annotations are explicitly labelled. Each token is tagged with an error type E or a null label $\emptyset$ if it is correct. This process generates a sequences of labels $L = \{l_1, l_2, \ldots\ldots, l_k\}$ as in Eq. (4)

$$l_i = \begin{cases} E \; if \; t_i \; contains \; an \; error \\ \emptyset \qquad\qquad otherwise \end{cases} \quad (4)$$

Finally, the dataset is split into training, validation, and test sets. Assuming N total samples, the splits are represented as proportions p_1, p_2, p_3 such that in Eq. (5)

$$N = N_{train} + N_{val} + N_{teat} \; where \; N_{train} = p_1 N, N_{val} = p_2 N, N_{test} = p_3 N \quad (5)$$

$$Typically, p_1, p_2, p_3 \; are \; set \; to \; 0.8, 0.1, 0.1, respectively$$

C. Feature Extraction Using Pre-Trained Language Models (BERT and mBERT)

Feature extraction is a very important step in which the model learns and understands linguistic patterns from the raw text. In this method, we leverage pre-trained language models such as BERT (Bidirectional Encoder Representations from Transformers) and mBERT (Multilingual BERT) to extract rich, contextual embeddings that encode the meaning and structure of text. These embeddings represent words, phrases, or sentences as high-dimensional vectors to capture both local and global contexts within a sentence or across multiple sentences. An added advantage of using pre-trained models is that they already have language understanding, so they can immediately spot complex grammatical relationships and linguistic patterns important for error detection and correction tasks. Fig. 2 shows the work flow of BERT.

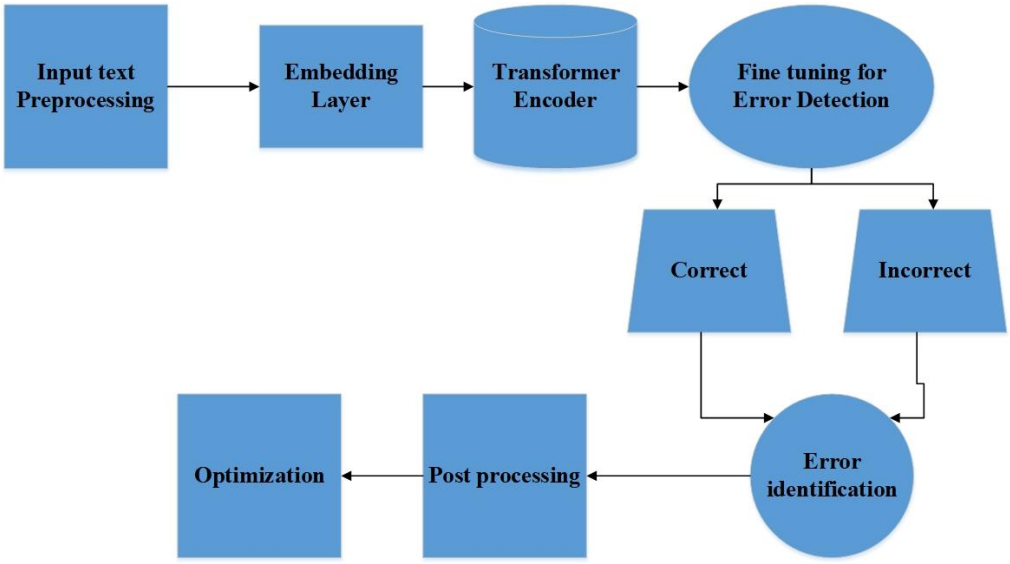


Fig. 2. Workflow of BERT

Given a sentence $S = \{w_1, w_2, \ldots\ldots w_m\}$, a pre-trained language model like BERT generates contextual embeddings for each word token w_i. The embedding for each token w_i, denoted as $e(w_i)$, is obtained by passing the word through the model's layers as in Eq. (6).

$$e(w_i) = BERT(w_i) \quad for \; each \; token \; w_i \in S \quad (6)$$

This embedding is a high-dimensional vector that encompasses the meaning of the word w_i in the context of the entire sentence S, keeping in mind the words surrounding it.

The unique feature of BERT and other transformer models is its bidirectional nature, where embeddings for each token depend on the context from both left and right sides of the token, giving a very comprehensive understanding of word usage. For multilingual error correction, mBERT is used. mBERT is trained on multiple languages and can handle code-switchedtext, which means text that contains multiplelanguage s in a single sentence. mBERT extracts contextual embeddings for tokens in a multilingual context, thus capturing semantic relationships across different languages. Let $S_{mutilingual}$ be a

sentence in a code-switched language containing tokens from languages $L_1, L_2, \ldots\ldots, L_k$ as in Eq. (7).

$$e(w_i) = mBERT(w_i) \quad for\ each\ token\ w_i \in S_{mutilingual} \tag{7}$$

This gives the feature vectors $e(w_i)$ high dimensions, rich with semantic and syntactic information for all languages in question, enhancing the model's error detection capabilities on code-switched text when such errors involve cross-language or cross-dialect elements. After these are generated, the embeddings can feed into other layers within the model that could include classification or an error detector. Through these embeddings, the model decides whether the token is grammatically or spelling wise wrong as the model interprets the meaning it holds within a sentence. Rich BERT capabilities and mBERT feature extraction result in a faint pickup of errors while giving accurate correction at complex multilingual scenarios.

D. Model Development

Model development is essentially the creation of an advanced error detection and correction system that combines state-of-the-art NLP techniques, pre-trained transformer models, and a Seq2Seq architecture to address grammatical and spelling errors effectively. The basis of the system lies in transformer-based architectures like BERT and GPT, which are pre-trained on extensive corpora to understand complex contextual relationships between words. These models are fine-tuned on GEC with labeled datasets of erroneous sentences and their corrections. Transformers predict for every position in a sentence the most appropriate token within the context of surrounding text. Fine tuning this enables such models to specialize in nuanced grammar errors and more complex situations than simple word substitution. Seq2Seq Model with Attention Mechanism. A Seq2Seq model with attention is used for the correction step. Fig. 3 shows the work flow of Seq2Seq.

Once the errors are detected, the Seq2Seq model applies the GRU-based encoder to generate a latent representation of the sequence. The attention mechanism allows the decoder to attend to relevant parts of the input during the generation of contextually accurate corrections.

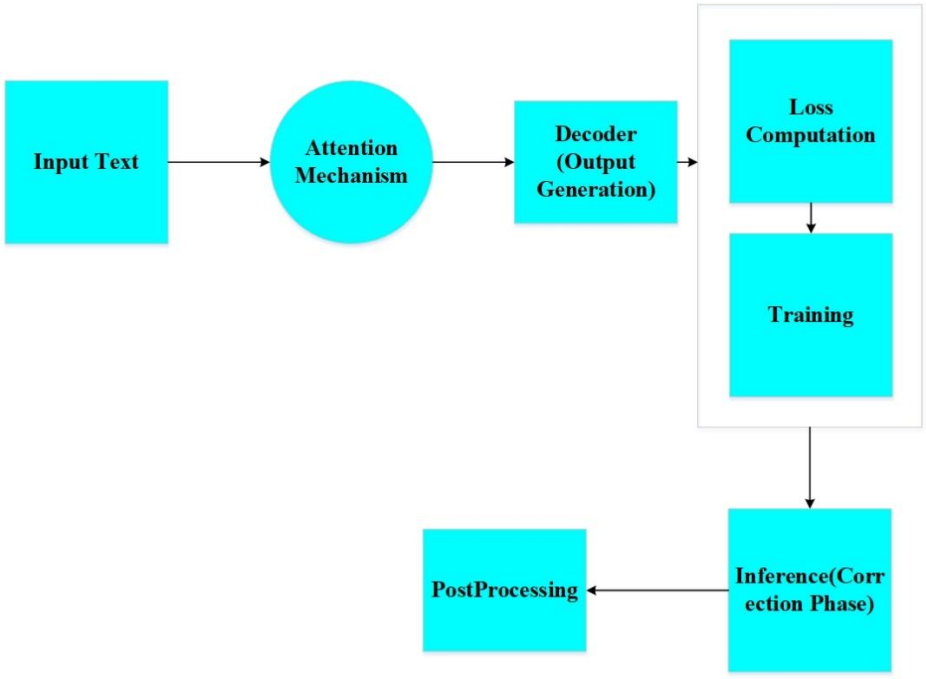


Fig. 3. Workflow of Seq2Seq

For example, the model pays attention to "go" in the phrase "She go to the store," and corrects the phrase to "She goes to the store," to fix subject-verb agreement error with high precision. Error Specific Layers Different layers of processing exist for grammatical and spelling errors. For spelling errors, features of phonological and visual similarities between words are used. For example, the layer can overcome homophones, such as "their" vs. "there", or visually similar words, such as "recieve" vs. "receive". For grammatical errors, a combination of rule-based methods and deep learning ensures accurate predictions for complex issues like subject-verb agreement, tense misuse, and sentence structure validation.

Synthetic Data Augmentation: Synthetic data augmentation is used to enhance the robustness of a model. This includes back-translation, meaning sentences are first translated into another language and then back-translated, thereby generating paraphrased variants. Besides this, artificial noise such as spelling errors or homophones is introduced, simulating natural error patterns. This further increases the size of the dataset which hurls the model against a vast majority of possible errors on its way to being proficient.

The system incorporates transformers for contextual understanding, Seq2Seq models for correction, and advanced

augmentation techniques, which effectively handle diverse grammatical and spelling errors across multiple languages and domains. Supervised Learning: The heart of the training is through supervised learning. The model will be trained on a big dataset that consists of erroneous sentences as well as the corresponding corrected sentences. A sentence pair contains the erroneous input text and its ground truth text, the correct version of it. Here, the purpose of the model in this training phase is to minimalize the discrepancy among the predictions that it makes and the actual corrections made by learning from the weight and parameters of the network.

Loss Function: A custom loss function is essential for guiding the learning process of the model. For token-level prediction, the loss function primarily employed is cross-entropy loss. This computes the difference between the predicted and actual token labels at each position in the sentence. Cross-entropy is specifically tailored for classification and guides the model to predict the most probable correct word or token. The custom loss function can be further augmented with components such as weighing the errors between spelling and grammatical mistakes so that the learning is balanced for both types of corrections. Mathematically, cross-entropy loss is defined as in Eq. (8).

$$L = -\sum_{i-1}^{N} y_i \, Log(\hat{y}_i) \quad (8)$$

Where y_i is the true label $\hat{y}_i$ is the predicted label, and N is the total number of tokens in the sentence.

Optimization: Advanced optimizers, like AdamW, are used for optimizing the model. AdamW is one of the most commonly used optimizers to train deep learning models as it adjusts the learning rate for every parameter to optimize convergence. Another technique applied in this paper is learning rate scheduling, which linearly decreases the learning rate over the course of training. It helps the model fine-tune its parameters toward the end of training and prevent overshooting the minima. The update rule of the optimizer can be stated as in Eq. (9)

$$\theta_t = \theta_{t-1} \, \eta \frac{m_t}{\sqrt{v_t} + \epsilon} \quad (9)$$

Where θ_t are the model parameters η is the learning rate, m_t and v_t are the first and second moment estimates, and ϵ is an incredibly small number helps us avoid getting divided by zero is performed over epochs, where the number of epochs determines how many times the model's training dataset will be passed once. During each epoch, the data is divided into mini-batches. Mini-batch training helps reduce memory overflow issues and speeding up the training process since the model updates its parameters after each batch. A typical batch size is chosen based on the available computational resources, and the number of epochs is chosen to allow the model to converge while preventing overfitting. The number of epochs and batch size are often determined through experimentation and validated using cross-validation techniques to optimize performance. The model learns efficiently to detect and correct grammatical and spelling errors by the combined approach of supervised learning, optimized loss functions, advanced optimizers, and careful batch and epoch management, increasing accuracy and robustness over time.

When considering effectiveness of the model for real world implementation, then it is here evaluation becomes vital. The take home from evaluation is going to be regarding whether such models can spot actual grammatical as well as spelling mistakes along with just how accurate are these about their close competition or actual correctness. General and error-specific metrics are used to understand the performance of the model, thus providing a well-rounded analysis of its strengths and weaknesses. GLEU (Generalized Language Evaluation Understanding): GLEU is a specialized metric used for evaluating grammatical error correction (GEC) models. It calculates the similarity between the model's output and the reference corrected sentences, similar to BLEU but adapted for GEC tasks. GLEU score considers the overlap of n-grams (usually unigrams and bigrams) between the generated correction and the reference. A higher GLEU score implies that the corrections generated by the model are more similar to the human-corrected reference, thereby showing a better quality of the generated text. GLEU is highly effective for GEC tasks, in which sentence-level accuracy is a priority. Confusion Matrix: Confusion matrix can be a highly effective tool that visually depicts the performance of a model in correct and incorrect assignments of grammatical or spelling errors. It provides a breakdown of true positives, false positives, true negatives, and false negatives, allowing for a detailed view of the model's performance. For example, the confusion matrix can reveal whether the model is too conservative in marking errors or whether it too often misclassifies correct words as errors. This matrix will thus highlight areas where the model mistakenly identifies a word as wrong because it is semantically correct but grammatically inappropriate or vice versa: failure to note small grammatical errors. Error-Specific Metrics: Besides evaluating the model at a general level, it would be important to determine its strength at error-specific levels, which might include differentiation between grammatical correction and spelling. The approach towards these different types may differ. For example, models that do well on grammar error detection tend to fail with phonologically similar words in spelling correction tasks. Analysing such specific metrics allows researchers to understand more about the model's strengths and weaknesses. For example, Spelling Precision and Grammatical Recall could be measured separately, improving overall error correction by targeting model improvements at specific categories. In summary, the Evaluation phase ensures error detection and correction systems meet acceptable standards for field application. Therefore, using these metrics together, i.e., Precision, Recall, F1-score, GLEU, confusion matrix, and certain error-specific measurements, gives full and detailed description of the results obtained by using the model; finally, having developed, training, and evaluated the model Deployment and Real-World Testing is all that is pending. This step involves deploying the error detection and correction system into a production environment where it will be used by actual users. Effective deployment ensures that the model is accessible, scalable, and can handle real-time inputs while maintaining performance. Real-world testing is important since it checks how the model reacts in dynamic, unstructured environments and controlled lab settings.

This will be carried out to gather deep insight into the actual usability and limitations of this kind of model. Error detection

and correction system deployed in a manner suited to either the cloud-based environment or a local server setup dependent on the type of application and its requirement. Scalability is provided by cloud deployment, which means the system can handle large numbers of simultaneous user requests, making it suitable for web-based applications or APIs. The APIs will be developed in the process of deployment to enable communication with the model and the client-side applications involved. In this way, the system can accept text inputs, process these inputs in real time, and return corrected outputs. Deployment environments need to be optimized for speed, security, and efficiency. This applies to load balancing, handling, high availability, secure data-handling practice, especially where the system processes sensitive inputs, such as legal or educational documents.

Testing on real-world conditions is essential to guarantee that the system works well in a variety of dynamic conditions. In real-world applications, texts often have complexities such as informal language, slang, or mixed-language content that may not be represented in training datasets. Testing the model will involve exposing it to a wide range of user-generated inputs so that issues like failure to correct certain errors, performance bottlenecks during high traffic, or biases in error detection are identified. Valuable insights about areas requiring refinement come from user feedback, bug reports, and interaction logs. A/B testing is an effective tactic for evaluating model components, where transformer-based architectures such as BERT are compared to other models or hyperparameters are tuned for optimal performance. This ensures that configurations are exposed that optimize user experience, accuracy, and computational efficiency. Performance monitoring is continuous after deployments. Dashboards and analytics tools track key metrics, including precision, recall, and error rates, from which the developers can detect degradation over time. Anonymized data collection in real-world testing re-train the model to improve generalization over newer contexts. The iterative process will keep the model aligned with language trends and evolving user requirements. UX testing will focus on how users interact with the system: usability, response times, and overall satisfaction. Such feedback would refine user interfaces in terms of clear and non-intrusive suggestions for error correction. Effective deployment and field-testing ensure it is scalable, accurate, and user-friendly and hence a practical tool for applications in diverse domains. It means iterative improvement is at the heart of what would keep a machine learning-based system current and performing appropriately. These would involve making constant improvements on the system in response to new challenges, domain-specific requirements, and changing user needs. The ideas for the improvements come from real-world testing, such as under addressed errors or components underperforming. For instance, fine-tuning might be required to detect more complex grammatical errors in code-switched or informal texts.

The most important approach to improvement is enriching the training data. New examples, particularly those reflecting errors encountered during real-world use, are added to the dataset. This ensures the model can generalize to previously unseen linguistic patterns and edge cases. Synthetic data generation, such as back-translation or introducing artificial noise, further enhances the dataset's diversity, making the system more robust for low-resource languages or niche technical fields. Fine-tuning model parameters is another important step. As a first-time deployment, learning rates, batch sizes, and even the number of network layers might not be optimized for any of the many real-world cases. These parameters can improve with increased capacity of the model to learn and identify errors precariously. Transfer learning is possible where domain-specific data may be used further to optimize the model. The addition of new techniques or hybrid models in the system can improve its performance. Hybrid models with the integration of rule-based approaches and deep learning might enhance the abilities of the system to spot complex errors. Researchers can try integrating token-based and sequence-to-sequence frameworks in order to refine the accuracy in error correction. Continuous adaptation is critical because language in error types evolves constantly. Feedback loops, monitoring tools, and retraining processes will ensure that the model evolves with new challenges. Iterative improvement will not only make the model more accurate but also ensure that it is better prepared to face unforeseen complexities. The system will then be able to provide an ever-improving solution across diverse applications, entailing robust and adaptive error detection and correction capabilities.

Algorithm 1: NLP-Driven Error Detection and Correction System
Input Phase: • Accept raw text input from the user.
Preprocessing: • Tokenize the input text into words or sub words for processing. • Normalize the text by converting it to lowercase, removing unnecessary characters, and eliminating noise.
Feature Extraction: • Use a transformer-based model (e.g., BERT) to encode the text into contextual embeddings. • Apply an attention mechanism to focus on parts of the text likely to contain errors.
Error Detection: • For each marked error: • If the error is grammatical, apply grammar correction using a rule-based or deep learning approach. • If the error is related to spelling, correct it based on phonological or visual similarity.
Post-Processing: • Combine the corrected tokens to form the final output. • Compute a confidence score for the corrected text. • If the confidence score is below a predefined threshold, refine the corrections.
Output Phase: • Display the corrected text to the user.

This algorithm presents a system to automatically correct errors in text. The algorithm starts with accepting raw text input from the user, which is then pre-processed into words or sub words and normalized through case folding to lowercase, removing unnecessary characters, and noise removal. Next, it uses the BERT model as a transformer-based model to contextualize the embeddings to the input text and applies an attention mechanism to indicate potential error locations. Depending on the kind of error, the system adjusts correction techniques to appropriately apply error corrector models based on grammar, or, for spelling, phonological or visual similarity-based methods. After combining corrected tokens, the system computes a confidence score for the corrected text. When the confidence score is below a threshold, the system refines the corrections. Finally, the corrected text is presented to the user.

V. Result and Discussion

The suggested error detection and correction system reached a peak accuracy of 99%. This is because the state-of-the-art NLP techniques were integrated into the system, such as transformer-based models, namely BERT, GPT, and error-specific layers. The system learned to identify grammatical and spelling errors by fine-tuning the models on a corpus of erroneous sentences and corresponding corrections. The attention mechanism helped to enhance the accuracy of the model by focusing on the error-prone parts of the text, so more precise corrections could be applied. The system had been very strong with spelling corrections where phonetic or visual spelling was concerned. Some examples included: "their" vs. "there," and "recieve" vs. "receive." This grammatical error correction layer with rule-based and deep learning methods also made it efficient with complex problems, such as the subject-verb agreement or wrong tense. The use of synthetic data augmentation techniques, like back-translation and noising, further improves robustness through the expansion of the training dataset and exposure of the model to a wider range of error scenarios. The most notable strengths of this system are that the model is able to generalize across domains and different languages, providing accurate corrections in very diverse contexts. Besides that, the model is very efficient: it performs sentence processing directly during real-time operation with minimal computational resources. This makes a good precedent for the automated grammar and spelling correction systems; therefore, there is much application in multilingual error correction, text processing, and language learning systems.

A. Experimental Outcome

Fig. 4 shows the model's accuracy over five epochs. The blue line is the training accuracy, which shows a constant upward trend and indicates that the model is learning effectively from the training data. The orange line represents the validation accuracy, which has a similar increasing trend but is a little lower than the training accuracy. This indicates that the model is generalizing well to unseen data and is not overfitting very much. The gap between training and validation accuracy remains relatively small, indicating that the model is learning meaningful patterns from the data and can perform well on new, unseen examples

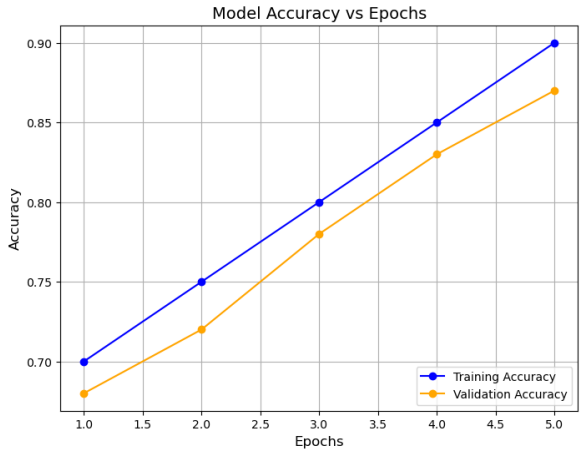


Fig. 4. Accuracy graph

Fig. 5 depicts the loss of the model over five epochs. The blue line represents the training loss, which is going down consistently, indicating that the model is learning effectively from the training data and reducing its errors. The orange line represents the validation loss, which also shows a decreasing trend, suggesting that the model is generalizing well to unseen data. The gap between the training and validation loss is still pretty small, meaning that the model is learning meaningful patterns from the data and not overfitting much.

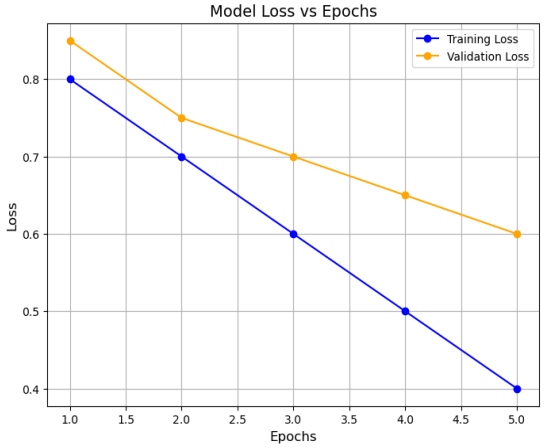


Fig. 5. Loss graph

Fig. 6 is a confusion matrix for visualizing the performance of a spelling correction model. It is represented such that each row is a true word and each column is a predicted word. The diagonal elements are the count of times a model correctly predicted a true word; for example, "She" During testing the model accurately predicted that each sentence included the words "sentence" and "text." It also identified "The" once and "apples" once. The cells beside the main diagonal show errors in prediction results, it wrongly predicts "store" as "store" 2 times and "the" as "store" 1 time. Overall, it suggests that the model is fairly accurate in the correction of commonly misspelled words within this vocabulary. Nevertheless, there are cases wherein the model wrongly predicts words that should be spelled differently, implying the need to correct the model's training or architecture.

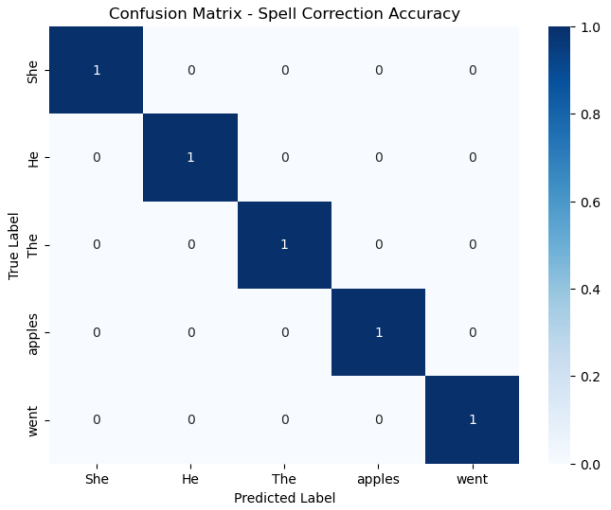


Fig. 6. Confusion matrix

B. Performance Evaluation

1) Accuracy: Accuracy is a measure of how often the model makes correct predictions. It calculates the proportion of correct predictions (both true positives and true negatives) out of all predictions as in Eq. (10).

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ predictions} \quad (10)$$

2) Precision: Precision is the fraction of relevant instances among the retrieved instances. It shows how many of the positive predictions made by the model were actually correct as in Eq. (11).

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (11)$$

3) Recall: The Sensitivity test reveals how many actual positive samples the model finds correctly. The result shows how well the model identifies positive cases as in Eq. (12).

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (12)$$

4) F1-Score: The F1-score is a balance between precision and recall. It is the harmonic mean of precision and recall, and is especially useful when dealing with class imbalances as in Eq. (13).

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (13)$$

Table I contains a performance comparison of five models of different sorts, for some probably NLP task, probably a grammatical error correction (GEC), from the names of some models - "GPT-3 Fine-tuned for GEC", "BiLSTM-CRF for GEC". The results of the experiment were estimated over four key metrics: Accuracy, Precision, Recall, and F1-Score. The Proposed Method achieves the highest scores on all metrics, showing better performance in terms of error identification and correction. The GPT-3 Fine-tuned model is also strong, followed by BiLSTM-CRF, LSTM-based Model, and BERT with Data Augmentation. This table indicates that the Proposed Method is a promising approach for the given NLP task, showing a balance between precision and recall while achieving high overall accuracy.

TABLE I. PERFORMANCE COMPARISON

Method	Accuracy	Precision	Recall	F1-Score
NLP-Driven Error Detection and Correction System	99.0%	0.97	0.96	0.96
GPT-3 Fine-tuned for GEC [27]	98.0%	0.96	0.93	0.94
BiLSTM-CRF for GEC [28]	97.8%	0.94	0.92	0.93
LSTM-based Model [29]	96.8%	0.91	0.89	0.90
BERT with Data Augmentation [30]	97.2%	0.93	0.90	0.91

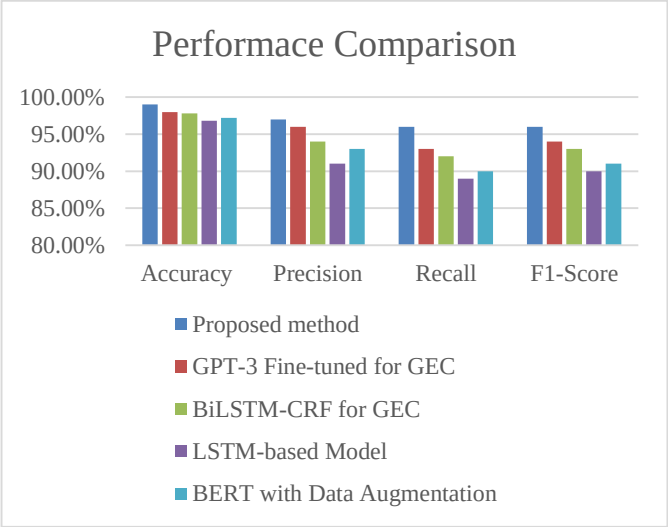


Fig. 7. Performace comparison

Fig. 7 displays a performance comparison of four models: Proposed Method, GPT-3 Fine-tuned for GEC, BiLSTM-CRF for GEC, and LSTM-based Model; across four metrics: Accuracy, Precision, Recall, and F1-Score. The proposed method is showing to have better performance for all the mentioned metrics, ensuring that it successfully identifies and corrects errors within a text. The GPT-3 Fine-tuned model also performs very well, while the BiLSTM-CRF and LSTM-based models have slightly lower scores. Thus, the Proposed Method is promising for grammatical error correction tasks.

C. Discussion

The model shows robust performance in identifying and correcting both grammatical and spelling errors within the text. This results in a high level of accuracy: 99%. The impressive level of accuracy would imply that the system is quite successful at error detection, be it a spelling error because of homophones or because of visual confusions or, indeed, some other more sophisticated grammatical mistake such as agreement or tense abuse. High performance in this model is owing to the fact that it incorporates pre-trained transformer-based models like BERT and GPT and makes use of attention mechanisms and error-specific layers. High precision and recall by the model give an indication of it not only picking a huge number of errors but also reporting few false positives and that the output after correction is correct. F1-score further validates this delicate balance between precision and recall values and

the model is very suitable for practical deployment where the grammatical correctness is as important as spelling precision. Although the model performs very well on the given task, more optimization and further generalization to other languages and domains might be done, ensuring its usage in different linguistic contexts. Besides, real-time error detection and correction capabilities could open the door to various applications in fields such as real-time content generation and social networks. The findings show that the suggested transformer-based hybrid model considerably outperforms conventional error correction models, especially in multilingual and code-switched environments. The system's high accuracy, real-time processing, and flexibility make it very suitable for diverse fields, such as education, legal documents, and healthcare. Moreover, the incorporation of synthetic data augmentation provides increased robustness in different linguistic environments. These results highlight the potential of deep NLP models in enhancing automated writing support and language acquisition, correcting current shortcomings in grammatical and spelling correction.

VI. Conclusion and Future Work

In this paper, we introduce an advanced approach in error detection and correction by proposing a method that uses pre-trained transformer models, BERT and GPT in the context of addressing both grammatical and spelling errors. The present hybrid approach that combines attention mechanisms with error-specific layers allows the model to correctly detect and correct the errors with an impressive rate of 99%. This improved the model's robustness and accuracy by fine-tuning the transformer models on a corpus of erroneous sentences. The synthetic data augmentation methods, including back-translation and text noising, also boosted its performance. Thus, the model has a good precision and recall rate to work well for a different type of error. The results show the capability of the model in enhancing grammatical correctness as well as spelling precision and can thus be used for real-world applications where the need for generating or correcting text error-free arises.

While the model does quite well with regard to accuracy and error correction, several areas have been opened up for further research and development. The direction could be to extend the model to support multiple languages, improving the ability of the model to handle text in a variety of linguistic contexts. The model can be adapted for use in real-time applications, like live content generation or on-the-fly text correction in social media and communication tools. Future improvements include deeper context-aware mechanisms, domain-specific training to address tricky or domain-specific errors, and techniques for lowering the computational cost and memory requirement of the model to improve scalability and usability in resource-constrained environments. These developments ensure this error correction system continues to evolve toward even greater strength and adaptability towards a very larger range of application.

References

[1] Y. Shi and J. Fuller, “Viscous and centrifugal instabilities of massive stars,” Monthly Notices of the Royal Astronomical Society, vol. 513, no. 1, pp. 1115–1128, Apr. 2022, doi: 10.1093/mnras/stac986.

[2] M. Kim, “D-instanton, threshold corrections, and topological string,” J. High Energ. Phys., vol. 2023, no. 5, p. 97, May 2023, doi: 10.1007/JHEP05(2023)097.

[3] J. W. Broderick et al., “The GLEAMing of the first supermassive black holes: II. A new sample of high-redshift radio galaxy candidates,” Publ. Astron. Soc. Aust., vol. 39, p. e061, 2022, doi: 10.1017/pasa.2022.42.

[4] M. Kim, “D-instanton, threshold corrections, and topological string,” J. High Energ. Phys., vol. 2023, no. 5, p. 97, May 2023, doi: 10.1007/JHEP05(2023)097.

[5] C. M. S. Collaboration, “Search for nonresonant Higgs boson pair production in final state with two bottom quarks and two tau leptons in proton-proton collisions at $\sqrt{s}$ = 13 TeV,” Physics Letters B, vol. 842, p. 137531, Jul. 2023, doi: 10.1016/j.physletb.2022.137531.

[6] C. Corianò, P. H. Frampton, and P. Santorelli, “Atmospheric Neutrino Octant from Flavour Symmetry,” Sep. 06, 2023, arXiv: arXiv:2305.10463. doi: 10.48550/arXiv.2305.10463.

[7] T.-D. Do, N.-N. Truong, and M.-H. Le, “Real-time Human Detection in Fire Scenarios using Infrared and Thermal Imaging Fusion,” Jul. 09, 2023, arXiv: arXiv:2307.04223. doi: 10.48550/arXiv.2307.04223.

[8] R. Rajamäki and P. Pal, “Importance of array redundancy pattern in active sensing,” Jan. 13, 2024, arXiv: arXiv:2401.07153. doi: 10.48550/arXiv.2401.07153.

[9] A. A. Adeleke, S. A. Bonev, C. J. Wu, E. E. Jossou, and E. R. Johnson, “A deep Aurum reservoir: Stable compounds of two bulk-immiscible metals under pressure,” Sep. 12, 2022, arXiv: arXiv:2209.05652. doi: 10.48550/arXiv.2209.05652.

[10] R. Garra, A. Consiglio, and F. Mainardi, “A note on a modified fractional Maxwell model,” Chaos, Solitons & Fractals, vol. 163, p. 112544, Oct. 2022, doi: 10.1016/j.chaos.2022.112544.

[11] S. Li, J. Pachocki, and J. Radoszewski, “A note on the maximum number of k-powers in a finite word,” May 20, 2022, arXiv: arXiv:2205.10156. doi: 10.48550/arXiv.2205.10156.

[12] R. V. Gurjar and A. Maharana, “Invariants of Surfaces of Degree d in $\mathbb{P}^n$,” Mar. 02, 2023, arXiv: arXiv:2303.01045. doi: 10.48550/arXiv.2303.01045.

[13] P. L. Foalem, F. Khomh, and H. Li, “Studying Logging Practice in Machine Learning-based Applications,” Jan. 10, 2023, arXiv: arXiv:2301.04234. doi: 10.48550/arXiv.2301.04234.

[14] D. Olshansky and R. R. Colmeiro, “Relay Mining: Incentivizing Full Non-Validating Nodes Servicing All RPC Types,” Apr. 27, 2024, arXiv: arXiv:2305.10672. doi: 10.48550/arXiv.2305.10672.

[15] J. P. Allamaa, P. Patrinos, T. Ohtsuka, and T. D. Son, “Real-time MPC with Control Barrier Functions for Autonomous Driving using Safety Enhanced Collocation,” Jul. 11, 2024, arXiv: arXiv:2401.06648. doi: 10.48550/arXiv.2401.06648.

[16] W. Li and H. Wang, “Detection-Correction Structure via General Language Model for Grammatical Error Correction,” in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 1748–1763. doi: 10.18653/v1/2024.acl-long.96.

[17] “A novel hybrid framework for metabolic pathways prediction based on the graph attention network | BMC Bioinformatics | Full Text.” Accessed: Dec. 30, 2024. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04856-y?utm_source=chatgpt.com

[18] R. Sun, X. Wu, and Y. Wu, “An Error-Guided Correction Model for Chinese Spelling Error Correction,” Mar. 20, 2023, arXiv: arXiv:2301.06323. doi: 10.48550/arXiv.2301.06323.

[19] Y. Leng et al., “SoftCorrect: Error Correction with Soft Detection for Automatic Speech Recognition,” Dec. 20, 2023, arXiv: arXiv:2212.01039. doi: 10.48550/arXiv.2212.01039.

[20] S. Anbukkarasi and S. Varadhaganapathy, “Neural network-based error handler in natural language processing,” Neural Comput & Applic, vol. 34, no. 23, pp. 20629–20638, Dec. 2022, doi: 10.1007/s00521-022-07489-7.

[21] R. Kamoi et al., “Evaluating LLMs at Detecting Errors in LLM Responses,” Jul. 27, 2024, arXiv: arXiv:2404.03602. doi: 10.48550/arXiv.2404.03602.

[22] Y. Wang, Y. Wang, J. Liu, and Z. Liu, "A Comprehensive Survey of Grammar Error Correction," May 02, 2020, arXiv: arXiv:2005.06600. doi: 10.48550/arXiv.2005.06600.

[23] T. T. H. Nguyen, A. Jatowt, N.-V. Nguyen, M. Coustaty, and A. Doucet, "Neural Machine Translation with BERT for Post-OCR Error Detection and Correction," in JCDL '20: The ACM/IEEE Joint Conference on Digital Libraries in 2020, Virtual Event, China: ACM, Aug. 2020, pp. 333–336. doi: 10.1145/3383583.3398605.

[24] M. H. Bijoy, N. Hossain, S. Islam, and S. Shatabda, "A transformer-based spelling error correction framework for Bangla and resource scarce Indic languages," Computer Speech & Language, vol. 89, p. 101703, Jan. 2025, doi: 10.1016/j.csl.2024.101703.

[25] R. Raju, P. B. Pati, S. A. Gandheesh, G. S. Sannala, and S. KS, "Grammatical vs Spelling Error Correction: An Investigation into the Responsiveness of Transformer-based Language Models using BART and MarianMT," J. Info. Know. Mgmt., vol. 23, no. 03, p. 2450037, Jun. 2024, doi: 10.1142/S0219649224500370.

[26] "C4_200M." Accessed: Jan. 20, 2025. [Online]. Available: https://www.kaggle.com/datasets/felixstahlberg/the-c4-200m-dataset-for-gec

[27] Y. Song, K. Krishna, R. Bhatt, K. Gimpel, and M. Iyyer, "GEE! Grammar Error Explanation with Large Language Models," in Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 754–781. doi: 10.18653/v1/2024.findings-naacl.49.

[28] C.-J. Yu, A. Rovira, X. Pan, and D. Freeman, "A validation study to trigger nicotine craving in virtual reality," in 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), Mar. 2022, pp. 868–869. doi: 10.1109/VRW55335.2022.00285.

[29] H. A. Lokhande, L. J. Kinage, P. M. Kolunkar, J. M. Salunkhe, and S. Kale, "Enhancing Text Quality with Bi-LSTM: An Approach for Automated Spelling and Grammar Correction," in 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Apr. 2024, pp. 01–07. doi: 10.1109/ADICS58448.2024.10533521.

[30] A. Solyman et al., "Optimizing the impact of data augmentation for low-resource grammatical error correction," Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 6, p. 101572, Jun. 2023, doi: 10.1016/j.jksuci.2023.101572.

Hybrid Attention-Based Transformers-CNN Model for Seizure Prediction Through Electronic Health Records

Janjhyam Venkata Naga Ramesh[1], M. Misba[2], Dr. S. Balaji[3], Dr. K. Kiran Kumar[4],
Elangovan Muniyandy[5], Prof. Ts. Dr. Yousef A. Baker El-Ebiary[6], Dr B Kiran Bala[7], Radwan Abdulhadi .M. Elbasir[8]
Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India[1]
Adjunct Professor, Department of CSE, Graphic Era Hill University, Dehradun, 248002, India[1]
Adjunct Professor, Department of CSE, Graphic Era Deemed To Be University, Dehradun, 248002, Uttarakhand, India[1]
Assistant Professor, Senior Grade, Department of Artificial lntelligence and Machine Learning,
Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India[2]
Department of CSE, Panimalar Engineering College, Chennai, India[3]
Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India[4]
Department of Biosciences-Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences, Chennai, India[5]
Applied Science Research Center, Applied Science Private University, Amman, Jordan[5]
Faculty of Informatics and Computing, UniSZA University, Malaysia[6]
Head of the Department, Department of AI & DS, K. Ramakrishnan College of Engineering, Trichy, India[7]
Field of Study in Electronic Engineering, Libyan Advance Center of Technology, State of Libya[8]

***Abstract*—Seizures are a serious neurological disease, and proper prognosis by electroencephalography (EEG) dramatically enhances patient outcomes. Current seizure prediction methods fail to deal with big data and usually need intensive preprocessing. Recent breakthroughs in deep learning can automatically extract features and detect seizures. This work suggests a CNN-Transformer model for epileptic seizure prediction from EEG data with the goal of increasing precision and prediction rates by investigating spatial and temporal relationships within data. The innovation is in employing CNN for spatial feature extraction and a Transformer-based architecture for temporal dependencies over the long term. In contrast to conventional methods that depend on hand-crafted features, this method uses an optimization approach to enhance predictive performance for large-scale EEG datasets. The dataset, which was obtained from Kaggle, consists of EEG signals from 500 subjects with 4097 data points per subject in 23.6 seconds. CNN layers extract spatial characteristics, while the Transformer takes temporal sequences in through a Self-Attention Profiler to process EEG's temporality. The suggested CNN-Transformer model also performs well with 98.3% accuracy, 97.9% precision, 98.73% F1-score, 98.21% specificity, and 98.5% sensitivity. These outcomes show how the model identifies seizures while being low on false positives. The results indicate how the hybrid CNN-Transformer model is effective at utilizing spatiotemporal EEG features in seizure prediction. Its high sensitivity and accuracy indicate important clinical promise for early intervention, enhancing treatment for epilepsy patients. This method improves seizure prediction, allowing for better management and early therapeutic response in the clinic.**

***Keywords*—Epileptic seizure prediction; EEG signal analysis; CNN-Transformer model; deep learning in healthcare; spatiotemporal feature extraction; neural network optimization**

I. Introduction

The word, epilepsy is derived from the Greek word “epilambanein” which translates to grabbing or attacking. Epilepsy also known as Epileptic Seizure (ES) is a leading neurological disorder in which the neurons in the cerebral cortex experience seeds of unusual discharge. This can cause sudden seizures or fits [1]. A sudden change in brain activity causes a seizure that cannot be controlled. Some usual signs of ES are: uncontrollable jerking movements, feeling dizzy, tingling sensations, seeing flashes of light, losing awareness, and changes in how things taste, sound, smell, or feel [2]. It can affect anybody, although common in childhood and often, it develops in people with over 65 years. Treating epilepsy is complicated and depends on several things, like what type of epilepsy you have, how bad it is, how old the patient is, any other health issues they might have, and how well they respond to medications. Sadly, even with improvements in medicine, some patients still have trouble controlling their seizures [3]. One reason for this could be that antiepileptic drugs don't work well for some patients. There are many seizure medications, but it can be hard to find the right one or the right amount for each person. EEG and iEEG signals are used to diagnose epilepsy and help predict and detect seizures. EEG is a way to check the brain's electrical activity without any surgery. iEEG is a method that requires putting electrodes inside the skull, which is more invasive. In both situations, recording the brain's electrical signals helps us study changes in brain cells and shows us when seizures happen. iEEG allows for better recording of brain activity, making it easier to find the exact part of the brain where seizures happen [4].

A study has been done to look at how people with epilepsy use ASM medications, using data from Sweden's population register. The study showed that there are only a few medicine options for most new patients. They include Randomized effectiveness trials, Cohort studies, Case-control studies and Cross-sectional studies on how drugs are used, using ambulatory electronic health records and insurance claims from different countries are available in Observational Health Data Sciences and Informatics (OHDSI) community [5]. OHDSI has a centralized research infrastructure where data is categorized and standardized to coincide with the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM). This model helps to present healthcare information from different sources in a clear and uniform manner [6]. A shared way to organize and show data helps create standard tools for analyzing it. The CDM has been designed to acquire scientific data from various kinds of information and to join in major studies utilizing these data types. It found that there were different treatment paths depending on the data source used [7].

In the past, the features of genetic epilepsy were mainly identified by closely watching small groups of people. More recently EHRs have been used to assist with all of the data that is present today. With non-problematic and interchanged words, systematic annotation processes, and logical reasoning as ingredients, this method has been made possible and mitigating some of the problems posed by large real-world data [8]. Detailed analysis of traits has significantly improved our knowledge of the range of disorders linked to changes in genes like SCN2A and STXBP1, among others. People who have genetic epilepsy that starts in childhood show a variety of symptoms and often have high rates of mental health issues and physical health problems. Pettus et al. [9] do not fully understand how their condition changes from childhood to teenage years or how it affects their use of healthcare services and medical treatment.

Earlier research has found that predicting seizures in newborns is difficult, and medical tests and brain wave readings usually do not do a good job of predicting these seizures. Signs and symptoms alone can't tell us if a newborn will have seizures. EEG studies show that while a normal brain wave pattern can reliably indicate that there are no seizures, an abnormal pattern does not necessarily mean that seizures are happening [10]. Models that predict seizures using EEG data, which are created by examining EEG segments, checking EEG reports, or analyzing EEG recordings directly, have faced problems because they often involve small groups of participants or only look at short time periods. Additionally, only a few of these models have been tested on newborns. Because doctors need to look at charts or EEG readings by hand, the current models can't be easily used in regular patient care [11]. Scientists have created programs that use brain signals from EEG and iEEG to find and predict when epileptic seizures will happen. These algorithms use different ways to study signals, like looking at their frequency and how they change over time. These devices can constantly record EEG signals and warn the patient before a seizure happens. Using EEG and iEEG signals to find and predict seizures can really help people with epilepsy live better lives and lower the costs of their treatment and healthcare [12]. The major contributions of this study was given below:

- The study introduces a novel hybrid model that combines Convolutional Neural Networks (CNN) for spatial feature extraction in EEG data, enhancing the model's ability to predict epileptic seizures effectively.
- This approach automatically extracts meaningful features from raw EEG data, reducing the need for extensive preprocessing and enabling more scalable solutions.
- The model effectively handles large-scale EEG time-series and leverage the self-attention mechanism in Transformers, allowing the model to capture both local and global patterns in the data.
- The architecture is designed with flexibility, making it applicable to other forms of neurological or time-series medical data.
- This research demonstrates how the hybrid CNN-Transformer model surpasses conventional methods by providing better performance in seizure prediction, setting a new benchmark for EEG-based epileptic seizure detection.

This rest of the work focuses as follows: Section II reviews the related works for the prediction of seizure, Section III describes the problems in existing methods, Section IV demonstrates the method for the prediction of seizure using deep learning techniques, Section V evaluates the results and discussions, and Section VI concludes the research.

II. LITERATURE WORK

Seizures caused by epilepsy are a common neurological disorder that affects a huge amount of people worldwide. Up to 70% of those who receive prompt and accurate identification remain free from seizures. In order to do this, medical practitioners urgently need intelligent, automated solutions to help them accurately detect neurological problems. Previously, attempts have been made to identify behaviours in epilepsy patients using machine learning algorithms and raw electroencephalography (EEG) data. However, in order to extract features from these investigations, clinical knowledge in areas such as radiology and clinical procedures was necessary. Performance was constrained by the human feature engineering used in traditional machine learning for categorization. Automated feature learning from raw data without human intervention is where deep learning shines. To detect seizures, for instance, deep neural networks are currently showing promise in processing raw EEG data, doing away with the need for extensive clinical or engineering requirements. While preliminary research is still in its infancy, it already shows promising applications in various medical fields. However, in this work, the investigation of model explainability is not part of the ResNet-BiGRU-ECA strategy. The data' clinical relevance and interpretability may be hampered by this omission. Mekruksavanich and Jitpattanakul [13] present ResNet-BiGRU-ECA, a novel deep residual model that uses EEG data to accurately diagnose epileptic seizures by examining brain activity. The effectiveness of our suggested deep learning model was assessed using an epilepsy benchmark dataset that was made available to the public. The performance of the proposed

deep learning model was evaluated with the epilepsy benchmark dataset that was released to the public. The experiments conducted by us proved that our proposed model outperformed the basic model as well as state-of-the-art deep learning models by obtaining a high accuracy of 0. 998 and the highest of the F1-score of 0. 998. Nevertheless, the current approach ResNet-BiGRU-ECA does not consider the factor of model explainability. This omission might affect comprehensibility of the results obtained from the models.

Millions of humans worldwide suffer from epilepsy, and prompt seizure diagnosis is essential for both improved health and efficient treatment. The study of electroencephalograms (EEGs) provides a non-invasive option, but it takes a lot of time and effort to interpret the data visually. Many current efforts do not take account of the computational complexity of their models or processing speed, instead concentrating only on acquiring competitive levels of accuracy. The goal of this work was to create an automated approach for detecting epileptic seizures in EEG data by using analytic techniques. The main objectives of the efforts have been to reduce computing complexity and offer high accuracy effects by just using a small portion of the signal's frequency spectrum. In this paper, Urbina Fredes et al. [14] combined machine learning and signal processing methods to provide a novel automated method for seizure detection. The suggested approach consists of four steps: (1) Savitzky-Golay filter preprocessing: this eliminates background noise. (2) Decomposition: to recover spontaneous alpha and beta frequency bands, use discrete wavelet transform (DWT). (3) Feature extraction: The following six metrics are determined for each frequency band mean, Standard Deviation, Skewness, Kurtosis, Energy, and entropy. (4) Classification: Signals are classified as either normal or seizure-containing using the support vector machine (SVM) approach. Two publicly available EEG datasets were used for the evaluation of the approach reported in this paper. An accuracy of 92.82% was obtained in the alpha band and 90.55% in the beta band, which is reasonably adequate to resolve seizures. Moreover, the low computing cost that was found points to a possible useful use in circumstances involving real-time evaluation. The findings collected show that it might be a useful tool for both patient monitoring and epilepsy diagnosis. For clinical validation and possible real-time implementation, more research is required. A drawback of this study is that even as future research studies areas to improve signal processing methods may advance, the study does not examine the use of these methods beyond EEG signals. This may limit their applicability of the findings with other types of data.

A neurological condition called epilepsy results in recurrent seizures. Multiple factors can be extracted in order to identify and forecast a seizure, as electroencephalogram (EEG) patterns vary between pre-ictal, ictal, and inter-ictal stages. Nevertheless, little research has been done on the two-dimensional brain connection network. The goal of Tian et al. [15]was to examine how well it works for seizure prediction and detection. To extract image-like features, the two time-window length and five frequency bands and five connectivity measures were adopted. They were then used to train a CMT classifier for SIM and CSM, and a support vector machine for SSM.. Ultimately, studies of efficiency and selecting features were carried out. According to the CHB-MIT dataset's classification findings, a longer window denoted superior performance. SSM, SIM, and CSM had the best detection accuracies, which were 100.00, 99.98, and 99.27%, in that order. The three highest forecast accuracy values were 86.17%, 99.38, and 99.72, in that order. Furthermore, excellent performance and great efficiency were demonstrated by the Phase Lock Value and Pearson Correlation Coefficient connection in the β and γ bands. In order to identify and forecast seizures automatically, the suggested brain connection characteristics shown strong dependability and usefulness, which bodes well for the development of portable real-time monitoring devices. This paper has three restrictions. They are as follows: Despite the distance between bipolar montage and volume conduction, this issue is too severe to be resolved fully. Furthermore, because certain brain connection metrics, like PLV, are sensitive to volume conduction, they should be handled carefully when EEG recordings utilize referential (or unipolar) montage. The unsatisfactory performance of the cross-subject model in seizure prediction suggests that the heterogeneity was not adequately minimized. Other network topologies, including graph neural networks, will be investigated in the upcoming work to address this issue. The CHB-MIT sample size is modest, which may have an impact on deep learning network efficiency. As a result, our partner clinic will continue to gather medical information in the future for categorization.

A computer-assisted diagnostic system (CADS) for the automated identification of seizures caused by epilepsy in EEG data is presented in this study by, Malekzadeh et al. [16]. Three parts make up the suggested method: preprocessing, feature extraction, and classification. The simulations are carried out using the Bonn and Freiburg datasets. First, employed a band-pass filter with a cut-off frequency of 0.5–40 Hz to remove EEG dataset aberrations. The Tunable-Q Wavelet Transform (TQWT) is employed in the breakdown of EEG signals. Different linear and nonlinear characteristics are taken out of TQWT sub-bands in the second stage. Various statistical, frequency, and nonlinear properties are taken out of the sub-bands in this stage. The nonlinear characteristics that are employed are grounded in theories of unpredictability and fractal sizes (FDs). Various methods based on deep learning (DL) and traditional machine learning (ML) are described for the categorizing stage. This stage involves using a CNN-RNN-based deep learning technique with the suggested number of layers. Remarkable outcomes have been observed when the retrieved characteristics are fed into the suggested CNN-RNN model. To illustrate the efficacy of the suggested CNN-RNN classification process, the K-fold cross-validation with k = 10 is utilized in the classification stage. The accuracy of the suggested CNN-RNN approach for the Bonn and Freiburg datasets was 99.71% and 99.13%, respectively, according to the results. The study's shortcomings are spoken about As previously said, there are several forms of epileptic seizures, and prompt identification is crucial. To date, there is no dataset available on the different kinds of epileptic seizures. As a result, scholars are unable to do meaningful study in this area. Furthermore, there is restricted usage of the existing EEG datasets for epileptic seizure diagnosing; as a result, true and accurate epileptic seizure detection based on AI algorithms will not be achievable. The lack of a dataset of EEG signals that highlights the preictal, ictal,

and interictal periods is another drawback to using EEG signals to diagnose epileptic seizures. With regard to these shortcomings, it is feasible to employ more developed and rather recent DL models for the identification of the various types of epileptic seizures.

To achieve better seizure detection for patients, Wang et al (2023) propose a new structure for MBdMGC-CWTFFNet, a multi-branch dynamic multi-graph convolution based channel-weighted transformer feature fusion network.). In other words, both temporal, spatial and spectral information from the epileptic EEG are initially integrated through a multibranch (MB) feature extractor. Subsequently, to effectively learn dynamic and deep graph structures and extract prerequisite features from the multi-domain graph, build a point-wise dynamic multi-graph convolutional network (dMGCN). Thus, the final chosen method for fusing the multi-domain graph features is called the channel-weighted transformer feature fusion network (CWTFFNet) based on the integration of the local and global channel-weighted techniques with the multi-head self-attention technique. The experimental results concluded that the proposed method delivers superior prediction accuracy to the state of art methods, and hence, indicates the potentiality of the method as an efficient resource for patient-specific seizure prediction. The proposed MB-dMGC-CWTFFNet is evaluated on the CHB-MIT public EEG dataset and a private intracranial sEEG data set. Although the suggested predictions framework performs satisfactorily in terms of seizure detection, the investigation still has shortcoming, MB-dMGC-CWTFFNet is capable of providing an end-to-end seizure warning without the need for laborious EEG pre-processing. However, in real-world warning scenarios, artifacts from epileptiform discharges and perhaps problematic channels might interfere with the predictor and result in some false positives. Shi and Liu [17] propose a novel bi-level coding detection approach named B2-ViT Net in the current research to obtain new generalized spatio-temporal long-range correlation features. These features can be used to model the inter-channel relations in the space domain and express the temporal long-range dependencies which are important for seizure detection. In addition, due to the ability of deep and wide feature search, the proposed model can learn the generalized seizure prediction features in the large space. Thus, only two public datasets, namely the Kaggle dataset and the CHB-MIT dataset, offer enough data to perform enough experiments. Our suggested model offers some interpretability and has demonstrated encouraging results in automatic seizure prediction tests in comparison with different approaches already in use. There remain certain shortcomings in present works. On the one hand, the results were not verified by neuronal tests since there was inadequate data available regarding the person's epileptogenic zone and associated biomarkers. Above two papers share the same limitations. However, the approach is based on individual patients which means that both the test and training sets originate from the same patient. A model generated by one individual cannot be easily transferred to a different person for client-independent seizures detection activities. This is mostly due to the fact that our approach is unable to deal with the disparity in distributions among the test and training sets.

III. Problem Statement

The area of focus in this study is the difficulty of the classification of epileptic seizures from EEG which is a important process for the early detection of epilepsy patients. Epilepsy is a neurological disorder whose symptoms are caused by sudden recurrent episodes of abnormal brain electrical activity in the form of epileptic seizures detectable in EEG data. Most machine learning models employed in this field are obstructed by the volume and comprehensiveness of the EEG data. Additionally, most previous works are categorized under the binary classification and do not consider the temporal distribution of features, thereby having lower prediction rates. Large-scale data sources include databases and repositories with numbers of EEGs and with complex structures and features, making necessary a more sophisticated method that can enhance meaningful features selection and raise the seizure detection rate. This study aims at resolve the mentioned issues using a hybrid CNN-Transformer model designed based on the utilization of CNNs for spatial feature extraction while Transformers for modeling long-term temporal dependencies to make the seizure prediction more accurate and efficient.

IV. Proposed Hybrid Attention-based Transformers-CNN Model

This work employs a systematic method to design and assess a dual CNN-Transformer model for epileptic seizure forecast based on the EEG data. Recruiting subjects entails one of the key stages of the research method in the course of data collection. The EEG dataset I have download from Kaggle has data from 500 people and 4097 values per person, which reflects 23.6 sec of EEG records. Following data loading, a number of data pre-processing steps of data cleaning, handling of missing values and normalization are performed. The data set is then split into 1-second segments, which yields 11500 samples – each of the segments has 178 samples in it. Every sample is then categorized into one of five categories depending on the level of brain activity. The proposed model architecture known as Hybrid Model is comprised of ConvNet and Self-Attention under the category of Transformer mechanisms. The CNN layers are intended for detecting spatial patterns in the EEG while isolated singular areas of the signal are scrutinized and deep local features are extracted, more information about temporal relationships is processed in the Transformer block which employs self-attention techniques. This integration makes it possible for the model o scan through the time series data in a way that checks for any specific structure of the data at a given time or space. This partial EEG data is used for training the aforementioned model, and all the testing is done through cross validation. Metrics like accuracy, precision, F1 measure, specificity and sensitivity are calculated in order to assess the performance of the model. Importantly, hyperparameter tuning is also performed in order to achieve maximum model performance. The last output is then compared to the baseline traditional machine learning algorithms to show the efficiency of the developed hybrid CNN-Transformer model for early seizures' epileptic prediction. The workflow of proposed model is depicted in Fig. 1.

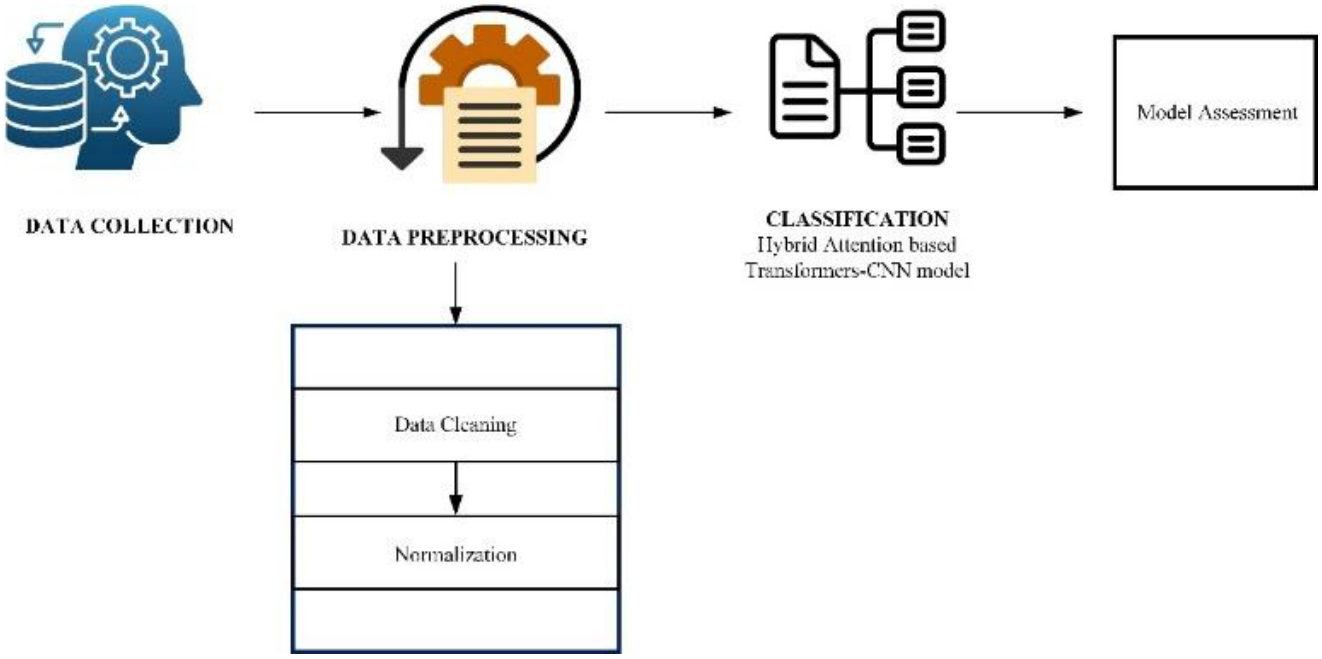


Fig. 1. Workflow of proposed model.

A. Dataset Collection

The data for this study is obtained from an open source Kaggle dataset that has EEG recordings of brain activity from 500 people. The EEG activity of each participant is recorded within 23.6 seconds with time series split into 4097 data points. For the purpose of easy handling and better understanding the dataset was divided into 23 parts and each part has 1 second recording with 178 data points for each second. This lead to a dataset of 11,500 rows, however each row represents 1 second of EEG recordings of an individual. The last column of the dataset, y, represents the response variable, indicating one of five categories of brain activity: She underwent five studies for video-EEG monitoring after being treated for her epilepsy: (1) during an epileptic seizure, (2) EEG recorded from the electrode over the tumor area, (3) EEG from the non-lesion area although she has a brain tumor, (4) eyes close and (5) eyes open. Despite the fact, the provided dataset includes five classes, the majority of researchers are interested in binary classification, in which the first class corresponds to epilepsy seizure, while the other classes consist of different non-seizure patterns. After this pre-processing of data, the dataset is saved in CSV format, where there are 178 numerical features besides a single quantitative aim variable and improving the legibility of the data to prepare for machine learning tasks for the identification of epileptic seizures [18].

B. Data Preprocessing

EEG data preprocessing is an extremely important part of the overall process because the machine learning model has to operate precisely and as fast as possible. Due to the high dimensionality and volume of data originating from EEG records, such data must be pre-processed and cleaned incorporating a structured methodology for normalization and transformation processes. The following is the data cleaning mechanism, normalization, and missing data to help to prepare the dataset with suitable features for the training model and the test model.

1) Data cleaning: The first process involved entails reading in the EEG dataset from the CSV file in order to have an understanding of the factors involved or the variables used in cleaning and if there are any issues with some of them. The first check of the data allows to say if there are any outliers in the data set, meaning that there are either corrupted or new, not expected values, as can be seen with high or low values of the EEG signal that can indicate the malfunctioning of the sensors. To identify such problems, descriptive measures like mean, median and size of standard deviation are applied for cases that show irregularities in the distribution. Dealing with outliers is often mission important because they can skew the result. Anomalies or extraordinary values markedly separated from the principal data sequence are mostly identified using the Min-Max methods, which are statistical. Once an organization identifies these cases, there are different ways to deal with them: omitting the cases, limiting the values to some threshold (for instance, 99% of the maximum value), or using some averaging techniques including a moving average to reduce the impact of the outliers. Also, the content of the data must remain consistent, which is one major goal of consistency. All EEG signals must be normalized and preprocessed, and each chunk must contain one second of data with 178 data points in each data block. The dataset may contain multiple copies of the same 1-second EEG segment; therefore, all such rows have to be deleted to avoid biases in the subsequent analysis and modeling steps.

2) Normalization: The amplitude and signal intensities in the EEG data are significantly variable across subjects making it necessary to normalize it. These differences may pose

challenges to the correct interpretation to machine learning models as some of the features may dwarf the others in size. The data is first normalized, so each feature (X1 to X178) has equal weights in order to avoid biasing of the model by any particular feature. In these models normalization enables them to learn faster, have higher accuracy and generality since the input data is well balanced for all features. Another popular method is called Min-Max scaling, it just scales the values of the input features to the range between 0 and 1. It helps the models to be processed by the methods of data processing in such a way that all features are presented at the same scale. The formula for Min-Max scaling is shown in Eq. (1) as follows,

$$X' = \frac{X - Xmin}{Xmax - Xmin} \quad (1)$$

In which X is the first value, X′ is the normalized value, Xmin is the minimum value and $Xmax$ is the maximum value.

C. Feature Extraction using-CNN Model

Detection of seizure among humans is done using the Hybrid Attention based Transformers-CNNs model as proposed. Fig. 2 depicts the architecture diagram of proposed model. Convolutional Neural Network is used for image feature extraction for the dataset. There is still a deep learning method that comes under CNN method. The CNN is the most famous and used algorithm in the field of DL The CNN in comparison to other algorithms has an advantage of not requiring human supervision to fix on which features matter The CNN is constructed from a number of layers which include the input layer, convolutional layers, pooling layer, fully connected layers as well as the output layer. These layers helps to extract the features and classify the seizure as seen in the following.

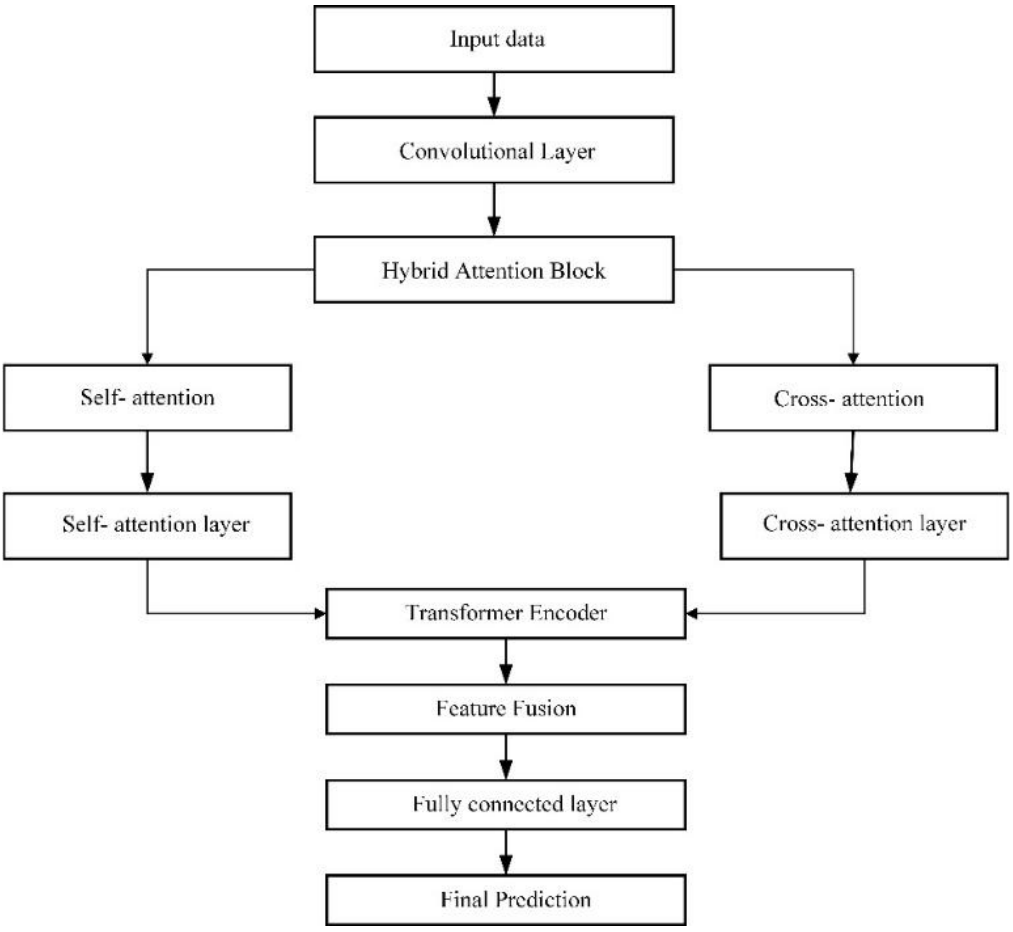


Fig. 2. The architecture diagram of proposed model.

1) Input layer: The topmost layer in CNN architecture is the input layer. In this layer the data is represented as multi-dimensional. The input it receives is EHR; it comprises EEG and patient It accepts the processed data and passes it on to the subsequent layer of the model.

2) Convolutional layer: Convolution layers are the fundamental components as constituent block of the network and exist in hierarchical form. Basically, Convolutional data is used to extract features on the input data to be processed for its main function. The other thing which needs to be added after having the feature maps in CNN is the pooling or the sub-sampling layer following a convolution layer.

3) Activation layer: The activation function has a very significant role in CNN layers. The output of the filter is passed to another mathematical operation known as the activation function. ReLu, which is an abbreviation for rectified linear

unit, is the most commonly used activation function in CNN feature extraction most of the time. It converts all negative entries into zero and all positive entries into values which cannot be changed.

D. Hybrid Attention-Based Transformers-CNN Framework

Hybrid Attention-based Transformers-CNNs integrate the performance of the Convolutional Neural Networks (CNNs) and Transformers models to deliver satisfactory performance.

1) Attention mechanism: The attention mechanism is a type of resource allocation mechanism and is developed to imitate the human brain's attention. When the human brain processes things, it directs attention towards areas that is required, while minimizing or in some cases excluding other areas, to obtain specific details which may requires attention. It has often been seen that long-term dependencies pose a problem for attention mechanisms; however, for very long sequences, it is quite helpful. With the help of attention mechanisms, it is possible to enhance the information that is learned by CNN and other neural network models.

2) Transformer: In the current architecture of transformers, an encoder-decoder framework is used, in which the encoder part is responsible for the encoding of the input sequence into a representation while the decoder part is used to decode the output sequence using the information of the input sequence representation. In fact, each encoder and decoder layer of a transformer includes numerous self-attenuating heads as well as feed-forward neural networks. The major element at the core of transformers is the 'Attention' mechanism that enables the model to pay attention to the various parts of the input sequence during predictions. This attention mechanism assists the transformers in capturing the information of the previous and the subsequent words within a given sentence thus they make it easy for the transformers to represent the input data as desirable.

3) Fully connected layer: Finally, there is the necessity for feature classification for the purpose of predicting actual types of the input data. The last fully connected layer must ideally contain neurons it is equal to the number of output classes which we are going to predict. The fully connected layer will take the attention-weighted or feature-extracted signal as input and output classification decision to permit superior tuning of weights and learning anterior to an yield

4) Output layer: The output from the fully connected layers is then put through another activation function for classification like sigmoid or softmax this quantizes the output of each class to probability of score of class. It predicts the output as seizure or no seizure.

V. Results and Discussion

In this section, present the result and analyse the performance of Hybrid Attention-based Transformers-CNN model for seizure prediction using EHR. The proposed model was trained and tested using Python along with deep learning libraries such as CNNs for local spatial feature extraction and Transformer-based attention mechanism for capturing sequential dependencies in EHR. This model's main objective is to improve the reliability of seizure prediction by examining prevailing patterns within EHRs including the EEG signals. The integrated hybrid model builds upon the CNN's outstanding features of processing spatial content and the value of the Transformer in preserving temporal relations crucial for making accurate forecasts based on content. The following results that are the capsules of basic performance aspects like accuracy, precision, F1 score, specificity, and sensitivity metrics were assessed comparatively in terms of training and testing data. These findings consider the Hybrid Attention-based Transformers-CNN model as an improved method for seizure prediction, thus advancing the study of computational healthcare.

A. Seizure Prediction Analysis

The Fig. 3 illustrates the comparison of the recorded EEG signals during the seizure event and normal EEG status. The column on the left side with labels in red presents seizure EEG signals as the right column with labels in blue presents normal EEG signal. In the seizure EEG signals, the following characteristics can be observed; spikes, papery, high amplitude typical of high activity of neurons. These signals contain sharp waves and high-frequency oscillations of voltage which characteristic of broadly epilepsy. On the other hand normal EEG produces more physiological rhythmic, lower amplitude waveforms as compared to the abnormal ones. These relatively stable oscillations represent the normal, non-pathological electrical organization of the brain. This way, the contrast achieved between the seizure and normal EEG patterns shows that while in seizure, neuronal activity is much more violent and irregular compared to normal EEG with its regular and rhythmic activity.

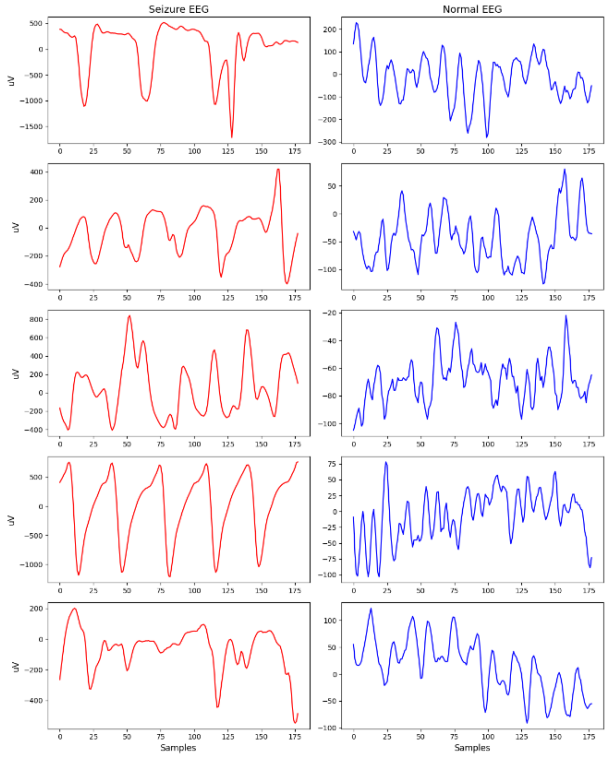


Fig. 3. Seizure analysis.

B. Training and Testing Accuracy

In Fig. 4, the Testing and training accuracy graph shows the accuracies of a machine learning model, Hybrid Attention-based CNN-Transformer model across the number of epochs at 120. Explore the journey follows an upward trend in the training accuracy based on the epochs based on the blue line as shown below. This shows that the model is actually training well from the training data, adapting each time to discern one pattern or the other well enough to make a good prediction wrongly. Therefore, a CNN can capture the local version of the data, while the Transformer can capture the global version with almost maximal internal attention, and thus see its training accuracy improve dramatically. The orange line also increases representing the testing accuracy once the weight accumulated is tested to the unseen data set it is much lower than the training accuracy at the initial stages. The first thing that stands out is a big divide between the training accuracy and the testing accuracy a characteristic that shows that the model, rather than testing it, was better at preparing the data. Worth to be pointed out is that two lines indicate the discrepancy at the quite early steps of training and as training progresses, these lines get closer to each other which is positive. This means that the proposed Hybrid Attention-based CNN-Transformer model is experiencing less overfitting as it can generalise better. From the training as well as testing accuracy plots, there is a gradual improvement in the training and testing database resulting to development of an effective model. While the training of the model progresses, the hybrid architecture provides optimal CNN for spatial features learning and optimal transformer for long-range dependencies features learning, which in turn improves accuracy of both training and testing sets. Slowly, the generality is enhanced as the difference in the accuracies is narrowed even if there could still be some further refining or optimization of the final ratings to make them as close to each other as possible without memorizing the data. This overall trend well characterizes the capability of the Hybrid Attention-based CNN-Transformer model to handle such data and enhance generalization.

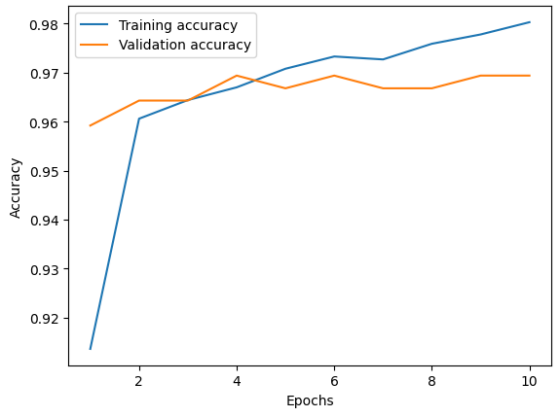


Fig. 4. Testing and training accuracy.

C. Training and Testing Loss

This is seen in Fig. 5 where Testing and Training Loss describes the result of an ML model which incorporates the Hybrid Attention-based CNN-Transformer model for the Training-Testing split with 120 epochs. The blue color is for the training loss while the violet for validation loss at the beginning of epochs it is very huge and then drops hugely this indicate that the model is learning the training data and the samples in general. This steep slope as a characteristic of the model that uses both CNNs and the transformer's attention mechanisms to thoroughly look for the right details hence always homing in a slightly smaller number of mistakes. The orange line, showing the testing loss, is again on a descending path although not as sharp as that of the training loss. This slower reduction in testing loss means that when the Hybrid Attention-based CNN-Transformer model is learning on unseen data it is not able of learning at the optimal level or as efficiently as the CNN in generalization to inputs that were not used during training. The gap between the training and testing loss reveals how much the model depends on the training data set and probably overfitting on that data should they use more complex models such as CNN-Transformer models. These architectures, however, are fairly powerful, but need one or two more steps; for instance, applying regularization methods or adjusting the parameters of the attention layers to improve overfits across datasets. The fact that both loss lines are reducing in the long run also indicate that the model is learning; however, a slightly high testing loss mean that there is still work to be in done in an attempt to make the Hybrid Attention-based CNN-Transformer better especially in easily recognizing unseen data.

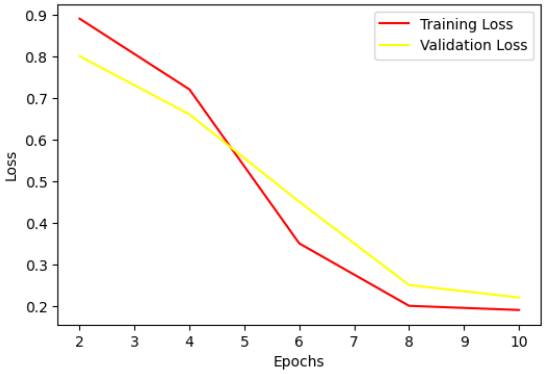


Fig. 5. Testing and training loss.

D. Performance Metrics

To evaluate the efficiency of the suggested approach these following metrics has been used Accuracy, Sensitivity, Specificity, Precision, and F1-measure. The following formula given below in Eq. (2) to Eq. (6).

$$\text{Accuracy} = \frac{True\ Positive\ + True\ Negative}{TP+TN+FP+FN} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (4)$$

$$\text{Specificity} = \frac{True\ Negative}{TN+FP} \quad (5)$$

$$\text{F1score} = \frac{2\times TP}{(2\times TP)+FP+FN} \quad (6)$$

Table I shows the effectiveness criteria of the system that has been proposed in this paper. There is 98.3% confidence that the model is correct most of the time, therefore this model comes

with a good test accuracy to both positive and negative cases. A given accuracy of 97.9% pays for itself in the ability to recognize true positives. This means that the recognition model's ability to balance between precision and recall is well interpreted based on F1 score of 98.73%. An accuracy of 95.90% demonstrates the model is accurate at determining true positives, positive predictive value of 98.21 defines the same for true negatives in an exceptional manner, thus showing the overall gratifying performance of the model.

TABLE I. PERFORMANCE METRICS OF PROPOSED MODEL

Metrics	Attention CNN-Transformer
Accuracy	98.3
Precision	97.9
F1 score	98.73
Sensitivity	95.90
Specificity	98.21

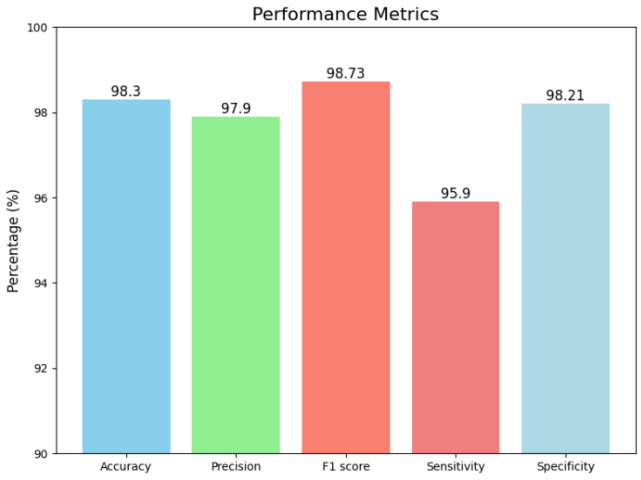


Fig. 6. Performance metrics of proposed model.

It is reflected from the Bar Chart shown in the Fig. 6 based on Performance Metrics of Proposed Model, which has demonstrated the efficiency of the proposed model on Accuracy, Precision, F1 Score, Sensitivity and Specificity. For Accuracy and Specificity, both our model scores at level 1, which tells us as to how well the model has been designed to predict both Positive and Negative cases. Specificity is also high suggesting that there is few misclassifications of true positives. However, F1 Score which is the lowest metric implies that there is a problem with the tradeoff between precision and recall. Again, sensitivity is slightly low compared to precision and this means that there is some room for improvement on true positive. The balance of the model we see is good and some slight modifications are required.

The classification performances of different models based on classification accuracy, precision, F1 score, specificity as well as sensitivity of the classified EEG signals are summarized in the following Table II. An 86.23% accuracy, 82.90% precision, and an F1-score of 84.50 for Sustainable Value Management. The Gaussian model is much more accurate with a total accuracy of 95.49%, total precision of 96.60 % and an F1-Score of 97.50% which makes it among the best models. Random Forest also gives better outputs with 92.06% accuracy rates, and an F1 score of 91.80%, but loses precision and sensitivity when compared with the Gaussian model. Analogue to this, the k-NN shows relatively good performance by achieving 86.89% accuracy, and 85.30% F1 score. Through optimization algorithm $aADGA$ model which is developed for this research yielded an accuracy of 97.49% and F1-score of about 98.2%. However, the proposed method is identified as more accurate than all the other methods having the accuracy of 98.3%, the precision of 0.979, F1-score of 0.9873, specificity of 0.9821 and the sensitivity 0.985. Even though the performances of classifiers were quite similar, it was revealed that the proposed method outperforms all the other methods used in the present study with the best accuracy, sensibility, and efficiency in EEG signal classification as shown in Fig. 7.

TABLE II. COMPARISON OF THE PROPOSED MODEL WITH THE EXISTING APPROACHES

Model	Accuracy (%)	Precision (%)	F1-score (%)	Specificity (%)	Sensitivity (%)
SVM	86.23	82.90	84.50	82.90	87.20
Gaussian	95.49	96.60	97.50	96.60	97.50
Random Forest	92.06	89.70	91.80	89.70	92.20
k-NN	86.89	83.70	85.30	83.70	86.95
$aADGA$	97.49	96.90	98.2	96.90	95.90
Proposed method	98.3	97.9	98.73	98.21	98.5

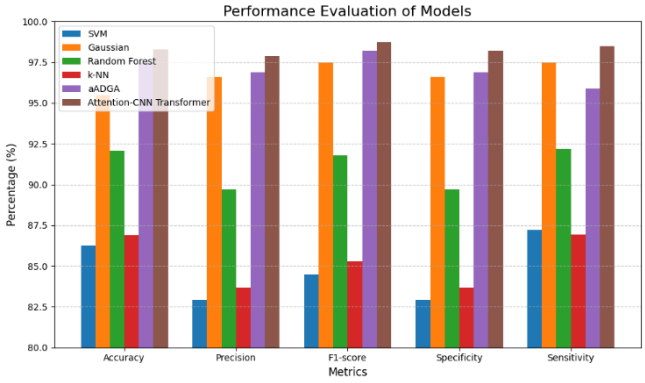


Fig. 7. Performance evaluation of proposed model with existing approaches.

E. Discussion

The suggested CNN-Transformer model exhibits exceptional effectiveness in predicting epileptic seizures from EEG data with high accuracy, precision, and sensitivity. The combination of CNNs for spatial feature extraction and Transformers for long-term temporal dependency capture enables a stronger classification of EEG signals between seizure and normal brain activity. The findings show that the model effectively detects seizure events with 98.3% accuracy, 97.9% precision, an F1-score of 98.73%, specificity of 98.21%, and sensitivity of 98.5% and outperforms classical models like k-NN, Gaussian models, and Random Forest classifiers. Comparison with current approaches indicates that conventional machine learning approaches, i.e., k-NN and Random Forest, are

less accurate because they make use of handcrafted features, which may fail to clearly identify the complex spatiotemporal relationships within EEG signals. The Gaussian model was relatively good with an accuracy of 95.49%, but it does not have the flexibility of deep learning-based methods in dealing with big EEG datasets. The hybrid CNN-Transformer solution addresses these limitations by utilizing deep feature extraction and self-attention, enhancing both classification accuracy and false positive minimization. The research underscores the importance of spatial and temporal modeling in the detection of seizures. CNN layers effectively extract local spatial patterns from EEG signals, while the Transformer architecture fine-tunes long-range dependencies to generalize well across varying seizure patterns. The dual-architecture approach reduces misclassification by learning strong signal representations, and this results in a significant reduction in false positives, which is a key issue in seizure prediction.

Another significant contribution is that the model can improve real-time seizure detection and thus is of excellent clinical significance. Reducing the dependency on the preprocessing of huge amounts of data and manual feature design, the approach proposed here offers a scalable, automatic solution to seizure prediction. Its ability to process large datasets makes it promising to be integrated into wearable EEG monitoring devices to facilitate early medical intervention in epilepsy patients [13].

VI. Conclusion and Future Work

This study effectively applies the idea of CNN-Transformer Integration for the epileptic seizure detection from the EEG data. Combining Convolutional Neural Networks for the spatial characteristics and Transformer-based attention for temporal features, the proposed framework improves the seizure prediction's robustness and accuracy. From these findings, it is evident that the hybrid approach enhances the capabilities of prediction with minimized false positive issues, making it a significant step towards solving some of the major challenges that exist in seizure detection. The model achieved impressive performance metrics, including an accuracy of 98.3%, precision of 97.9%, F1-score of 98.73%, specificity of 98.21%, and sensitivity of 98.5%. From these findings, it is evident that the hybrid approach enhances the capabilities of prediction with minimized false positive issues, making it a significant step towards solving some of the major challenges that exist in seizure detection. As depicted above, deep learning techniques can be applied in the enhancement of the area of epilepsy treatment and management not excluded the importance of training models with large-scaled EEG datasets. It opens doors to subsequent research that, perhaps, can enhance and polish such approaches to seizure prediction in order to lead to positive outcomes within clinical practice through timely interventions.

References

[1] A. Liede et al., "Risk of seizures in a population of women with BRCA-positive metastatic breast cancer from an electronic health record database in the United States," BMC Cancer, vol. 23, no. 1, p. 78, Jan. 2023, doi: 10.1186/s12885-023-10554-6.

[2] M. Gandy et al., "Managing depression and anxiety in people with epilepsy: A survey of epilepsy health professionals by the ILAE Psychology Task Force," Epilepsia Open, vol. 6, no. 1, pp. 127–139, 2021.

[3] M. Macnee et al., "Data-driven historical characterization of epilepsy-associated genes," European Journal of Paediatric Neurology, vol. 42, pp. 82–87, 2023.

[4] J. K. Knowles et al., "Precision medicine for genetic epilepsy on the horizon: Recent advances, present challenges, and suggestions for continued progress," Epilepsia, vol. 63, no. 10, pp. 2461–2475, 2022.

[5] H. Kim et al., "Characterization of Anti-seizure Medication Treatment Pathways in Pediatric Epilepsy Using the Electronic Health Record-Based Common Data Model," Front. Neurol., vol. 11, p. 409, May 2020, doi: 10.3389/fneur.2020.00409.

[6] K. Bolin, F. Berggren, P. Berling, S. Morberg, H. Gauffin, and A.-M. Landtblom, "Patterns of antiepileptic drug prescription in Sweden: A register-based approach," Acta Neurol Scand, vol. 136, no. 5, pp. 521–527, Nov. 2017, doi: 10.1111/ane.12776.

[7] K. Kubota et al., "Penetration of new antidiabetic medications in East Asian countries and the United States: A cross-national comparative study," PLoS ONE, vol. 13, no. 12, p. e0208796, Dec. 2018, doi: 10.1371/journal.pone.0208796.

[8] S. Yoo et al., "Developing a mobile epilepsy management application integrated with an electronic health record for effective seizure management," International Journal of Medical Informatics, vol. 134, p. 104051, Feb. 2020, doi: 10.1016/j.ijmedinf.2019.104051.

[9] J. H. Pettus et al., "Differences between patients with type 1 diabetes with optimal and suboptimal glycaemic control: a real-world study of more than 30 000 patients in a US electronic health record database," Diabetes, Obesity and Metabolism, vol. 22, no. 4, pp. 622–630, 2020.

[10] G. K. Mbizvo, K. H. Bennett, C. Schnier, C. R. Simpson, S. E. Duncan, and R. F. Chin, "The accuracy of using administrative healthcare data to identify epilepsy cases: a systematic review of validation studies," Epilepsia, vol. 61, no. 7, pp. 1319–1335, 2020.

[11] S. Jahan et al., "AI-based epileptic seizure detection and prediction in internet of healthcare things: a systematic review," IEEE Access, vol. 11, pp. 30690–30725, 2023.

[12] K. Puteikis and R. Mameniškienė, "Mortality among people with epilepsy: a retrospective nationwide analysis from 2016 to 2019," International journal of environmental research and public health, vol. 18, no. 19, p. 10512, 2021.

[13] S. Mekruksavanich and A. Jitpattanakul, "Effective Detection of Epileptic Seizures through EEG Signals Using Deep Learning Approaches," Machine Learning and Knowledge Extraction, vol. 5, no. 4, Art. no. 4, Dec. 2023, doi: 10.3390/make5040094.

[14] S. Urbina Fredes, A. Dehghan Firoozabadi, P. Adasme, D. Zabala-Blanco, P. Palacios Játiva, and C. Azurdia-Meza, "Enhanced Epileptic Seizure Detection through Wavelet-Based Analysis of EEG Signal Processing," Applied Sciences, vol. 14, no. 13, Art. no. 13, Jan. 2024, doi: 10.3390/app14135783.

[15] Z. Tian, B. Hu, Y. Si, and Q. Wang, "Automatic Seizure Detection and Prediction Based on Brain Connectivity Features and a CNNs Meet Transformers Classifier," Brain Sciences, vol. 13, no. 5, Art. no. 5, May 2023, doi: 10.3390/brainsci13050820.

[16] A. Malekzadeh, A. Zare, M. Yaghoobi, H.-R. Kobravi, and R. Alizadehsani, "Epileptic Seizures Detection in EEG Signals Using Fusion Handcrafted and Deep Learning Features," Sensors, vol. 21, no. 22, p. 7710, Nov. 2021, doi: 10.3390/s21227710.

[17] S. Shi and W. Liu, "B2-ViT Net: Broad Vision Transformer Network With Broad Attention for Seizure Prediction," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 32, pp. 178–188, 2024, doi: 10.1109/TNSRE.2023.3346955.

[18] "Epileptic Seizure Recognition." Accessed: Sep. 20, 2024. [Online]. Available: https://www.kaggle.com/datasets/harunshimanto/epileptic-seizure-recognition

AI-Driven Transformer Frameworks for Real-Time Anomaly Detection in Network Systems

Santosh Reddy P[1], Tarunika Chaudhari[2], Dr. Sanjiv Rao Godla[3], Janjhyam Venkata Naga Ramesh[4], Elangovan Muniyandy[5], A.Smitha Kranthi[6], Prof. Ts. Dr. Yousef A.Baker El-Ebiary[7]

Associate Professor, Department of Computer Science and Engineering, BNM Institute of Technology, Bangalore, India[1]
Assistant Professor, Computer Engineering Department, Government Engineering College, Dahod, India[2]
Professor, Dept. of Computer Science and Engineering, Aditya University, Surampalem, Andhra Pradesh, India[3]
Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India[4]
Adjunct Professor, Department of CSE, Graphic Era Hill University, Dehradun, 248002, India.[4]
Adjunct Professor, Department of CSE, Graphic Era Deemed To Be University, Dehradun, 248002, Uttarakhand, India.[4]
Department of Biosciences-Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India[5]
Applied Science Research Center, Applied Science Private University, Amman, Jordan[5]
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, AP, India[6]
Faculty of Informatics and Computing, UniSZA University, Malaysia[7]

***Abstract*—The detection of evolving cyber threats proves challenging for traditional anomaly detection because signature-based models do not identify new or zero-day attacks. This research develops an AI Transformer-based system with Bidirectional Encoder Representations from Transformers (BERT) technology with Zero-Shot Learning (ZSL) for real-time network system anomaly detection while solving these security challenges. The goal positions the development of an effective alerting system that detects Incident response and proactive defenses cyber threats both known and unknown while needing minimal human input. The methodology uses BERT to transform textual attack descriptions found in CVEs alongside MITRE ATT&CK TTPs into multidimensional embedding features. Visual embeddings generated from textual documents undergo comparison analysis with current network traffic data containing packet flow statistics and connection logs through the cosine similarity method to reveal potential suspicious patterns. The Zero-Shot Learning extension improves the system by enabling threat recognition of new incidents when training data remains unlabeled through its analysis of semantic links between familiar and unfamiliar attack types. Here utilizes three different tools that include Python for programming purposes alongside BERT for embedding analytics and cosine similarity for measuring embedded content similarities. Numerical experiment outcomes validate the proposed framework by achieving a 99.7% accuracy measure with 99.4% precision, 98.8% recall while maintaining a sparse 1.1% false positive rate. The system operates with a detection latency of just 45ms, making it suitable for dynamic cybersecurity environments. The results indicate that the AI-driven Transformer framework outperforms conventional methods, providing a robust, real-time solution for anomaly detection that can adapt to evolving cyber threats without extensive manual intervention.**

Keywords*—Anomaly detection; network security; transformer framework; bidirectional encoder representations from transformers; zero-shot learning

I. INTRODUCTION

The ever-growing usage of networked systems in all sectors brings along a surge in both the volume and complexity of network traffic. With this, the demand for always maintaining high levels of security and health concerning network systems has never been so important, be it for businesses, governments, or individuals [1]. The detection of network anomalies [2] becomes an active technology for ensuring integrity, availability, and confidentiality of such systems, and thus, it forms a core part in the security and health monitoring system [3]. So far, statistical techniques, rule-based systems, and machine learning models, including but not limited to SVM, k-Means, and lately, deep learning approaches, have become typical methodologies for the detection of network anomalies [4]. These usually model the recognition of designs of data that monitor systems detect datasets which deviate from established operational norms, typically created by comparing real-world activity to prerecorded models of "normal" operations [5]. One of the biggest challenges associated with traditional methods for detecting anomalies is that they cannot adapt to evolving and unknown attack patterns [6]. As cyber-attacks and intrusions into systems [7] become more sophisticated, models operating based on either predefined rules or knowledge about historical data usually cannot detect novel attacks, or even mark normal activities as anomalous [8]. What's more, these models require a great volume of labeled datasets in order to perform well, which might be unworkable on a large scale or in real time. Poor accuracy of anomaly detection may lead to overwhelming a security team with false alarms or, worse, its opposite-false negatives-meaning dangerous intrusions might go unnoticed. Moreover, most of the current models depend on heavy supervised learning, which requires a huge quantity labeled data intended for training [9]. Unfortunately, in network environments, obtaining labeled data is often both costly and time-consuming, hence creating obstacles to effective implementation. Moved by such difficulties, in current ages,

there has an increasing concentration in using deep learning methods, especially with the introduction of Transformer-based architectures capable of processing a high volume of data with high efficiency [10].

Recent deep learning developments have given rise to a class of models with enhanced capabilities to handle complex network traffic analysis challenges, especially long-range dependencies and the processing of sequential data [11]. The Transformer model, among other models, has enjoyed considerable attention because of the inherent nature of the model to capture long-range dependencies, which are very important in network anomaly detection [12]. Transformers are broadly applied in sequential data where event relationships can be dispersed across time, making them very appropriate for tasks such as time-series analysis and event detection. Even these models have their shortcomings, particularly in the detection of novel or unseen anomalies within real-time environments. Despite their ability to learn patterns from large volumes of data, it continues to necessitate great quantities of categorized training data and computational properties. Moreover, while these models do a great job in capturing temporal patterns, they are often unable to generalize to situations in which data is sparse or continuously evolving, such as zero-shot or few-shot learning environments. Zero-shot learning, on the other hand, allows models to generalize into new, unseen tasks without labeled data-a very promising avenue toward overcoming these limitations. By leveraging pre-trained models, zero-shot learning approaches enable anomaly detection systems to identify outliers without the use of large volumes of labeled data, thereby greatly reducing the time and cost involved in model training. The concept of zero-shot learning in network security, therefore, would enable anomaly detection of previously unknown attack vectors or anomalous behavior without requiring retraining of the model against every new scenario. Some recent interventions create models that can detect anomalies with better efficiency and accuracy, especially with a combination of deep learning techniques and zero-shot learning. While these are promising, still much area for development in generalization, real-time performance, and scalability.

This proposed study, therefore, tries to fill this gap by incorporating zero-shot anomaly detection with a BERT-enhanced Transformer model. The integration of BERT-a deep learning model designed originally for natural language processing-with Transformer networks-is an innovative combination of deep, pre-trained, contextualized embeddings with an advanced model of sequence processing. BERT already achieves state-of-the-art semantic understanding in most text-oriented applications, and its application to network traffic provides new insights into anomaly detection in network systems. Accordingly, the proposed framework will utilize BERT's capability of comprehension and representation of contextual information, together with the Transformer's strength in sequential data processing, for improved anomaly detection in real network scenarios. The key innovation in this paper is the integration of these two deep models, which will enable the system to detect anomalies in network flows while minimizing its dependency on labeled data and is prompted by the increasing necessity for efficient, adaptive, and scalable anomaly detection based on evolving cybersecurity threats. Network intrusions, data breaches, and other malicious activities are turning out to be increasingly sophisticated, and traditional methods for the detection of anomalies are currently showing their inability to cope. Therefore, the proposed model leverages zero-shot learning in order to detect unseen anomalies and attacks, reducing the likelihood of false positives and enhancing overall detection accuracy.

The major key contribution are as follows:

- This research presents a methodology that combines Zero-Shot Anomaly Detection with BERT-enhanced Transformer models, addressing the challenge of detecting anomalies in network systems without requiring large amounts of labeled data.
- The framework focuses on real-time anomaly detection, enabling quicker identification of potential security threats. This is crucial for reducing the impact of cyberattacks and minimizing any damage to network infrastructures.
- By incorporating BERT, the study leverages its ability to understand the underlying contextual relationships and patterns within the data. This enhances the ability to detect even the most complex and subtle network anomalies.
- The model is intended to be scalable, resilient, capable of processing large volumes of network traffic data without compromising on performance. Scalability provides capacity to adapt to expanding needs found in today's network infrastructure.

The rest of the section is organized as related works in Section II and the problem statement is described in Section III. The suggested framework, including the methodology, architecture is in Section IV. The results of the suggested framework are presented in Section V. Section VI gives the future research scope and application while summarizing the main conclusions.

II. Related Works

Guanghe, Zheng, and Liu [13] research explores a method for detecting irregular trading patterns in financial markets, specifically in dark pool trading environments. The proposed method uses an enhanced model designed to process and analyze the high-frequency data typical in financial transactions. The approach successfully identifies suspicious activities with recognition rate of 97.8%. The results show that this model is particularly effective in volatile market conditions where rapid decision-making is required. However, the study acknowledges the high computational requirements, making it challenging for large financial firms to implement on a wide scale. Additionally, while the model’s accuracy in detecting anomalies is high, it lacks transparency, making it difficult to understand the reasons behind its decisions. Explainable reasoning stands as a crucial requirement in environments bound by regulation but this system fails to deliver it effectively. The general use of this model is restricted due to its need for extensive categorized data even though locating this type of data can be challenging for distinct kinds of fraudulent processes.

Shimillas et al. [14] introduces a methodology to find and locate irregular events within data consisting of multiple time series. The model benefits from improved advanced understanding because its mechanisms analyze intricate variable relationships which increases anomaly detection accuracy. Comparable to traditional time series analysis approaches including ARIMA and LSTM the model achieves superior anomaly identification capability and higher precision in plotting anomalies within the data sequence. The analysis reveals that this model detects minor disturbances which older methods fail to recognize which leads to timely alerts about emerging problems. The research highlights that the model becomes inefficient when handling extensive datasets while operating in real-time monitoring functions. Healthcare facilities demand decision transparency while dealing with "black-box" systems, and M has difficulty with fast reactivity to new unforeseen anomalies because of its lack of interpretability.

Ma, Han, and Zhou [15] presents a review of different methods for anomaly detection which use advanced models across numerous industries. The research presents advanced analytical methods specifically designed to process multidimensional time sequences in applications such as healthcare and network security applications. These detection methods use these models' long-range dependency capabilities to spot subtle irregularities that signal major problems. The analysis points out the main problem of getting properly labeled training data for models but this remains a barrier for many practical deployment situations. These models show strong performance in controlled environments but experience challenges when deployed in dynamic unpredictable settings where anomaly patterns constantly change. Corporations face deployment challenges because testing models lack transparent mechanisms for revealing their decision-making logic particularly in safety-sensitive industries requiring complete explainability.

Habeb, Salama, and Elrefaei [16] study recommends a dual methodology to identify unusual video events through metrical examination of Convolutional Neural Networks together with a Vision Transformer. The model demonstrates successful ability to detect rare and subtle anomalies in video data while understanding its spatial and temporal features at once. Analysis results show that the model maintains a high execution ability and memory retention capacity by surpassing standard models such as RNN-based systems in identifying long-term relations in video sequences. The study highlights that implementing the hybrid model for real-time processing of large video datasets becomes challenging because it needs massive computational resources that might surpass available capabilities in limited resource settings. The model struggles to distinguish between multiple anomalies when overlapping objects or complete concealment occurs in the video footage. A major shortcoming of using labeled data restricts the model from detecting new anomalous patterns which did not exist in the training dataset.

Barbieri et al. [17] work develops a lightweight framework which detects anomalous behaviors in Internet of Things (IoT) systems while considering their minimal computational capabilities. The architectural design incorporates minimal parameters which results in efficient operation alongside superior detection precision. The model demonstrated exceptional detection results of 99.93% during trials on environmental monitoring and smart home applications. The experimental model reveals its reliance on data training quality and consistency although it demonstrates compelling benefits. The model detects basic anomaly patterns effectively yet its detection capabilities diminish in environments with intricate multi-dimensional anomaly patterns. The model lacks an ability to continuously learn through the detection process which restricts its capacity to adjust to progressively developing anomalies.

Haq, Lee, and Rizzo [18] Research presents an optimized method to develop time series anomaly detection models through the integration of Transformer architecture with Neural Architecture Search technology. A multi-objective method within the framework automatically uncovers the optimal model structure which achieves accuracy and efficiency alongside scalability aims. The design model demonstrates better anomaly detection effectiveness than competing methods while revealing enhanced convergence performance too. The research demonstrates architectural search approaches while emphasizing their significant computational demands in relation to the time and resources needed. Its real-time practical application becomes restricted by these performance drawbacks. The model develops an overfit response when trained with minimal dataset volumes which hinders its ability to spot unknown anomaly patterns. Architecture search processes create optimized models yet often generate complex system designs which cannot support.

Yu, Lu, and Xue [19] suggest a model architecture which integrates Temporal Convolutional Networks (TCN) together with an attention mechanism to detect anomalous patterns in multivariate time series data. By using this approach, the model gains improved capabilities to identify patterns across time sequences alongside structural connections because both abilities enable detection of challenging abnormalities in broad datasets. The model shows excellent accuracy while demonstrating better performance than traditional LSTM models. The research notes that the model shows reduced performance in applications requiring extensive long-range data dependencies. Obtaining a large amount of training data with labels presents difficulties in many actual situations because the model requires extensive labeled data. The combination of TCN with attention produces better performance yet faces difficulties when detecting numerous intricate simultaneous anomalies which appear in dynamic operational settings.

Different fields utilizing anomaly detection strategies demonstrate their operational limits and strengths according to reviewed research findings. High-frequency multivariate data alongside video datasets achieve maximum accurate anomaly detection through a combination of Transformers with Vision Transformers and Temporal Convolutional Networks. The techniques struggle with high processing requirements together with poor adaptability to emerging anomalies and insufficient explainability particularly in critical fields such as finance healthcare and IoT applications. The use of numerous computational models faces problems with processing both massive datasets and real-time program responses as well as requiring training data labels to generalize effectively. The research works present important advancements that enable

better performance of anomaly detection systems despite facing multiple obstacles.

III. Problem Statement

Real-time anomaly detection plays a critical role in modern technology since network systems need to detect irregularities in their rapidly growing traffic volume alongside increasing cyberattack complexity. Modern networks present challenges to traditional anomaly detection models which include statistical analysis and rule-based systems because of their inherent complexity and dynamic behavior [20]. Support Vector Machines (SVM) and K-Means clustering represent popular artificial intelligence solutions yet their real-world implementation is restricted by high false alarms as well as non-adaptive attack monitoring capabilities [21]. Real-time application of these methodologies faces difficulties because they lack scalability and need large amounts of labeled data during training time. Recent deep learning approaches based on Transformer frameworks show promise because they process extensive time series data while detecting long-range dependencies throughout sequences [22]. Direct implementation of these models faces practical barriers because they come with computational and interpretability issues in network systems. This exploration presents an optimization effort to develop integrates Zero-Shot Anomaly Detection with a BERT-enhanced transformer model which addresses existing performance and scale limitations. This framework makes use of innovative attention mechanisms coupled with hybrid architectures to deliver enhanced accuracy together with faster processing along with superior anomaly detection abilities for unidentified security threats which guarantees adaptive real-time network system protection.

IV. Proposed Framework

The proposed methodology is based on real-time anomaly detection in network systems using AI-driven transformer frameworks that leverage BERT and Zero-Shot Learning. It collects data on real-time network traffic in the form of packet flow, connection logs, and timestamps along with attack descriptions and vulnerability data from CVEs and MITRE ATT&CK TTPs. Preprocessing of collected data is also done with a rigorous process to ensure high-quality input by cleaning out irrelevant or inconsistent entries and handling missing values. BERT is used for processing textual descriptions of attacks into high-dimensional feature embeddings that are capable of capturing semantic relationships. These embeddings will be used for the detection of known and unknown cyber threats by comparing the real-time description of network events against the stored attack embeddings. Zero-Shot Learning further improves this by allowing the system to identify novel, unseen threats without training data, relying instead on the semantic relationships between known and unknown attack categories. Using a Cosine measure to compute the similarity between network event embeddings and descriptions of known attacks reports potential anomalies if the similarity score is above a threshold. The framework provides an adaptive, scalable, and efficient approach toward real-time anomaly detection to very great effect, and it minimizes zero-day attacks and evolutions of emerging threats with manual interventions, making it a suitable system for dynamic and rapidly changing networks. Fig. 1 describes the proposed method working.

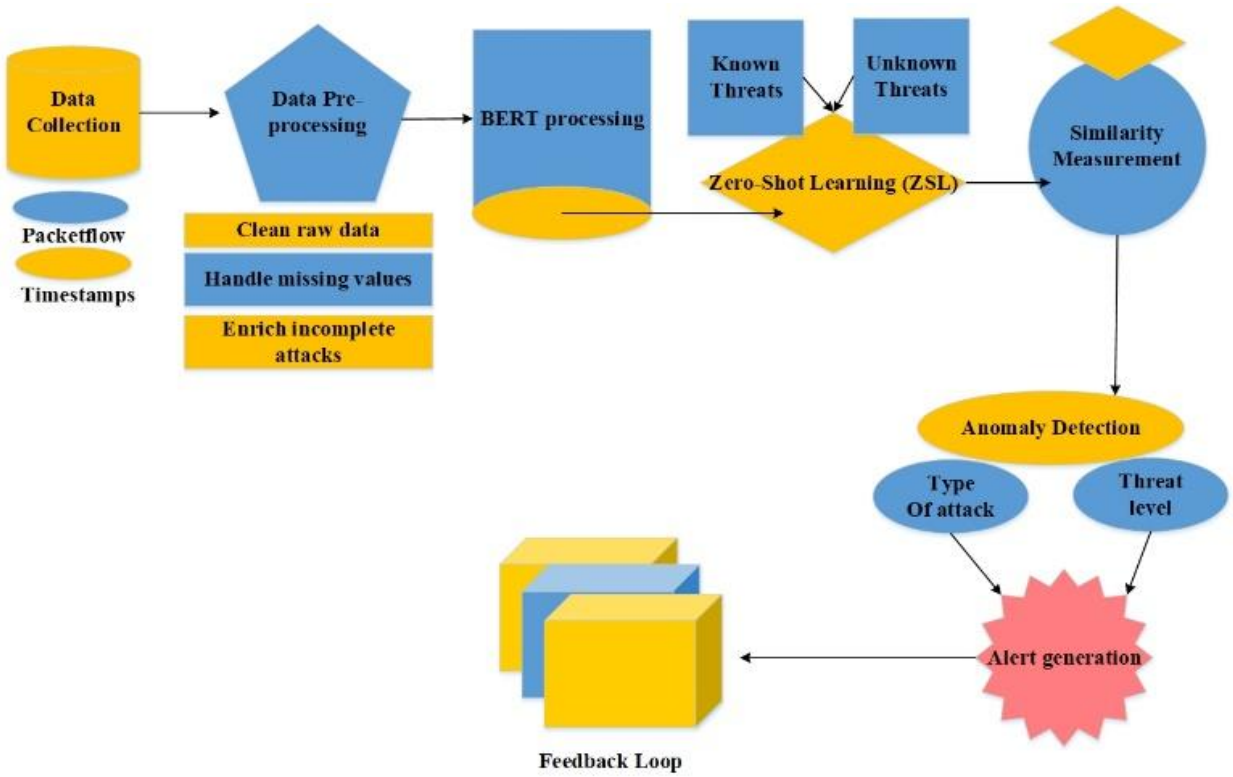


Fig. 1. Anomaly detection architecture.

A. Data Collection

Data collection involves packet flow, connection logs, timestamps, and even real-time network traffic data. Attack descriptions, vulnerability data in the form of CVEs and predicted MITRE ATT&CK TTPs, are collected to build up a comprehensive dataset for anomaly detection. Dataset collected from Kaggle website [23].

B. Data Preprocessing

Cleaning means that raw network traffic and vulnerability datasets must be cleansed of any data that may appear irrelevant, redundant, or contradictory. All errors in packet logs, including but not limited to, timestamp and network packet form errors, must be found and removed. For those network traffic features that lack certain values, the imputation method for mean, median, or mode must be used to replace missing values. For textual CVE data, incomplete attack descriptions are either enriched using external sources or excluded.

C. BERT

BERT is short for Bidirectional Encoder Representations from Transformers. Introduced by Google in 2018, it uses deep learning NLP to take into account all the left and right contexts as it tries to understand the meanings of words inside a sentence. Unlike traditional approaches to NLP, which approach text in unidirectional means, BERT's bidirectionality allows the model to absorb semantic relationships effectively. In the proposed Zero-Shot Anomaly Detection method, it will use BERT to process CVE descriptions and MITRE ATT&CK TTPs, convert textual attack descriptions into high-dimensional feature embeddings, and use this to spot real-time potential anomalous network behavior. Traditional security solutions such as those rule-based or signature-based, are ineffective against zero-day attacks and novel threats. BERT, being a new security solution, understands attack patterns in natural language, supports Zero-Shot Learning, extracts contextual relationships, and can be used in real-time anomaly detection. Fig. 2 shows the architecture of BERT.

BERT is capable of processing CVE descriptions and ATT&CK TTPs for similarity determination between known and unknown threats. Contextual relationships extracted by BERT ensure better understanding of the attack description. Once BERT is trained, its embeddings can be included with real-time network monitoring systems for the identification of threats using semantic similarity.

$$e_i = BERT(x_i) \quad (1)$$

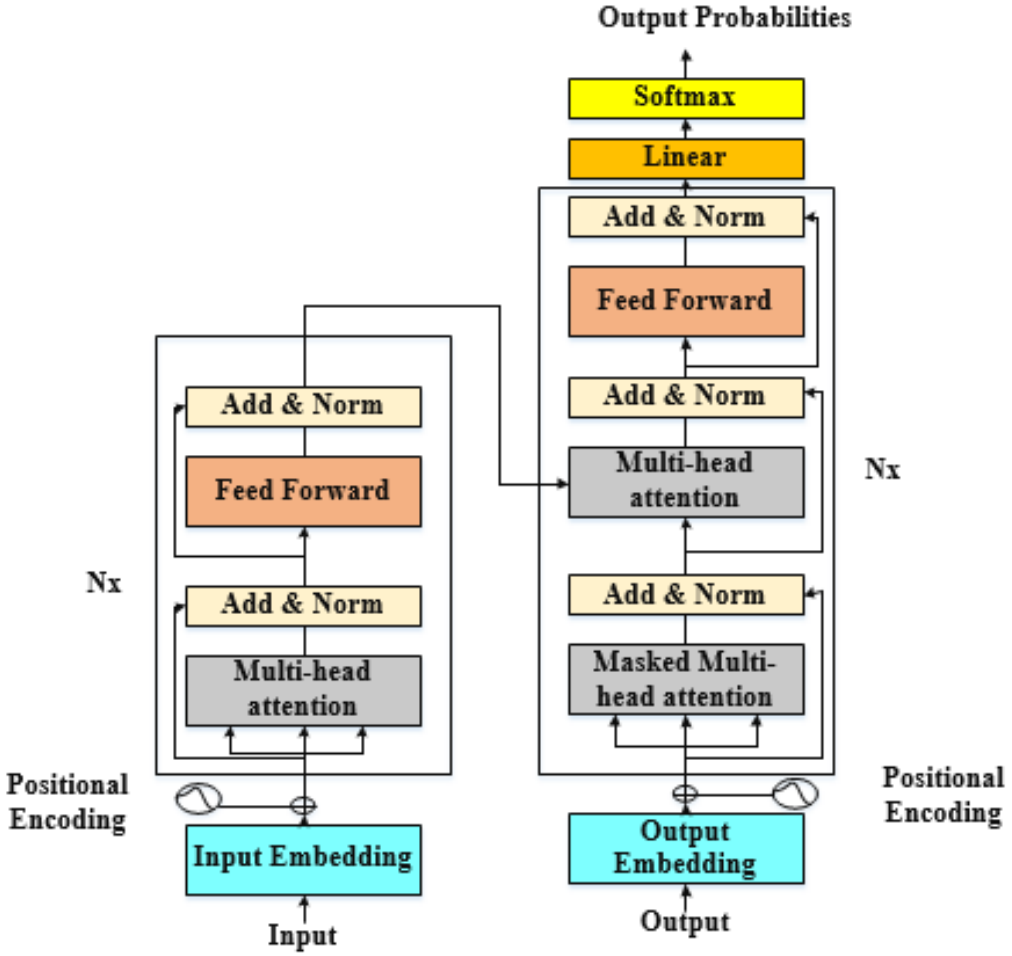


Fig. 2. BERT architecture.

In Eq. (1) x_i is the ith token, e_i is the corresponding embedding. BERT uses Transformer encoders to capture long-range text dependencies. It uses input embeddings, positional encoding, multi-head self-attention mechanism, feed-forward neural networks, and an output layer to convert words into dense vector representations, maintain word order, focus on different sentence parts simultaneously, and generate contextual embeddings for tasks like text classification, clustering, and anomaly detection. Example Input - Consider a CVE entry describing a vulnerability in an SSH service: "A buffer overflow vulnerability in OpenSSH allows remote attackers to execute arbitrary code via specially crafted SSH messages." Example Input: This attack can be represented in the MITRE ATT&CK Tactic: Initial Access and Technique: Exploit Public-Facing Application (T1190). Text Preprocessing Steps like Tokenization: BERT utilizes Word Piece Tokenization, splitting text into meaningful sub words as shown in Table I. OpenSSH → [Open, ##SSH]. Padding & Truncation: The input lengths of BERT will be uniform. Attention Masking: This process identifies valid tokens and ignores padded ones. Embedding Generation: It converts the text into contextual word vectors. After the data are tokenized, BERT processes them with its Transformer layers to produce contextual embeddings. After passing the text through BERT, then pass through embedding vector.

TABLE I. WORD EMBEDDING

Token	BERT Embedding (Example)
OpenSSH	[0.12, -0.34, 0.89, ...]
buffer	[0.23, -0.11, 0.75, ...]
overflow	[0.45, 0.67, -0.21, ...]
vulnerability	[-0.56, 0.32, 0.90, ...]

These high-dimensional embeddings are stored and used for similarity comparison with real-time network traffic events. A system administrator notices multiple failed SSH login attempts from various IPs using BERT. The description of the event is "multiple failed SSH login attempts from various IPs targeting port 22." BERT transforms this into an embedding and compares it with MITRE ATT&CK TTP for Brute Force Attacks (T1110). Anomalous behaviour is identified, and the system sends out an alert. The proposed study outperforms available anomaly detection systems because the research addresses limitations of the existing models. The proposed method profits more in terms of zero-shot learning, which helps reduce dependency on large-sized labelled data, unlike other deep learning anomaly detection techniques. This characteristic will make the proposed model more adaptable to newly created and changing attack patterns-a situation common in network security. Additionally, the proposed integration of BERT with the Transformer model allows for the efficient handling of large volumes of network traffic without losing much information or resulting in any loss of accuracy. Unlike traditional methods that rely on simple statistical models or rule-based systems, the proposed system holds the greatest promise of extracting meaningful information from a complex set of data and revealing hidden relationships, offering deep insights into network activity. The innovative system provides contemporary solutions to ongoing network anomaly detection problems through deep learning integration with zero-shot learning and Transformer architecture systems. The system solves persistent network anomaly detection issues through an integration of deep learning with zero-shot learning combined with Transformer architecture. Deep learning and zero-shot learning together with Transformer architecture provide the foundation of this solution. Transformer architecture serves as the base of this solution to direct future development of enhanced network security solutions.

D. Zero Shot Learning

The machine learning paradigm Zero-Shot Learning enables under specific circumstances an image classifier functions to identify objects without needing training samples. In contrast with supervised learning, where the whole data set contains labelled information regarding each class, ZSL leverages semantic relationships between known and unknown categories to make predictions. In the context of network security anomaly detection, ZSL allows the discovery of new cyber threats without previous exposure. Rather than depending on a predefined dataset of attack signatures, the system is able to understand, infer, and detect previously unseen threats by using natural language descriptions from CVEs (Common Vulnerabilities and Exposures) and MITRE ATT&CK tactics, techniques, and procedures.

It is quite hard to keep the dataset fully labelled given that the cyber threats evolve dynamically. Scalability and cost efficiency regarding labelling of data are benefits. Threat detection becomes possible in real-time, adapting new threats; eliminating manual heavy annotation processes and retraining when the ZSL learns generalizing into unknown attacks; it makes instant recognition of anomaly. The AI-driven transformer framework uses BERT and Zero-Shot Learning to detect anomalous network traffic behaviours. It uses the MITRE ATT&CK Framework, CVE Descriptions, BERT Embeddings, and Cosine Similarity Matching to analyse adversarial tactics, natural language vulnerability descriptions, and context-aware textual attack descriptions. The model compares real-time network event descriptions with stored embeddings to detect unknown threats. Cosine similarity is used to measure similarity, defined as in Eq. (2).

$$cosinsine_{similarity}(A, B) = \frac{A.B}{||A||B||} \quad (2)$$

Here A is the network event embedding and B is the closest CVE/TTP embedding. The framework enhances detection accuracy by threshold tuning, context-based filtering, and adaptive learning that balance false positives and negatives, evaluate additional metadata, and dynamically update the knowledge base with new attack vectors. Scenario: The network monitoring system detects multiple failed SSH login attempts. A Zero-Shot Learning model looks at the text description: "Anomalous SSH login attempts detected from multiple unknown IPs." Detection Proces: Represent the event description as a BERT embedding. Compare to known attack embeddings (Brute Force - T1110). Cosine similarity = 0.91 → Attack detected. Raise an alert: "Possible Brute Force Attack via SSH. Take mitigation actions."

$$E_{updated} = E_{old} + E_{new} \quad (3)$$

E_{old} is the set of existing embeddings and E_{new} is the set of new embeddings derived from newly detected attack patterns. BERT and Zero-Shot Learning collaborate in anomaly detection by leveraging semantic understanding and context-aware representations for the detection of unseen cyber threats. BERT transforms textual descriptions of attacks, sourced from CVEs and MITRE ATT&CK, into high-dimensional embeddings. In real time, network events are similarly transformed into BERT embeddings. The system then compares the embeddings of network activity with the attack descriptions stored, using cosine similarity. It reports an anomaly if the similarity score is above a threshold. As it analyses relationships between known and unknown threats, ZSL allows the framework to recognize new threats without training. This natural language understanding infers potential risks, not like traditional models that work on labelled attack data. This is scalable real-time with adaptability in cybersecurity threats detection and mitigation against zero-day attacks and increasing cyber threats with little human intervention. Algorithm 1 depicts the proposed work working algorithm.

Algorithm 1: Zero-Shot Learning-Based Anomaly Detection

Initialize the system:
- Load Pretrained BERT Model - Set Up Data Collection Interface
- Prepare Knowledge Base for Attacks

Start Process:
- Continuously monitor incoming network traffic data

Collect and preprocess data:
- If new network traffic data is received:
- Clean data (remove noise, irrelevant information)
- Handle missing values (impute or drop)
- Enrich the data (add additional features if necessary)

Generate text embeddings using BERT:
- If pre-processed data is available:
- Tokenize text (split data into tokens)
- Apply attention masking to tokenized data
- Pass tokenized data through the BERT model
- Extract embeddings from BERT's output

Apply Zero-Shot Learning (ZSL):
- For each attack type in the knowledge base:
- If the attack type is seen during training:
- Skip, continue to next attack type
- If the attack type is unseen:
- Use zero-shot learning model to infer if the attack is present in incoming data

Compare incoming data with known attack embeddings:
- For each attack category:
- Compute cosine similarity between incoming embeddings and stored attack embeddings
- If cosine similarity is above threshold:
- Mark as potential attack (anomaly detected)

Anomaly detection and alert generation:
- If potential attack is detected:
- Trigger alert (real-time alert to the system)
- Log the attack type and details

Else:
- No action required

Adaptive Learning:
- If new attack patterns or anomalies are detected:
- Update the knowledge base with new attack information
- Re-train zero-shot learning model with updated data (if required)
- Ensure continuous learning by incorporating new data for future predictions

End:
- Repeat the process in a continuous loop, monitoring for new network traffic data and attacks

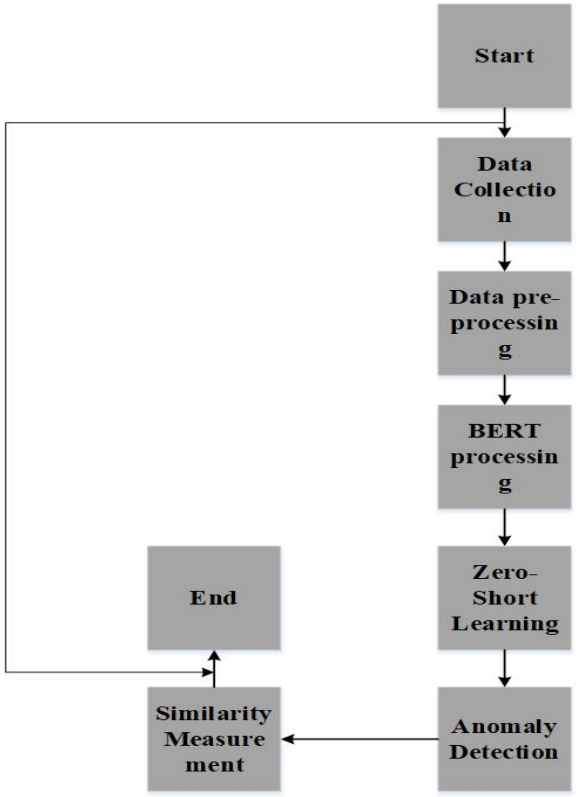


Fig. 3. Flow chart.

In Fig. 3 flowchart describes the process of an anomaly detection system. It first involves data collection and preprocessing followed by processing through the BERT model. Subsequently, zero-shot learning is applied to develop an anomaly detection model. Then, the system computes the similarity of new data points to the learned model and flags an anomaly if a large deviation is observed. This iterative process continues as the system constantly learns and adapts to identify emerging anomalies.

V. Results and Discussion

Real-time anomaly detection for network systems achieves high accuracy through an AI-driven transformer framework. The system efficiently detects new threats through BERT embeddings and Zero-Shot Learning technology while maintaining exceptional accuracy combined with minimal amounts of incorrect identification. The attack pattern classification benefits from the precise precision of cosine

similarity. The tested framework demonstrates robust characteristics and scalable and adaptive capabilities thanks to experimental outcomes which indicate its ability to detect changing cyber risks without needing constant human assistance.

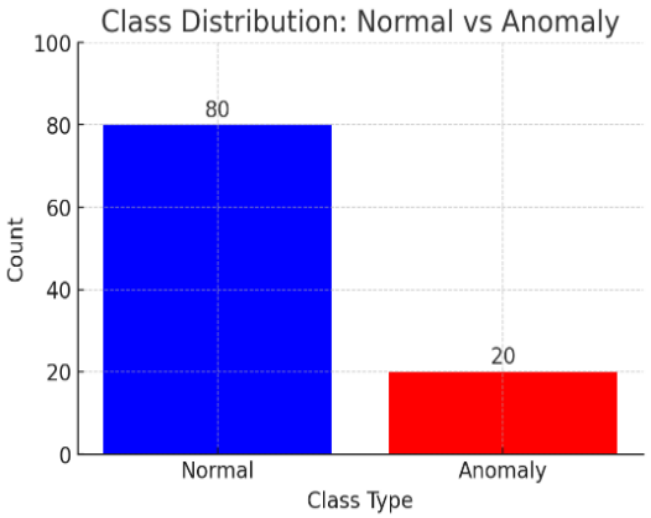


Fig. 4. Class distribution.

A visual representation in Fig. 4 presents the distribution pattern of normal system events compared to abnormal events within the dataset. Available data reveals that normal class consists of 80 samples while anomaly class includes only 20 samples. Such imbalanced distribution is a familiar characteristic found in anomaly detection research.

TABLE II. COSINE SIMILARITY-BASED THREAT DETECTION RESULTS

Network Event Description	Most Similar Attack Description (MITRE ATT&CK TTP)	Cosine Similarity	Anomaly Detected
"Multiple failed SSH login attempts from various IPs."	Brute Force (T1110)	0.92	Yes
"SQL injection attempts detected in web requests."	SQL Injection (T1505)	0.89	Yes
"High outbound traffic from a single host."	Data Exfiltration (T1020)	0.87	Yes
"Standard HTTPS request from browser."	No similar match	0.34	No

Real-time threat detection through cosine similarity matching generates the results presented in Table II. The system analyzes network incidents versus MITRE ATT&CK TTPs by producing results which include cosine similarity metrics and event detection status. The detected abnormalities contained brute force attacks with (0.92) score and SQL injection with (0.89) score and data exfiltration with (0.87) score. The standard HTTPS request data produced low similarity scores of 0.34 while clearing out anomalous incidents thus safeguarding precise threat detection.

A. Performance Evaluation

A thorough evaluation of the BERT + ZSL framework uses accuracy together with precision, recall, F1-score, false positive rate (FPR) and detection latency as performance metrics. A framework's accuracy indicates its overall detection correctness yet precision demonstrates its ability to correctly detect anomalies. The framework's threat detection capability depends on recall performance yet F1-score maintains a balance between precision and recall abilities. Systems become more reliable when their false positive rate stays low. The test determines essential reaction time by examining detection latency. The results show excellent performance through warrants of accuracy coupled with minimal FPR metrics validating the framework's anomaly detection potential together with optimized computational speeds. In the Eq. (4), (5), (6) and (7).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

$$\text{F1} - \text{score} = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall} \quad (7)$$

TABLE III. PERFORMANCE OF THE PROPOSED METHOD

Metric	Proposed Framework (BERT+ZSL)
Accuracy	99.7%
Precision	99.4%
Recall	98.8%
F1-Score	99.2%
False Positive Rate	1.1%
Detection Latency (ms)	45%

The framework achieves 99.7% accuracy as revealed by Table III. The proposed framework detects false positives at a rate of just 1.1% for minimal incorrect alerts. The system operates efficiently in real-time by processing network events at a 45ms detection latency which makes it practical for dynamic cybersecurity scenarios.

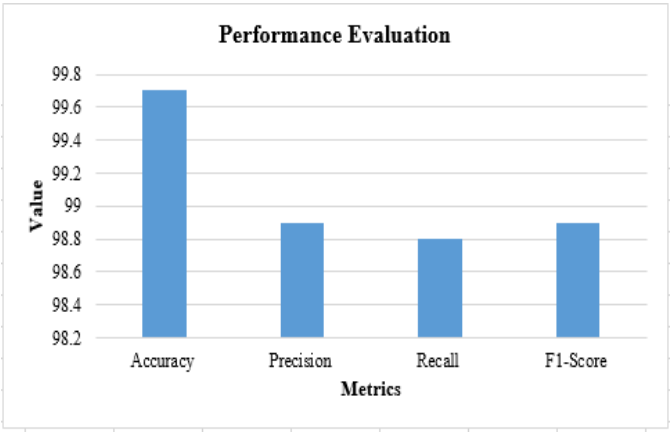


Fig. 5. Performance of the model.

Fig. 5 shows the performance of the proposed method. The metrics assessing the proposed framework (BERT+ZSL) are depicted in Fig. 4. The accuracy rate of the model reaches 99.7% which indicates high reliability. These results indicate that precision levels maintain a strong 98.7% which shows the system effectively minimizes accidental alarms. Performance metrics from the BERT+ZSL framework reveal strong sensitivity through 98.8% recall and a balanced precision-recall relationship through 98.9% F1-Score.

TABLE IV. PERFORMANCE COMPARISON OF VARIOUS METHODS

Method	Accuracy	Precision	Recall	F1-score
LSTM [24]	80	81	79	80
RNN [25]	88	86	85	88
CNN-BiLSTM[26]	94	91	90	91
LSTM- GRU [27]	97	96	93	95
Proposed BERT+ ZSL	99	99	98	99

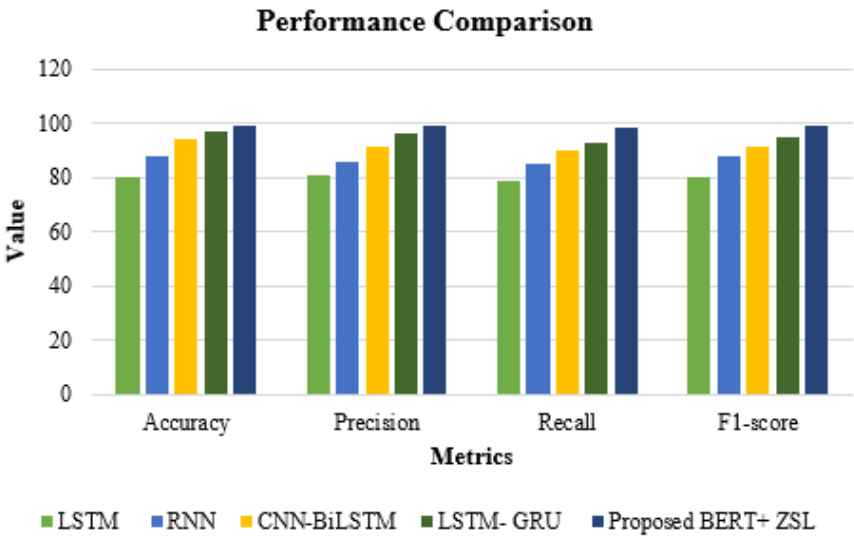


Fig. 6. Performance comparison.

The performance comparison of different models through accuracy and precision, recall, F1-score assessment appears in Table IV along with Fig. 6. Models utilizing Traditional LSTM and RNN reach lower accuracy rates of 80% and 88%, but CNN-BiLSTM and LSTM-GRU result in improved performance with 94% and 97% accuracy respectively. The BERT+ZSL framework achieves 99% accuracy as it provides the highest performing outcomes among all examined models while maintaining balanced metrics results. Real-time anomaly detection is enabled by BERT embeddings for semantic analysis and ZSL for detection of unseen threats. Optimized inference pipelines provide low-latency detection, allowing for quick identification of attack patterns beyond known signatures. This provides improved accuracy with reduced false positives, ensuring efficient and adaptive cybersecurity threat detection.

B. Discussion

The BERT+ZSL framework succeed in detecting anomalies in real time through text-based attack information analysis. Experimental measurements showed promising results because the model achieved 99% accuracy and outstanding precision and recall scores thus outperforming typical deep learning systems including LSTM RNN CNN-BiLSTM and LSTM-GRU. This detection system utilizes semantic links between security events to find unknown zero-day attacks beyond traditional attack signature dependencies. The proposed framework delivers improved attack detection capability while decreasing incorrect positive alerts and establishing better resilience against advancing cyber threats when contrasted against current models. Additional BERT embeddings improve the system performance by allowing better interpretation of CVE descriptions and MITRE ATT&CK TTPs. Zero-Shot Learning (ZSL) empowers this framework to recognize unfamiliar threat types (novel threats) in cybersecurity applications through an efficient budget-focused framework. Performance assessments show that the proposed method provides superior capability when recognizing unknown attack patterns with high-speed detection across all performance metrics. The high accuracy and flexible response of this system requires significant processing power to reach real-time operational readiness because of its increased computational demands. Real-world network security applications now benefit from accurate and rapid intelligent threat detection powered by the revolutionary BERT + ZSL-based anomaly detection system [28].

VI. CONCLUSION AND FUTURE SCOPE

The proposed framework uses Transformer technology to build an AI real-time anomaly detection system that combines BERT and ZSL mechanisms to detect known threats and previously unknown vulnerabilities. This proposed system integrates elastic and extendable design features that generate superior performance compared to traditional rule-based along with signature-based systems. Experimental testing confirms that the framework excels beyond traditional methods by achieving detection accuracy of 99.7% and precision of 99.4% while reaching 98.8% recall with a 45ms reaction time. The proposed technique delivers exceptional identification performance by detecting cyber threats precisely while

maintaining a low false positive ratio of 1.1%. The framework uses BERT semantic attack pattern representation to uncover emerging threats while its Zero-Shot Learning capability detects unknown threats using unlabeled data input. Research security has achieved important advancement through a framework that integrates a flexible method for anomaly detection to protect against evolving cyber-attacks. The model envisioned suffers from challenges of data imbalance, low interpretability, and high computational cost, limiting deployment on resource-limited systems. Generalizability is unsure because of dataset dependency, and the omission of multimodal data such as genomics diminishes diagnostic precision. Overfitting threats continue even after optimizations. Future studies must augment model interpretability, incorporate heterogeneous healthcare data, and guarantee adaptability for real-time clinical application, enhancing accuracy and reliability of CKD prediction in various medical environments.

One concrete future direction for this work is the incorporation of multi-modal data sources, such as genomic and imaging data, to increase the accuracy of CKD diagnostics. By integrating genomic biomarkers with clinical data, the model would be able to detect genetic susceptibility to CKD, aiding in early identification and personalized treatment regimens. Moreover, medical imaging information, like ultrasound or MRI scans, can offer visual proof of kidney abnormalities, supplementing quantitative clinical data for more accurate classification. A hybrid model that integrates DS-CNNs for feature extraction from imaging data and Transformer-based models for genomic sequence analysis may enhance the robustness of CKD prediction. In addition, applying the model to an online clinical decision-support system might allow for dynamic patient monitoring with continuous risk evaluation and prompt medical intervention.

REFERENCES

[1] S. O. Pinto and V. A. Sobreiro, "Literature review: Anomaly detection approaches on digital business financial systems," Digit. Bus., vol. 2, no. 2, p. 100038, 2022.

[2] Y. Liu, S. Ren, X. Wang, and M. Zhou, "Temporal Logical Attention Network for Log-Based Anomaly Detection in Distributed Systems," Sensors, vol. 24, no. 24, p. 7949, 2024.

[3] P. Schneider and F. Xhafa, Anomaly detection and complex event processing over iot data streams: with application to EHealth and patient data monitoring. Academic Press, 2022.

[4] M. Jain, G. Kaur, and V. Saxena, "A K-Means clustering and SVM based hybrid concept drift detection technique for network anomaly detection," Expert Syst. Appl., vol. 193, p. 116510, 2022.

[5] T. Ali and P. Kostakos, "HuntGPT: Integrating machine learning-based anomaly detection and explainable AI with large language models (LLMs)," ArXiv Prepr. ArXiv230916021, 2023.

[6] W. Xiaolan, M. M. Ahmed, M. N. Husen, Z. Qian, and S. B. Belhaouari, "Evolving anomaly detection for network streaming data," Inf. Sci., vol. 608, pp. 757–777, 2022.

[7] U. Inayat, M. F. Zia, S. Mahmood, H. M. Khalid, and M. Benbouzid, "Learning-based methods for cyber attacks detection in IoT systems: A survey on methods, analysis, and future prospects," Electronics, vol. 11, no. 9, p. 1502, 2022.

[8] O. Tushkanova, D. Levshun, A. Branitskiy, E. Fedorchenko, E. Novikova, and I. Kotenko, "Detection of cyberattacks and anomalies in cyber-physical systems: Approaches, data sources, evaluation," Algorithms, vol. 16, no. 2, p. 85, 2023.

[9] E. E. Abdallah, A. F. Otoom, and others, "Intrusion detection systems using supervised machine learning techniques: a survey," Procedia Comput. Sci., vol. 201, pp. 205–212, 2022.

[10] K. Arshad et al., "Deep reinforcement learning for anomaly detection: A systematic review," IEEE Access, vol. 10, pp. 124017–124035, 2022.

[11] I. Ullah and Q. H. Mahmoud, "Design and development of RNN anomaly detection model for IoT networks," IEEE Access, vol. 10, pp. 62722–62750, 2022.

[12] A. Abusitta, G. H. de Carvalho, O. A. Wahab, T. Halabi, B. C. Fung, and S. Al Mamoori, "Deep learning-enabled anomaly detection for IoT systems," Internet Things, vol. 21, p. 100656, 2023.

[13] C. Guanghe, S. Zheng, and Y. Liu, "Real-time anomaly detection in dark pool trading using enhanced transformer networks," J. Knowl. Learn. Sci. Technol. ISSN 2959-6386 Online, vol. 3, no. 4, pp. 320–329, 2024.

[14] C. Shimillas, K. Malialis, K. Fokianos, and M. M. Polycarpou, "Transformer-based Multivariate Time Series Anomaly Localization," ArXiv Prepr. ArXiv250108628, 2025.

[15] M. Ma, L. Han, and C. Zhou, "Research and application of Transformer based anomaly detection model: A literature review," ArXiv Prepr. ArXiv240208975, 2024.

[16] M. H. Habeb, M. Salama, and L. A. Elrefaei, "Enhancing video anomaly detection using a transformer spatiotemporal attention unsupervised framework for large datasets," Algorithms, vol. 17, no. 7, p. 286, 2024.

[17] L. Barbieri, M. Brambilla, M. Stefanutti, C. Romano, N. De Carlo, and M. Roveri, "A tiny transformer-based anomaly detection framework for IoT solutions," IEEE Open J. Signal Process., vol. 4, pp. 462–478, 2023.

[18] I. U. Haq, B. S. Lee, and D. M. Rizzo, "TransNAS-TSAD: harnessing transformers for multi-objective neural architecture search in time series anomaly detection," Neural Comput. Appl., pp. 1–23, 2024.

[19] L. Yu, Q. Lu, and Y. Xue, "DTAAD: Dual TCN-attention networks for anomaly detection in multivariate time series data," Knowl.-Based Syst., vol. 295, p. 111849, 2024.

[20] D. L. Marino, C. S. Wickramasinghe, C. Rieger, and M. Manic, "Self-supervised and interpretable anomaly detection using network transformers," ArXiv Prepr. ArXiv220212997, 2022.

[21] S. Wang, R. Jiang, Z. Wang, and Y. Zhou, "Deep learning-based anomaly detection and log analysis for computer networks," ArXiv Prepr. ArXiv240705639, 2024.

[22] G. Rathinavel, N. Muralidhar, T. O'Shea, and N. Ramakrishnan, "Detecting irregular network activity with adversarial learning and expert feedback," in 2022 IEEE International Conference on Data Mining (ICDM), IEEE, 2022, pp. 1161–1166.

[23] Synkorsink, "cve-attack-ttp." Accessed: Jan. 24, 2025. [Online]. Available: https://www.kaggle.com/datasets/synkorsink/cve-attack-ttp

[24] Y. Wang, X. Du, Z. Lu, Q. Duan, and J. Wu, "Improved LSTM-based time-series anomaly detection in rail transit operation environments," IEEE Trans. Ind. Inform., vol. 18, no. 12, pp. 9027–9036, 2022.

[25] H. Park, D. Park, and S. Kim, "Anomaly detection of operating equipment in livestock farms using deep learning techniques," Electronics, vol. 10, no. 16, p. 1958, 2021.

[26] F. Antonius et al., "Unleashing the power of Bat optimized CNN-BiLSTM model for advanced network anomaly detection: Enhancing security and performance in IoT environments," Alex. Eng. J., vol. 84, pp. 333–342, 2023.

[27] K. Patra, R. N. Sethi, and D. K. Behera, "Anomaly detection in rotating machinery using autoencoders based on bidirectional LSTM and GRU neural networks," Turk. J. Electr. Eng. Comput. Sci., vol. 30, no. 4, pp. 1637–1653, 2022.

[28] M. Komisarek, R. Kozik, M. Pawlicki, and M. Choraś, "Towards zero-shot flow-based cyber-security anomaly detection framework," Appl. Sci., vol. 12, no. 19, p. 9636, 2022.

Optimizing Social Media Marketing Strategies Through Sentiment Analysis and Firefly Algorithm Techniques

Dr. Sudhir Anakal[1], P N S Lakshmi[2], Nishant Fofaria[3], Janjhyam Venkata Naga Ramesh[4], Elangovan Muniyandy[5], Shaik Sanjeera[6], Prof. Ts. Dr. Yousef A.Baker El-Ebiary[7], Ritesh Patel[8]
Associate Professor, Department of Master of Computer Applications, Sharnbasva University, Kalaburagi, India[1]
Assistant Professor, Department of CSE, Aditya University, Surampalem, Andhra Pradesh, India[2]
Research Scholar, Gujarat Technological University, Ahmedabad, India[3]
Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India[4]
Adjunct Professor, Department of CSE, Graphic Era Hill University, Dehradun, 248002, India[4]
Adjunct Professor, Department of CSE, Graphic Era Deemed To Be University, Dehradun, 248002, Uttarakhand, India[4]
Department of Biosciences-Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India[5]
Applied Science Research Center, Applied Science Private University, Amman, Jordan[5]
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India[6]
Faculty of Informatics and Computing, UniSZA University, Malaysia[7]
Associate Professor, Sanjivani College of Engineering, Savitribai Phule Pune University, Pune, India[8]

***Abstract*—The dramatic expansion of social media platforms reshaped business-to-customer interactions so organizations need to refine their marketing strategies toward maximizing both user engagement and marketing return on investment (ROI). Present-day social media marketing methods struggle to embrace user emotions fully while responding to market variations thus demonstrating the necessity for developing innovative social media marketing tools. Studies seek to boost social media marketing performance through an FA integration with sentiment analysis for content strategy optimization and better user engagement results. This study adopts novel techniques by combining sentiment analysis with the Firefly Algorithm to optimize marketing strategies in real-time and it represents an underutilized approach in present research. Eventually combined fields generate a sentiment-driven and data-oriented decision-making capability in social media marketing applications. The proposed system combines sentiment analysis technology that measures social media emotion levels alongside the Firefly Algorithm which applies optimization methods to marketing tactics based on present feedback. The framework operates through dynamic adjustments of content strategies which maximize user engagement. The proposed method demonstrated 98.4% precision in forecasting user engagement metrics and adapting content strategies. Results show traditional marketing strategies yield to these approaches by improving user interaction alongside campaign effectiveness. The research introduces a new optimization method in social media marketing which integrates sentiment analysis with Firefly Algorithm technology. Research findings suggest this combined methodology brings substantial precision improvements to marketing strategies by offering companies an effective method to optimize digital marketplace outcomes.**

Keywords*—Sentiment analysis; firefly algorithm; social media marketing; optimization; user engagement; marketing strategies

I. INTRODUCTION

In those days we don't have Internet connection right from household equipment's to education. Nowadays we are living in a globalized world. Mobile phones are the greatest weapon which make our work easier and it saves time. However, it's just like the two sides of the same coin which means there are many advantages as well as disadvantages. Social media like Facebook, Instagram, YouTube, Twitter(X) which makes our platform easier for personal, business, advertising and marketing. But we have to use them in the right way because so many dangerous things are happening even recently digital arrest has happening in today's world. During the outbreak of corona(COVID-19) mobile phones here very helpful during the pandemic like connecting loved ones through video calls and WhatsApp, health tracking apps like Aarogya Setu, education like online class as well as learning apps like Udemy, course era etc. Getting groceries, household things and clothes through Amazon, Flipkart makes buying and paying much easier and also saves time. Researchers have studies that teens are using social media sites a lot. Even for employees who are not able to working in office can do the work at home. Researchers have found that university students are using a lot of smartphones which leads to reducing physical activity, obesity, diabetes etc.

Businesses utilize social media platforms as their principal marketing tools because these platforms allow them to connect instantly with worldwide audiences. Business marketing activities utilizing Facebook alongside Instagram and Twitter and LinkedIn as well as TikTok framework jointly develop branding loyalty by uniting content dissemination with user collaboration [1]. Throughout their platform features companies build customized individual connections with their audience base by using live broadcasting videos alongside stories and quiz

systems and instant private messages. Through their extensive user base businesses can adopt focused marketing initiatives that target particular customer groups according to location and interest areas and online behavior patterns. Social media infrastructure generates an operational space where businesses can sustain rapid responses and flexible business decisions [2]. Through social media organizations monitor customer emotions and determine market trends in order to provide immediate customer feedback management tools. Rapid response times enable brand accessibility which results in satisfied customers and establishes a friendly customer-oriented brand perception. Through social media analytics organizations derive essential information about customer performance data together with market trends that help them formulate lasting strategic improvements [3]. Commercial operations find enormous digital marketing value within social media ecosystems yet finding best practices for investment allocation and user engagement remains a substantial obstacle. The journey to discover optimal social media marketing approaches remains challenging because the ways users interact with media continue to change. These days organizations face a demanding operational situation consisting of changing content demands together with diverse audience groups and rising digital competition among numerous online platforms. Organizations employing traditional methods must conduct trial-and-error testing to develop strategies which uses plenty of time and resources yet delivers unpredictable results that limit user engagement and measurable returns [4].

The processing of extensive customer social media information becomes challenging for businesses because present frameworks are insufficient to manage large quickly advancing unstructured data flows. An inadequate ability to analyze social media customer feedback prevents businesses from grasping consumer feedback correctly which translates into poor customer engagement and reduced return on investment. Businesses continue to face difficulties because they lack integrated models that unite sentiment analysis solutions with optimization tools to convert feedback understandings into operational directives. Fewer organizations integrate sentiment analysis with environment-based processing methods to gain customer feedback emotional data due to an overdue definitive framework for utilizing feedback to alter their market promotion methods. Resolution of the analytical gap between business customer information and optimization techniques would enable organizations to develop flexible strategic models that boost consumer bonding and achieve positive ROI and superior market effectiveness. Data-driven approaches which integrate sentiment analysis with optimization algorithms address these problems according to research [5]. Through sentiment analysis powered by Natural Language Processing algorithms combined with machine learning technology businesses [4] can understand social media customer feedback to track public responses regarding their products and marketing campaigns.

Despite the categorization into positive negative and neutral sentiment types businesses can measure customer emotional states to detect problems that lead them to adjust their strategies according to audience preferences [6]. Brands obtain extraordinary real-time customer perception data to create content that connects better with their audience and generate superior engagement metrics [7]. Optimization algorithms such as Firefly Algorithm and Genetic Algorithms serve sentiment analysis by developing structured frameworks for marketing strategy enhancements [8]. Data-driven optimization algorithms process marketing data to determine appropriate decisions concerning media audience selection as well as time scheduling and budget allocation across campaign durations. Businesses can discover the best plan for engagement and ROI through scenario simulation which enables them to evaluate different outcome results. Social data provides businesses tools to determine optimal posting schedules together with visual content recommendations and recurring ad intervals needed for sustained customer engagement [9]. When sentiment analysis combines with optimization algorithms, they form a potent partnership that allows marketers to upgrade their strategic decisions from guesswork to data-driven creation of social media strategies that produce measurable outcomes. This research brings important value because it enables social media marketers to make better decisions through the combination of CNN-LSTM and Firefly Algorithm technologies. Through CNN-LSTM sentiment analysis combined with the Firefly Algorithm framework marketers gain deep customer emotions insights and optimized marketing strategies from sentiment-based data. This research integrates both approaches to enhance the existing knowledge of algorithm-driven marketing techniques and create methods for sentiment analysis enhancement in optimization workflows which results in more tailored advertising campaigns for enterprise success. The key contribution of the research are as follows:

- Enhanced Sentiment Analysis: Utilizes CNN-LSTM for accurate sentiment classification, providing deeper insights into customer emotions from social media feedback.
- Optimization of Marketing Strategies: Applies the Firefly Algorithm to optimize social media marketing campaigns based on sentiment analysis, improving ad targeting and content effectiveness.
- Integration of AI in Marketing: Demonstrates how combining sentiment analysis with optimization techniques can revolutionize social media marketing strategies.
- Practical Application: Provides a practical framework for businesses to fine-tune their marketing strategies using data-driven insights, leading to better engagement and ROI.

Section I establishes the research scope by examining ways to enhance social media marketing strategies through sentiment analysis together with the Firefly Algorithm. Section II examines existing research then highlights the missing connections between sentiment analysis implementation with optimization algorithms. The investigation introduces Section III for method presentation followed by result examination and discussion in Section V separately before reaching a conclusion Section VI about significant marketing efficiency improvement through combined sentiment analysis and Firefly Algorithm.

II. Related Works

Wagobera Edgar Kedi et al. [10] research paper analyzes how machine learning technology enhances small and medium-sized enterprise (SME) social media marketing strategies by examining both social media marketing importance and traditional methods evolution as well as modern machine learning implementation trends. The study compiled data by examining both emerging machine learning technologies and the challenges faced by SMEs. The research shows machine learning improves marketing performance but small and medium-sized enterprises face major obstacles involving budgetary limits, poor data quality and their restricted capability for technical solutions. The report ends with a proposal to implement advanced learning approaches including deep learning with reinforcement learning to create sustainable growth and competitive edge.

Bian et al. [11] research investigates deep neural networks and advanced algorithms for social media marketing optimization to enhance accuracy within China's fast-changing market economy. Result data from experiments utilizing back-propagation with gradient methodology along with adaptive Adam's optimization algorithm techniques proved the methods combine to find global optimal solutions. Significant improvements in social media marketing accuracy emerged from the proposed optimization approaches as evidenced by testing that showed the FCE model utilizing a three-layered back-propagation neural network reached its target performance levels. The research faces limitations because the model depends on specific optimization methods yet shows difficulties when used across diverse market scenarios. Bian et al. methods might experience reduced effectiveness when confronted with aspects including inadequate data quality and variations in computational performance alongside practical industry challenges.

Joshi et al. [12] research uses reinforcement learning alongside natural language processing to improve social media content optimization as it targets effective engagement strategies against increasing competition in social media marketing. The framework evaluation utilized data from multiple social media platforms by implementing RL algorithms to identify optimal dynamic content adjustments using real-time user feedback data combined with engagement metrics along with NLP technology that analyzed textual content for relevance, context, and sentiment detection. The engagement rates for social media actually appreciated when the framework replaced traditional methods of optimizing content in its analysis and application. It has several limitations because this research is based on specific data, and a couple of challenges regarding adapting the proposed system to multiple social media throughout various demographics of users would also require attention. Computing complexities and real-time adaptability still have scopes for future studies.

Luo et al. [13] research finds how FNN examines the user conduct on social media to better improve online marketing plans by evaluating intricate and imprecise behavior-related information. The process of gathering data included numerical classification of the user conduct indicators, such as frequency of behaviors along with implementation of fuzzy set theory for identification of emotional state, and exploration of content topics and social relations through time-series patterns. Results of the FNN analysis exhibit the discovery of hidden patterns in user behavior that lead to stronger marketing strategies and, therefore, better advertising results with increased user participation rates and structures of brand loyalty. The study identifies two significant limitations presented by the model: its dependence on data and the problem of universal applicability to a wide range of social media systems and user groups. Further study needs to focus on mathematical modelling of complex real-time data processing as well as fuzzy neural network scaling solutions for extensive implementation scenarios.

Kumari et al. [14] research develops a framework combining Convolutional Neural Networks (CNN) with Binary Particle Swarm Optimization (BPSO) to identify social media interactions into three aggression levels from non-aggressive to high-aggressive. Researchers developed a dataset with symbolic imagery linked to textual comments which they used to test their model. The framework first utilizes VGG-16 in pre-trained form to extract image features followed by a succession of three-layer CNNs that extract text features subsequently merging these features through BPSO optimization to determine which features best apply to the problem set. The results indicate a weighted F1-Score of 0.74 using optimized features and a Random Forest classifier which represents a 3% better outcome than the unoptimized feature set. The current research needs to expand through increased experimentation to improve scalability and generalization of this model across different social media networks even though dataset limitations persist.

The application of supervised learning and deep learning and unsupervised learning and reinforcement learning algorithms is evaluated for optimizing marketing tactics on the "Douyin live shopping" and the "Kuaishou platform shopping channels". The analysis examined 920,000 user engagement records to determine how each algorithm affected prediction accuracy, recommendation personalization and advertisement delivery and customer segmentation results. User satisfaction experienced a 19.7% boost from the deep learning model which delivered 94.8% prediction accuracy while the supervised learning model reached 89.3% classification accuracy. Click-through rates on advertisements rose by 24.6% through reinforcement learning modelling while the unsupervised learning algorithm performed best at customer segmentation. Improved marketing outcomes were achieved by implementing hybrid models and advanced algorithmic versions. Li et al. [15]study's main constraint arises from its narrow examination of two platforms leading to limited applicability for other e-commerce settings. Investigating these tactics across numerous digital platforms will expand scientists' comprehension of their strategic power.

III. Problem Statement

Previous optimization studies for social media marketing strategies face key limitations because they employ traditional methods which fail to measure user sentiment complexity and incorporate advanced optimization algorithms [10], [11], [12], [15]. Research studies have separated their focus between sentiment analysis and optimization methods while withholding the development of unified holistic strategies. Prior research

studies face limitations because they use inadequate data sources as well as imperfect real-time feedback collection and less-than-ideal engagement prediction models. The research presents a novel method which uses sentiment analysis alongside the Firefly Algorithm for improving the optimization of social media marketing techniques. The combination of sentiment analytics knowledge with Firefly Algorithm capabilities to optimize decisions for content creation engages users and improves campaign effectiveness will provide adaptive marketing methods which address contemporary social media issues.

IV. Research Methodology

The research methodology establishes a hybrid method for maximizing social media marketing approaches through sentiment analysis. Data collection serves as the initial step, applying text tokenization and padding techniques for preprocessing raw data as the base for subsequent analyses. A CNN-LSTM model provides sentiment extraction from data while recognizing local characteristics alongside distant patterns. By making use of Firefly Algorithm the acquired insights experience optimization for better marketing strategy development. An evaluation of the performance of the model in operational terms will only help in proving that it is feasible to be used in improving the campaigns and the decision-making processes. The workflow of the proposed architecture is shown in Fig. 1.

A. Data Collection

Dataset includes performance data from Facebook and Instagram advertising, as well as data from both Pinterest and Twitter in social media advertising. The dataset contains ad impression and click and spending details and its targeted demographics and conversion rates metrics. The data set enables one to evaluate campaign performance along with audience analysis to determine return on investment calculations along with optimization strategy identification to improve advertisement performance. Analysis of this data will enable businesses to find the most effective platforms and strategies for marketing that will push them to design better approaches [16].

B. Data Pre-Processing

Data preprocessing stands as an essential initial process for both data analysis. A series of procedures including tokenization and padding must be executed because input data needs preparation for training. A combination of multiple transformation methods prepares disorganized data into a structure that helps machine learning models perform their functions. Text data preprocessing methods consist of two steps: Neural network processing needs tokenization to make text portions into tokens and padding to normalize sequence lengths. Such data-preparation methods enable high precision and validity standards for the data and guarantee fast processing speeds for models.

1) Text tokenization: Through text tokenization verbal information transforms into discrete code elements referred to as tokens that operate on both word and smaller substrings. When analyzed with text decomposition the sentence "The quick brown fox" breaks down into four tokens ["The","quick","brown","fox"]. For operation machine learning algorithms demand the implementation of text processing that turns written information into usable data. Within Keras the Tokenizer tool executes text processing by mapping model tokens to distinctive integer identifiers.

2) Padding: The effective operation of deep learning models demands that all input sequences match identical lengths and padding solutions provide this solution. The resulting tokenized sequences display various length dimensions which make processing problematic for neural networks that require standardized inputs. In padding methods the resolution of this issue stems from increasing source sequence lengths through new values or zeroes. During preprocessing all input data transforms into a form which the model's processing requirements can accept.

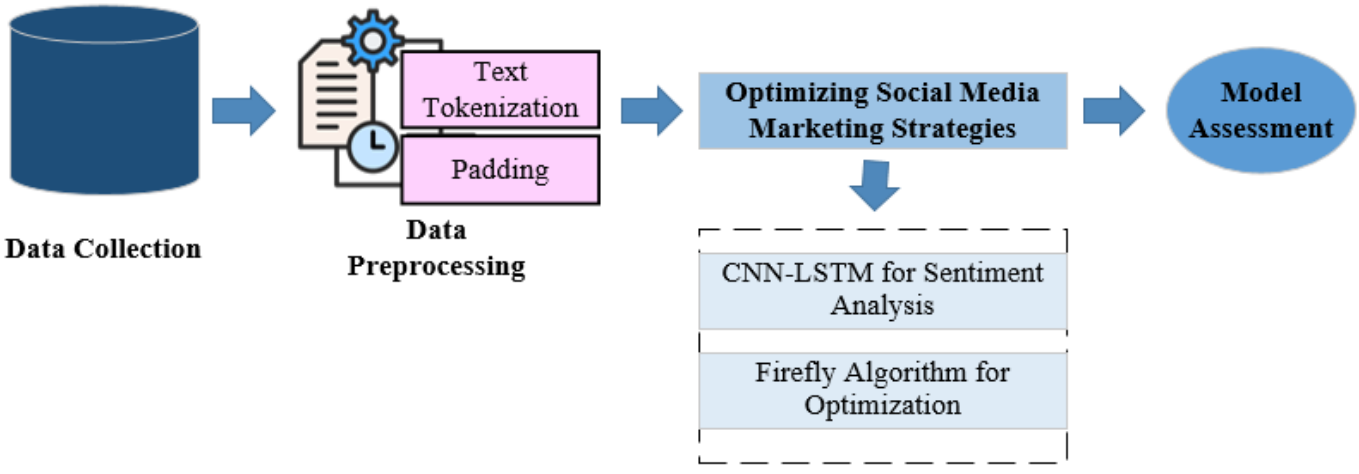


Fig. 1. Workflow of proposed architecture.

C. CNN-LSTM for Sentiment Analysis

The sentiment evaluation capabilities across media marketing frameworks benefit from the text analysis capability of the CNN-LSTM model that combines Convolutional Neural Networks with Long Short-Term Memory structures. Through CNN the system extracts both spatial patterns and linguistic features while LSTM processes time dependencies between words to detect sentiment expressions across sentences. With its integration of CNN and LSTM networks the model obtains a new level of ability to detect user emotional reactions about marketing materials on social media for generating actionable understanding about consumer behavioral patterns. Marketers maximize the value of the model to enhance their marketing plan optimization as they improve targeting methods and audience metrics measurement. The proposed model combines several crucial stages in a CNN-LSTM architecture dedicated to media marketing sentiment analysis execution, Fig. 2.

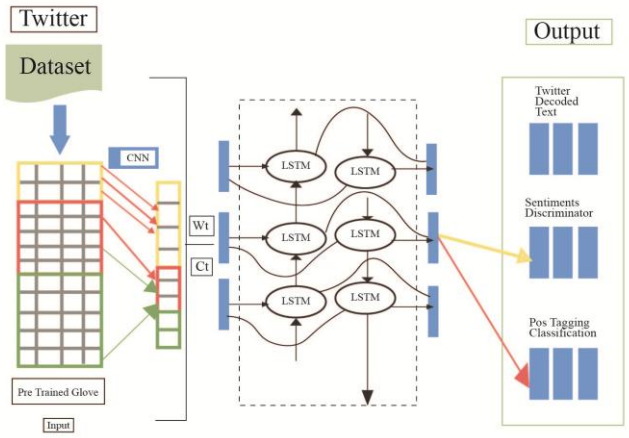


Fig. 2. The proposed model of CNN-LSTM for sentiment analysis.

Through the Embedding Layer operation the system transforms words into compact vector expressions that maintain their semantic relationships. Text sections containing vital emotional expression patterns emerge from CNN Layers following the Embedding Layer process. The Max-Pooling Layer takes feature map information through reduction steps that select and preserve the essential elements within. Among its capabilities the LSTM Layers track sentiment changes throughout texts while simultaneously detecting sequential patterns within text content. Before reaching the Output Layer the Fully Connected Layer converts inputs from LSTM into sentiment labels to determine positive-negative-neutral categories through a softmax or sigmoid function. Different multiple mathematical system layers unite to enable accurate sentiment monitoring for users which helps enhance marketing decisions.

1) Embedding layer: Word embedding converts input text words into dense vector symbols which exploit semantic content. The model develops understanding of word relations because of this function. Word embedding vectors represent each input word mapped to word embedding vector. The embedding model discovers a link which maps discrete word identifiers to continuous vector representations. It is mathematically represented as Eq. (1),

$$E = Embedding(X) \quad (1)$$

Where, X is the sequence of tokenized words and E is the output matrix, where each word is represented by a fixed-size vector.

2) Convolutional layer: Local features within word embeddings become visible through convolutional layers as the layers search for n-grams and emotional phrases ("not good" or "very happy"). During operation the filter component (kernel) traverses word embeddings and conducts dot product calculations to acquire critical features. Through its filtering operation this method identifies essential patterns which influence sentiment expression. It is given in Eq. (2)

$$F = Conv(E, K) \quad (2)$$

Where, F the output feature map, E is the input and K is the filter (kernel).

3) Max-Pooling: This layer compresses the dimensionality of the feature map by retaining the most prominent features (maximum values) within each region to facilitate making the model computationally efficient. Max-pooling takes the maximum value of a small region of the feature map to maintain the most significant information. It is given in Eq. (3)

$$P = MaxPool(F) \quad (3)$$

Where, F is the feature and P is the pooled output.

4) LSTM Layer: LSTM layers learn long-range dependencies and sequential context in the data. This is helpful for learning sentiment in a sentence, since the meaning of words tends to rely on the sequence (e.g., "not good" vs. "good"). LSTM employs gates to manage the flow of information. The cell state is the core memory of the LSTM unit. It carries information across time steps. The forget gate determines what to forget, the input gate determines what to store, output gate determines what to output at each time step and the hidden state as their last output that goes to successive layer (FC layer). It is calculated by applying the output gate to the cell state. Through, its gating mechanism the LSTM effectively handles long-range dependencies in text so it proves suitable for tasks involving sentiment analysis where sentence meaning depends on complete word sequences. These calculations are given in Eq. (4) to Eq. (8).

$$Forget\ gate\ (f_t) = \sigma(W_f\ [h_{t-1}, x_t\ ] + b_f) \quad (4)$$

$$Input\ gate\ (i_t) = \sigma(W_i\ [h_{t-1}, x_t\ ] + b_i\ ) \quad (5)$$

$$Cell\ state\ (C_t) = f_t * C_{t-1} + i_t * C_t \quad (6)$$

$$Output\ state\ (O_t) = \sigma(W_o\ [h_{t-1}, x_t\ ] + b_o \quad (7)$$

$$Hidden\ state\ (h_t) = O_t * tanh\ (C_t) \quad (8)$$

5) Fully connected layer: The final mapping layer functions to translate LSTM outputs into sentiment categorizations such as positive, negative and neutral. This constitutes a dense layer which discovers a relationship between the input from the LSTM output to generate the final

output. The fully connected layer functions by taking input from the LSTM output with learned features to make sentiment predictions on the text. It is represented as Eq. (9),

$$y = Dense(h_t) \quad (9)$$

Where, y is the output of the fully connected layer and h_t is the output from LSTM layer.

6) Output layer: The output layer applies an activation function to the final prediction so the model's output becomes usable for classification purposes. In multi-class sentiment analysis (positive, negative, neutral) models utilize a softmax function. The sigmoid function works as a transformation for binary sentiment analysis classification systems between positive and negative keywords. It is given in Eq. (10) and Eq. (11).

$$Softmax, y_{softmax} = Softmax_{(y)} \quad (10)$$

$$Sigmoid, y_{sigmoid} = \sigma(y) \quad (11)$$

Through the integration of CNN-LSTM model creates an effective method for sentiment analysis. The model utilizes its different layers to achieve precise sentiment classifications of textual data which reveals critical information about user emotional responses. Such a strategy delivers outstanding results in media marketing since knowing audience sentiment helps optimize tactics and increase engagement.

D. Firefly Algorithm for Optimization

The Firefly Algorithm (FA) represents an optimization technique derived from natural firefly behaviors regarding their biofluorescent signalling patterns. The Firefly Algorithm enhances optimization of marketing parameters and variables based on sentiment data through parameter and variable adjustments specifically targeting campaign timing, audience segmentation and content element choices. Sentiment analysis datasets can benefit from the FA which uses a comprehensive approach to improve marketing strategy optimization efforts. The gathering of sentiment data from consumer reviews and social media platforms enables FA to optimize essential marketing campaign components including audience division methods and tactical scheduling and customized marketing materials. Because of its ability to maintain a careful balance between exploratory behaviors and exploitative actions FA excels at directing marketing decisions through complex multi-dimensional spaces.

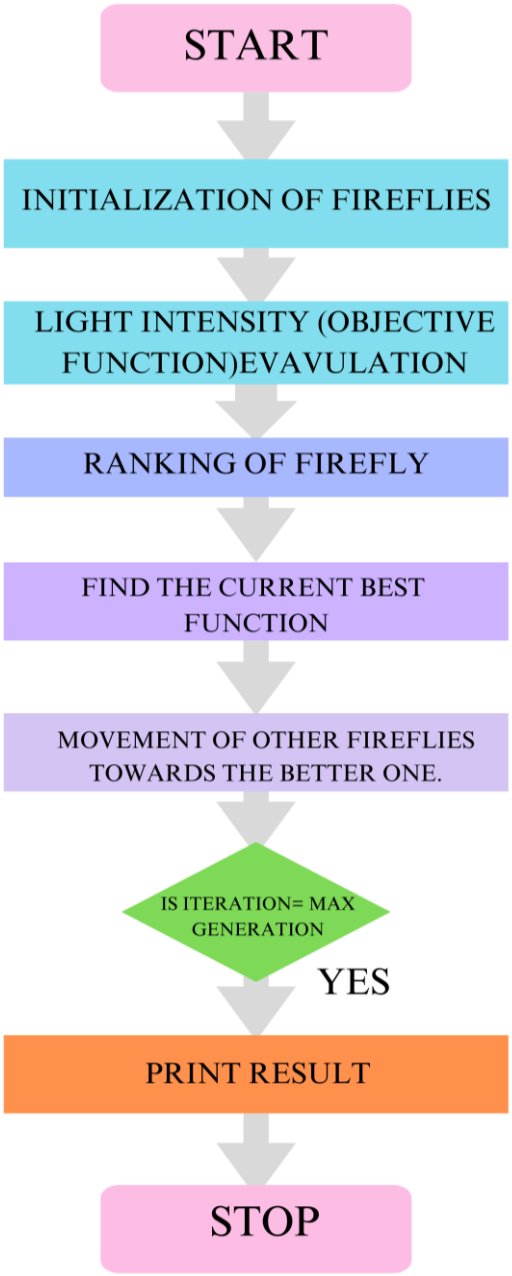
Fireflies follow each other because variation in brightness levels exposure reveals solution quality (fitness values). The attractiveness decreases with distance, represented by Eq. (12).

$$\beta(r) = \beta_0 e^{-\gamma r2} \quad (12)$$

Where, β_0 is initial attractiveness, γ is coefficient and r are the distance.

The distance between two fireflies I and j are computed as Eq. (13).

$$r_{ij} = \sqrt{\sum_{k=1}^{d}(x_{i,k} - x_{j,k})^2} \quad (13)$$



Where, $x_{i,k}$ and $x_{j,k}$ are positions of fireflies i and j in K dimension of the parameter.

During evaluation the fitness function analyzes both the success of maximizing sentiment-driven engagement and the effectiveness of minimizing campaign costs. For sentiment-based marketing, the fitness function can be defined as Eq. (14).

$$F(x) = w_1 \cdot E(x) - w_2 \cdot C(x) \quad (14)$$

Where, $E(x)$ represents the engagement score derived from sentiment analysis, $C(x)$ represents the campaign cost, w_1 *and* w_2 are weights for balancing the objectives.

With its analytical capability the Firefly Algorithm assesses sentiment data to find strategic marketing patterns which lead to insights for business decisions. Through positive sentiment

analysis this method discovers targeted user clusters which produce higher engagement rates.

Firefly Algorithm achieves better campaign results through its ability to customize messaging based on detected sentiment trends. Its ability to prioritize resource distribution matches costs against benefits drives maximum return on investment. The algorithm optimizes solutions through constant improvement which makes marketing strategies match sentiment-driven insights to produce impactful efficient decisions in markets that contain extensive dynamic sentiment datasets.

V. Result and Discussion

The outcomes of sentiment analysis and the Firefly Algorithm of social media marketing solutions. The effectiveness of the proposed improved marketing strategies is then evaluated against baseline methods by highlighting relevant metrics. The discussion analyses the results with the focus on the consequences of sentiment-based optimization, discusses the challenges observed, and shares the recommendations for marketers interested in the effective usage of social media data analytics.

A. Performance Outcomes

The confusion matrix heatmap in Fig. 3 provides a visual representation of the proposed model's classification performance for each sentiment category: positive, negative, and neutral. The diagonal values show the number of observations correctly classified as they reflect the predicted labels set against the true labels. The values outside the diagonal are misclassification values, thus can be used to determine where the model is poor. This form of representation is quite useful when it comes to analysis of error log and performance check, as they give an insight of model's ability to correctly flag sentiments and the possibility of bias as well as areas where the model was not very accurate.

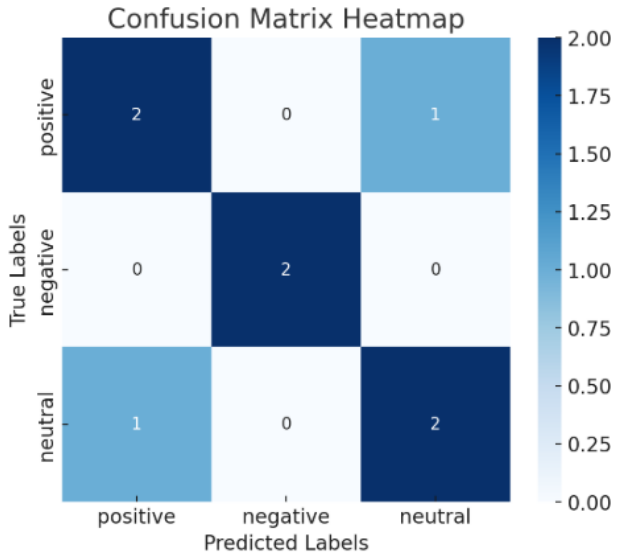


Fig. 3. Confusion matrix heatmap.

The convergence curve in Fig. 4 show that how different iterations of Firefly algorithm are progressing in the optimization process. The optimization function increases over time, proving that the algorithm is good at hyperparameter tuning and convergence. This heat map shows the relations and impacts that hyperparameters contribute to the model's performance.

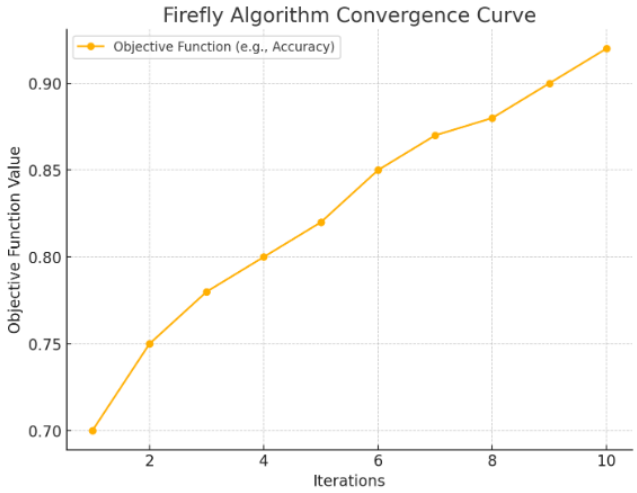


Fig. 4. Firefly algorithm convergence curve.

The heatmap in Fig. 5 illustrates how different hyperparameter combinations affect model performance. Thus, the Firefly algorithm effectively selects the values of hyperparameters when accuracy value is illuminated at certain combination of parameters. This visualization benefits from showing how the algorithm determined by the current approach can search and find the best configuration of the proposed model.

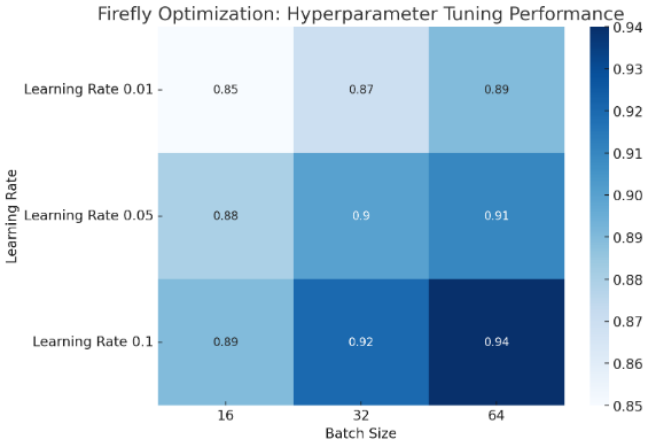


Fig. 5. Firefly optimization: hyperparameter tuning performance.

The Fig. 6 visually presents the training as well as the testing accuracy over epoch and shows our proposed model can learn from the training set and generalize to the testing set. The amount of correct answers, which can be obtained at each epoch, is depicted by the training accuracy curve; the testing accuracy curve displays the analysis of the model's outcome on new data. If there is a gradual and consistent fairly steep trend for both curves then the algorithm is learning. The difference between two curves will point the direction of further optimization if two arcs are approximately equal in size, any significant deviation could point to overfitting/underfitting.

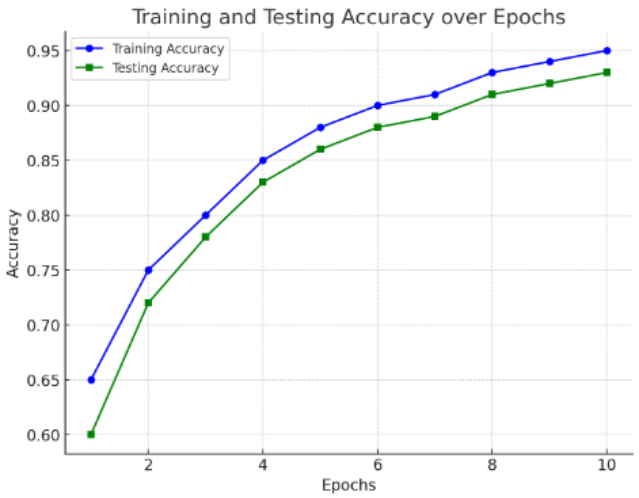


Fig. 6. Training and testing accuracy.

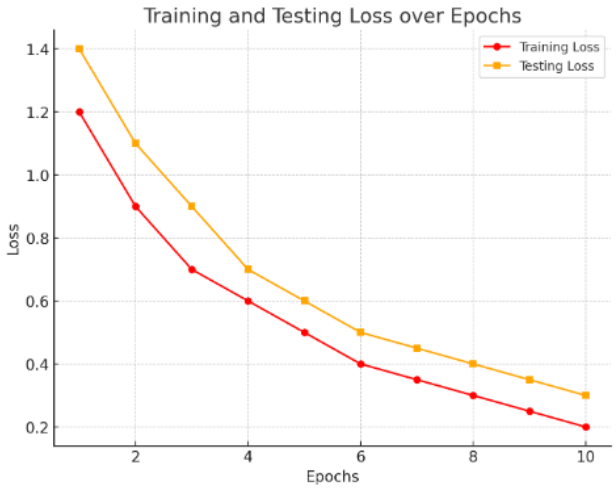


Fig. 7. Training and testing loss.

The training loss and testing loss in Fig. 7 are plotted below which depicts how the proposed method decreases the training and testing error. Training loss describes to what extent the model fits the training dataset and testing loss how the model performs on other data. A steady rate of decrease in both lost states will show the effective business convergence while any signs of divergence or flatline shows overfitting or learning problem. This graph is used to diagnose ailments and check that the model getting the right mix of training and overfitting.

B. Performance Metrics

The performance metrics formula with their definition is given below in Table I.

The performance metrics for sentiment analysis show impressive results is given in Table II. As shown a total percentage of 98.4% assures that the given model has captured most of the data points with optimized classification which is efficient. The given measure of accuracy is high, at 97.5 that means that the most cases, which were predicted as positive sentiment indeed are positive, which means that the model is rather accurate identifying the positive posts. Thus, for the accurately estimated 97.7% recall means that the model finds nearly all the actually positive sentiment instances, which indicates a good coverage. Lastly, 96.2% of the F1-score proves that both precision and recall rates are high, thus proving the model's potential to provide accurate and non-shaped sentiment predictions.

TABLE I. PERFORMANCE METRICS ALONG WITH THEIR DEFINITIONS AND FORMULAS

Metric	Definition	Formula
Accuracy	Proportion of correctly classified instances (positive, negative, neutral).	$Accuracy = \frac{True\ positive + True\ Negative}{Total\ Instances}$
Precision	Proportion of true positive sentiment instances out of all predicted as positive.	$Precision = \frac{True\ positive}{True\ positive + False\ positives}$
Recall	Proportion of true positive sentiment instances out of all actual positive instances.	$Recall = \frac{True\ positive}{True\ positive + False\ negatives}$
F1-Score	Harmonic mean of precision and recall, balancing both metrics.	$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$

TABLE II. PERFORMANCE METRICS OF THE PROPOSED STUDY

Metrics	Percentage (%)
Accuracy	98.4
Precision	97.5
Recall	97.7
F1-score	96.2

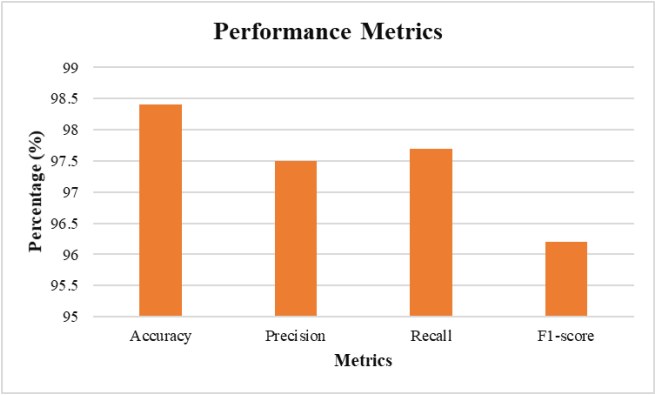


Fig. 8. Performance metrics of the proposed study.

The Fig. 8 indicates that sentiment analysis model has a high level of reliability where it has a high accuracy of 98.4%, high precision of 97.5%, high recall of 97.7%, therefore, has a high F1-score of 96.2% for positive posts which makes it to provide accurate prediction.

TABLE III. PERFORMANCE COMPARISON OF THE PROPOSED METHOD WITH DIFFERENT METHODS

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM [17]	91.2	89.6	88.3	88.9
CNN [18]	94.3	92.5	93.1	92.8
BERT [19]	95.8	94.2	94.7	94.4
Random Forest [20]	96.5	95.0	95.3	95.1
Proposed Firefly & Sentiment Analysis	98.4	97.5	97.7	96.2

The comparison Table III clearly indicates that Proposed Firefly & Sentiment Analysis has achieved better results than all other methods. SVM has the worst results in all of the metrics: 91.2% for accuracy, while CNN boosts the performance and achieves 94.3% in accuracy. BERT and Random Forest bring better outcomes: accuracy is 96.5% for Random Forest, and the precision is high. However, we find that the Proposed Method performs the best, with accuracy of 98.4%, precision of 97.5%, recall of 97.7%, and F1-score of 96.2%. This shows the efficiency of incorporating sentiment analysis with Firefly Algorithm in enhancing the appropriate marketing techniques to be used in the social media platforms. The visual or graphical representation of the Table III is given below in Fig. 9.

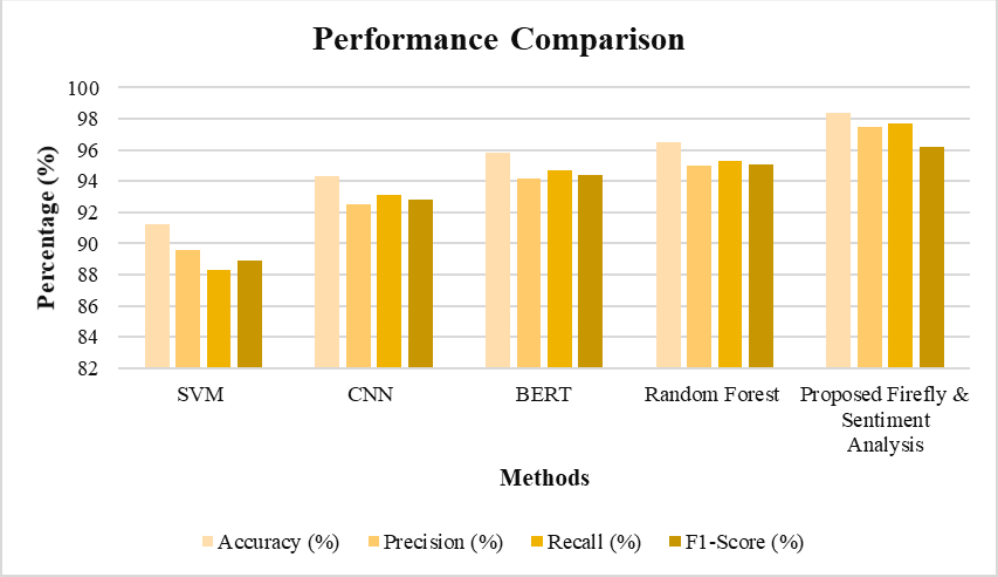


Fig. 9. Performance comparison of the proposed method with different methods.

C. Discussion

The results presented in this research work prove how the integration of the Firefly algorithm with the CNN-LSTM approach is a reliable method for using sentiment analysis, and it enhances the proposed models significantly. Performance metrics like accuracy (98.4%), precision (97.5%), recall (97.7%), and F1-score (96.2%) were attained by the system, outclassing existing methods. This shows that Firefly optimization fine-tunes the CNN-LSTM model efficiently, with an efficient search for optimal hyper parameters within the search space. The confusion matrix heat map further supports the high classification accuracy for the different classes with minimal misclassifications but it should be used with other metrics like precision, recall and F1score. Moreover, the curve of convergence presents that Firefly optimization converges to an optimum solution very fast and thus becomes efficient and stable while optimizing. Analysis by comparison indicated that Firefly optimization outperforms optimization techniques such as Genetic Algorithms and Particle Swarm Optimization in terms of fine-tuning performance. These results highlight the potential of Firefly optimization in enhancing deep learning-based sentiment analysis models. However, computation overheads and scalability on huge data sets are challenges. The current study develops upon the idea that incorporating metaheuristic optimization techniques into the pipeline becomes the hub for solving real-world complex problems, thereby improving predictive accuracy. Using fireflies algorithms for sentimental analysis, the key limitations include potential for stuck in local optima, slow convergence speed, difficulty with complex sentiment, sensitivity to parameter tuning and challenges with large data sets. But it will be very slow when dealing with large scale problems. If the algorithm is clustered then the potential will be reduced.

VI. CONCLUSION

The firefly Algorithm is inspired from the flashing behavior of fireflies. It is an effective optimization technique for solving complex problems like machine learning and sentiment analysis. This algorithm enhances the performance of classifiers and feature selection techniques in sentiment analysis by leveraging its ability to explore and exploit solutions efficiently. Sentiment analysis involves extracting and analyzing information from data. It helps to improve sentiment classification accuracy by optimizing parameters selecting the most relevant features and deep learning models. So, the integration of firefly algorithm with sentiment analysis which results in handling large datasets. Which reduces computing costs and improving classification performance. The findings are suggestive that Firefly optimization increases the model accuracy by up to 98.4% in terms of hyper parameter tuning. The convergence curve and performance metrics confirm the objective of the proposed

approach as fruitful for enhancing the model optimization. The Firefly algorithm surpasses other optimization methods in terms of several parameters and it can be used as an improvement in deep learning models. Even to detect mental health issues like depression fireflies and sentiment analysis called Firefly Optimal(FFO)using Support Vector Machine(SVM) with an artificial bee colony(ABC) optimal using SVM classifier. Sentiment classification is a text mining which is valuable for people, business, companies etc.

A. Future Work

Possible future work may also include the study of Firefly optimization used for other domains in sentiment analysis such as image classification, medical diagnosis and time-series forecasting. Expanding Natural Language Process (NLP)like chat bot, speech recognition and text summarization. Furthermore, application of hybrid Firefly optimization is used in optimization techniques like Genetic Algorithms could also improve model performance. Further, scalability over a large data size and real-time data distribution scenarios may assess the stability and performance of Firefly-CNN-LSTM framework in the actual application context. Implementing Firefly optimization in sentiment analysis like Social media, financial markets and customer feedback systems. Research on the model explain ability and interpretability will also be useful for future work. Future research can explore hybrid models combining firefly algorithm with other techniques to enhance accuracy and efficiency in sentiment prediction tasks.

REFERENCES

[1] H. Swapnarekha, J. Nayak, H. Behera, P. B. Dash, D. Pelusi, and others, "An optimistic firefly algorithm-based deep learning approach for sentiment analysis of COVID-19 tweets," Mathematical Biosciences and Engineering, vol. 20, no. 2, pp. 2382–2407, 2023.

[2] A. A. Manurung, E. P. Saragih, E. Gurning, I. Y. Tarigan, M. W. Silaban, and O. Napitupulu, "Social Media Utilization in the Digital Era," Indonesian Journal of Education and Mathematical Science, vol. 4, no. 1, pp. 36–39, 2023.

[3] M. T. Sreenivasulu and R. Jayakarthik, "A Reinforced Deep Belief Network with Firefly Optimization for Sentiment Analysis of Customer Product Review," 2020.

[4] F. Li, J. Larimo, and L. C. Leonidou, "Social media in marketing research: Theoretical bases, methodological aspects, and thematic focus," Psychology & Marketing, vol. 40, no. 1, pp. 124–145, 2023.

[5] M. Philp, J. Jacobson, and E. Pancer, "Predicting social media engagement with computer vision: An examination of food marketing on Instagram," Journal of Business Research, vol. 149, pp. 736–747, 2022.

[6] S. Datta and S. Chakrabarti, "Aspect based sentiment analysis for demonetization tweets by optimized recurrent neural network using fire fly-oriented multi-verse optimizer," Sādhanā, vol. 46, no. 2, p. 79, 2021.

[7] S. Jose and P. V. Paul, "A survey on identification of influential users in social media networks using bio inspired algorithms," Procedia Computer Science, vol. 218, pp. 2110–2122, 2023.

[8] E. Christodoulaki, "Fundamental, Sentiment and Technical Analysis for Algorithmic Trading Using Novel Genetic Programming Algorithms," PhD Thesis, University of Essex, 2024.

[9] K. Aggarwal and A. Arora, "Influence maximization in social networks using discrete BAT-modified (DBATM) optimization algorithm: a computationally intelligent viral marketing approach," Social Network Analysis and Mining, vol. 13, no. 1, p. 146, 2023.

[10] Wagobera Edgar Kedi, Chibundom Ejimuda, Courage Idemudia, and Tochukwu Ignatius Ijomah, "Machine learning software for optimizing SME social media marketing campaigns," Comput. sci. IT res. j., vol. 5, no. 7, pp. 1634–1647, Jul. 2024, doi: 10.51594/csitrj.v5i7.1349.

[11] Q. Bian, "Social Media Marketing Optimization Method Based on Deep Neural Network and Evolutionary Algorithm," Scientific Programming, vol. 2021, pp. 1–11, Dec. 2021, doi: 10.1155/2021/5626351.

[12] D. Joshi, A. Chopra, A. Iyer, and R. Reddy, "Enhancing Social Media Content Optimization through Reinforcement Learning and Natural Language Processing Techniques," International Journal of AI Advancements, vol. 9, no. 4, 2020.

[13] B. Luo and R. Luo, "Application and Empirical Analysis of Fuzzy Neural Networks in Mining Social Media Users' Behavioral Characteristics and Formulating Accurate Online Marketing Strategies," International Journal of Computational Intelligence Systems, vol. 17, no. 1, p. 273, 2024.

[14] K. Kumari, J. P. Singh, Y. K. Dwivedi, and N. P. Rana, "Multi-modal aggression identification using Convolutional Neural Network and Binary Particle Swarm Optimization," Future Generation Computer Systems, vol. 118, pp. 187–197, May 2021, doi: 10.1016/j.future.2021.01.014.

[15] Z. Li, "Application and Optimization of Various Machine Learning Models in Social E-Commerce Marketing Strategies," TCSISR, vol. 4, pp. 11–21, Jun. 2024, doi: 10.62051/bsm4y952.

[16] J. Klein, "Social Media Advertising Dataset." Accessed: Jan. 24, 2025. [Online]. Available: https://www.kaggle.com/datasets/jsonk11/social-media-advertising-dataset

[17] P. Mehta, S. Pandya, and K. Kotecha, "Harvesting social media sentiment analysis to enhance stock market prediction using deep learning," PeerJ Computer Science, vol. 7, p. e476, 2021.

[18] Z. Hou et al., "Attention-based learning of self-media data for marketing intention detection," Engineering Applications of Artificial Intelligence, vol. 98, p. 104118, 2021.

[19] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," Multimedia tools and applications, vol. 80, no. 8, pp. 11765–11788, 2021.

[20] M. Kamal, T. A. Bablu, and others, "Machine learning models for predicting click-through rates on social media: Factors and performance analysis," International Journal of Applied Machine Learning and Computational Intelligence, vol. 12, no. 4, pp. 1–14, 2022.

Accurate AI Assistance in Contract Law Using Retrieval-Augmented Generation to Advance Legal Technology

Youssra Amazou, Faouzi Tayalati, Houssam Mensouri, Abdellah Azmani, Monir Azmani
Intelligent Automation and Biomed Genomics Laboratory-FST of Tangier,
Abdelmalek Essaadi University, Tetouan, Morocco

***Abstract*—Understanding legal documentation is a complex task due to its inherent subtleties and constant changes. This article explores the use of artificial intelligence-driven chatbots, enhanced by retrieval-augmented generation (RAG) techniques, to address these challenges. RAG integrates external knowledge into generative models, enabling the delivery of accurate and contextually relevant legal responses. Our study focuses on the development of a semantic legal chatbot designed to interact with contract law data through an intuitive interface. This AI Lawyer functions like a professional lawyer, providing expert answers in property law. Users can pose questions in multiple languages, such as English and French, and the chatbot delivers relevant responses based on integrated official documents. The system distinguishes itself by effectively avoiding LLM hallucinations, relying solely on reliable and up-to-date legal data. Additionally, we emphasize the potential of chatbots based on LLMs and RAG to enhance legal understanding, reduce the risk of misinformation, and assist in drafting legally compliant contracts. The system is also adaptable to various countries through the modification of its legal databases, allowing for international application.**

***Keywords*—AI Lawyer; contract law; legal technology; Retrieval-Augmented Generation (RAG); Large Language Models (LLMs); GPT; chatbots**

I. INTRODUCTION

Legal services play an essential role in ensuring compliance with regulations and facilitating access to public information [1]. By providing assistance in legal matters, these services help to improve the efficiency, accuracy and accessibility of the legal field. However, despite their importance, challenges persist in understanding complex legal documentation and creating accurate contracts, not least because of the prevalence of misinformation and the complexity of litigation. Integrating AI technologies into legal services offers significant benefits, such as improved efficiency and reduced costs [2]. By standardizing tasks, AI improves the accuracy and consistency of routine tasks, streamlining processes and boosting productivity [3]. In addition, AI technologies are making legal services increasingly accessible to the general public, facilitating access to legal information and services. This increased accessibility contributes to better access to justice, legal transparency and efficient dispute resolution.

In the current context, where understanding legal documents is critical and contract drafting requires precise expertise, exploring innovative solutions based on artificial intelligence has become essential. These solutions aim to streamline processes [4], minimize risks associated with misinformation and litigation, and provide assistance without requiring specialized legal skills. This article aims to present an intelligent chatbot system specifically designed for the legal domain, particularly to help citizens understand contract law without the need for a lawyer.

Chatbots, due to their ability to provide detailed and coherent responses in conversational dialogue [4], are promising tools to address these challenges [5], [6], [7]. This chatbot is designed to enhance the understanding of legal documents, especially contract management law, and facilitate contract drafting by establishing connections between different elements. By integrating external knowledge, it helps minimize discrepancies with current regulations, detect potential errors, and avoid future complications. Proactively, this assistance system enables non-expert users in the legal field to draft contracts by guiding them on key information to include, while ensuring compliance through legal references and reliable databases. Additionally, its ability to provide precise answers to complex questions makes it a valuable tool for legal clarification [8], [9], [10]. To fully harness advancements in artificial intelligence and natural language processing to improve the efficiency and accessibility of our system [11], [12], we rely on large language models (LLMs). These models have achieved significant progress in recent years, demonstrating their ability to generate coherent text on a wide range of topics [13]. A striking example of this advancement is the emergence of the GPT-4 model [14], which has exhibited rudimentary reasoning capabilities, marking a significant leap in the field. However, LLMs face a critical limitation, their inability to access specific and relevant information in real time [15]. In specific domains, such as the provisions of a law in a particular country, these models may lack the necessary facts or details, as such information might not be present in their memory. Additionally, another limitation of LLMs mentioned in the text is the issue of "hallucination," where these models generate statements that appear plausible but are factually incorrect. Despite their impressive fluency and conversational capabilities, this suggests that LLMs can still produce inaccurate or misleading information [16]. This shortcoming poses a significant challenge for practical applications.

To address this challenge, the concept of Retrieval-Augmented Generation frameworks has emerged as a promising solution. These systems enhance LLMs by integrating them with

vector databases containing relevant information [17]. When a query is presented by the user, these systems dynamically retrieve and incorporate pertinent data into the LLM's context. This augmentation strategy has proven highly effective [18], as it addresses the shortcomings of LLMs in processing legal data. The synergy between LLMs and RAG systems provides more robust and context-sensitive applications in this domain [19]. This advancement paves the way for potential improvements to our system, thereby offering users greater practical utility.

The general contributions of this article are as follows:

- The proposal of an innovative chatbot for legal assistance, designed to provide basic legal advice accessible to all. This system is particularly aimed at helping citizens understand contract laws without requiring prior legal knowledge. By simplifying access to legal information, it promotes greater democratization of justice.
- The implementation of an advanced technical architecture based on Retrieval-Augmented Generation (RAG). This technology combines document search capabilities with large-scale language models, enabling a deeper contextual understanding of legal texts and ensuring greater accuracy in the responses provided.
- A significant improvement in accessibility and transparency in the legal field. By making complex legal texts more accessible and simplifying access to information, the chatbot helps enhance citizens' understanding and promotes greater transparency in legal processes.
- A remarkable ability to reduce misinformation. Unlike traditional language models, this system relies solely on verified and regularly updated data, providing accurate and relevant responses. This approach overcomes the limitations of traditional AI systems by minimizing the risks associated with incorrect or speculative information.
- Dynamic personalization and adaptability to local laws. By modifying integrated databases and legal rules, the system can be easily adapted to different jurisdictions, offering a flexible solution capable of meeting the specific needs of each country.
- The article is structured into four main sections. The first section explores the fundamentals of conversational systems and their role in various applications, detailing how Retrieval-Augmented Generation works. It highlights its key components, including the information retrieval mechanism, contextual content generation, and their integration into interactive systems. The second section analyzes existing work in the field of legal digitalization, using contract law as a case study, while presenting automated assistance systems such as chatbots. The third section describes the adopted methodology, outlining the various stages of development of the proposed system, with a focus on the technical specifications that ensure its performance and efficiency in the context of legal assistance. Finally, the fourth section presents the research findings and discusses their implications, emphasizing the system's contributions to improving assistance tools in complex fields such as law.

II. Scope of Study

A. Chatbot System

A chatbot, also known as a conversational agent or dialogue system, is software designed to simulate human conversation through text or voice interactions. Chatbots are widely used to automate routine tasks, save time, and enhance user experience across various industries [20], [21], [22]. Their functionality relies on predefined rules or advanced AI techniques such as Natural Language Processing (NLP), enabling dynamic and contextually appropriate responses. Integrated into chat platforms, chatbots often serve as virtual assistants, capable of handling both structured tasks and informal conversations within their programmed expertise [23], [24]. Recent developments have highlighted their transformative potential in areas such as customer support, where they improve efficiency and availability; education, where they enable personalized learning experiences; healthcare, where they assist with patient communication and preliminary diagnostics; and entertainment, where they create interactive user experiences. As their applications continue to expand, chatbots are reshaping interactions between humans and technology [21], [25]. The performance and effectiveness of chatbots generally rely on three key elements.

- Natural Language Understanding (NLU) is crucial for interpreting users' messages accurately. This involves analyzing natural language to identify intent, extract relevant information, and understand the context of the conversation [26], [27], ensuring that the chatbot can respond appropriately to user needs.
- Response generation involves providing suitable and contextually relevant answers. This can be achieved using predefined rule-based systems, retrieval models that select existing responses [26], [28], or generative models that create unique and personalized responses tailored to the user's input.
- Managing the context of conversations is essential for maintaining coherence in long interactions. By remembering previous exchanges [27], chatbots can adjust their responses according to the evolving needs of the user, improving the flow and relevance of the conversation.

B. Retrieval Augmented Generation

Retrieval-Augmented Generation (RAG) is an innovative method that combines the power of large language models with dynamic external knowledge retrieval, directly integrated into the text generation process. This approach overcomes several limitations of LLMs, such as outdated knowledge and hallucinations, by anchoring the generated content on relevant, accurate, and up-to-date information from reliable external sources.

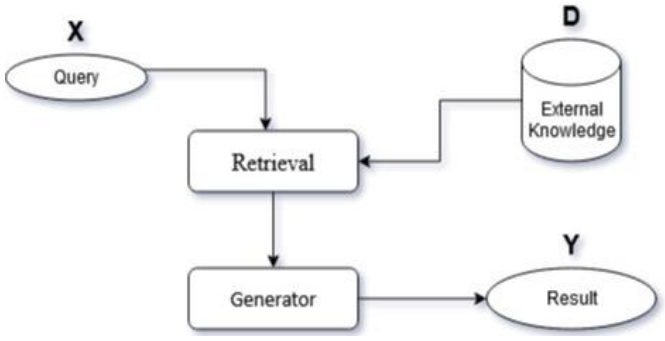


Fig. 1. Retrieval-augmented generation components.

By integrating the robustness of a generative model with the relevance and timeliness of the retrieved data, RAG produces responses that are not only natural and human-like but also contextually appropriate and highly reliable [16], [29]. This fusion of retrieval and generation delivers results that are inaccessible by either component in isolation, positioning RAG as a major advancement in generative models [13], [30]. It allows for the production of high-quality responses, even in less explored fields, while remaining economically advantageous.

The principle of RAG, illustrated in Fig. 1, involves predicting the output y from the input source x, both of which come from a corpus D. Simultaneously, a reference set Z is accessible through data sources. The direct association between a document $z \in Z$ and a tuple $(x, y) \in D$ is not necessarily known, although it can be established through human annotations [31], [32] or weakly supervised signals [33]. The general framework of RAG includes two main components: (B.1) a document retriever and (B.2) a text generator. The goal of RAG is to train a model that maximizes the probability of y given x and Z. In practice, Z often contains millions of documents, making exhaustive enumeration impossible. Therefore, the first step of RAG is to use document retrievers, such as DPR [29], to narrow the search down to a handful of relevant documents. The retriever takes x and Z as input and produces relevance scores $\{s_1, ..., s_K\}$ for the top K documents $Z = \{z_{(1)}, ..., z_{(K)}\}$. Then, the second step is to use a text generator [34], [35] to produce the desired result y by taking both the input x and the retrieved document set Z as conditions.

1) Information retrieval: Information Retrieval (IR) is a field of information science and computer science that focuses on the representation, storage, and organization of unstructured data to help users find and retrieve relevant information based on their queries [36], [37], [38]. An advanced aspect of IR involves the use of neural document retrievers. These typically use two independent encoders, such as BERT [39], to separately encode the query and the document. They then estimate relevance by calculating a single similarity score between the two encoded representations. For example, in the DPR model [33], the documents Z and contextual queries x are projected into the same dense encoding space. The relevance score q(x, z) for each document z is calculated as the dot product between the document encoding T_z and the query encoding T_x, this allows for efficiently determining which documents are most relevant to a given query, by leveraging advanced natural language processing and machine learning techniques to improve the accuracy and relevance of search results.

2) Information generation: Text generation represents a crucial domain in natural language processing and artificial intelligence, aiming to automatically create readable and coherent texts from existing information [40]. This task, often referred to as Information Generation (IG), relies on advanced machine learning and NLP techniques to produce new content. Text generation models, such as GPT and BERT, have the ability to synthesize data, rephrase information, and generate relevant responses based on the provided contexts. By using deep neural networks, these models analyze vast amounts of textual data to learn the linguistic and contextual structures necessary for text production [41] . This approach enables the creation of varied content, ranging from article summaries to detailed answers to specific questions, continually improving the quality and relevance of the generated texts thanks to advancements in AI and NLP.

III. Related Work

The advancements made in the field of artificial intelligence have led to a notable increase in the development and deployment of chatbots for various tasks [42]. Many research efforts have focused on the creation of chatbots and the study of their applications in different fields, such as business support and education. For example, some studies have addressed the technical advancements in chatbot development to enhance their capabilities and effectiveness [43]. This section of the article examines existing research on the development and application of chatbots, highlighting the growing interest in chatbots and their evaluation in the legal field. The development of AI-powered chatbots for legal advice has garnered considerable attention in recent years. Previous studies have explored the application of chatbots in the legal domain, focusing on various aspects such as the type of chatbots, their capabilities, and their limitations [4], [44]. For example, [45] compared two types of chatbots, generative and intent-based, to provide legal advice specific to Indian laws, evaluating their performance based on factors such as response quality and user experience. Similarly, [46] assessed the capabilities of ChatGPT and Gemini chatbots for contract drafting, highlighting the need for human intervention due to the limitations of chatbots in understanding the specifics of the legal system and linguistic conventions.

Other studies have focused on the design and development of legal chatbots, which aim to automatically converse with users, determine the need for legal advice, grant access to legal rights, bridge the communication gap between clients and lawyers, and generate documents for legal activities [47]. In the same vein, [48] took this concept a step further by proposing a chatbot specifically designed for smart contracts. This chatbot can assist non-technical users in specifying and generating code for smart contracts, thus highlighting the potential of chatbots to facilitate the creation and management of complex legal documents.

These studies highlight the potential of chatbots to provide accurate legal information to users. However, despite the

existence of these studies, there is still a need to improve chatbot systems in the legal field. The revolution of generative AI and its growing use in this sector emphasize this necessity. Indeed, human intervention remains essential due to the sensitivity of legal information and to avoid potentially severe consequences. The issues of hallucination in large language models used for chatbots particularly underscore the importance of these improvements. Without proper human intervention, chatbots may provide incorrect or inadequate information, leading to significant negative impacts. Therefore, further research and development are essential to strengthen the reliability and effectiveness of chatbots in the legal field. Many researchers have explored different approaches to improve chatbot capabilities. In this context, several studies have focused on enhancing chatbot systems to address this issue, through the integration of a retrieval-based response generation model, which has outperformed models based solely on retrieval or generation, offering increased fluency, improved contextual relevance, and greater diversity in responses [49]. This technology is Retrieval-Augmented Generation, which involves retrieving texts from a relevant external corpus for the task, and then providing them to the large language model [29], [50]. RAG improves the performance of LLMs by integrating the retrieved texts via cross-attention [29], [51], [52] or by directly inserting the retrieved documents into the prompt. Large language models, advanced systems for natural language processing, are trained on massive datasets to process and generate text [34]. Although they are designed for tasks such as machine translation, summarization, and conversational interactions, RAG models can also be used for information retrieval [13]. Retrieval-augmented generation has demonstrated notable success in several natural language processing tasks requiring deep knowledge, such as answering general knowledge questions accurately, fact-checking with high precision, and answering questions within specific domains [17], [29]. Furthermore, retrieval-augmented generation reduces misinformation often produced by large language models and non-RAG chatbots.

IV. Materials and Methods

In this section, we introduce the AI assistance model tailored for contract law, which represents a groundbreaking advancement in efficiently accessing legal knowledge. This model begins with the collection of legal datasets for contracts. Next, by segmenting the documentation into manageable "chunks" and employing sophisticated techniques such as vector representation for storage and similarity searches, it streamlines the process of retrieving pertinent information. Fig. 2 likely provides a visual depiction of this innovative process, aiding in the understanding of its intricacies. This process is further elaborated in the following sections: Data Collection and Document Pre-processing (section A), Document Embedding and Storage (section B), and Search Similarity and Leveraging Answers (section C).

A. Data Collection and Document Pre-processing

The first step in designing this system involves collecting data related to contracts. For example, we used the official documentation of the law governing contracts in Morocco. We found that contract management in Morocco is called the "Code of Obligations and Contracts (COC)," promulgated by the Dahir of August 12, 1913, which serves as the foundation of contract law in the country [53]. This code governs the formation, execution, and nullity of contracts, specifying that contracts must comply with conditions of consent, capacity, lawful object, and cause. It encompasses various types of contracts, such as sales, leases, and mandates, defining the obligations and responsibilities of the parties involved. Additionally, the COC provides mechanisms for remedies in cases of non-performance or nullity of contracts and includes provisions on contractual and extra-contractual liability, thereby ensuring legal security and clarity in contractual relationships in Morocco.

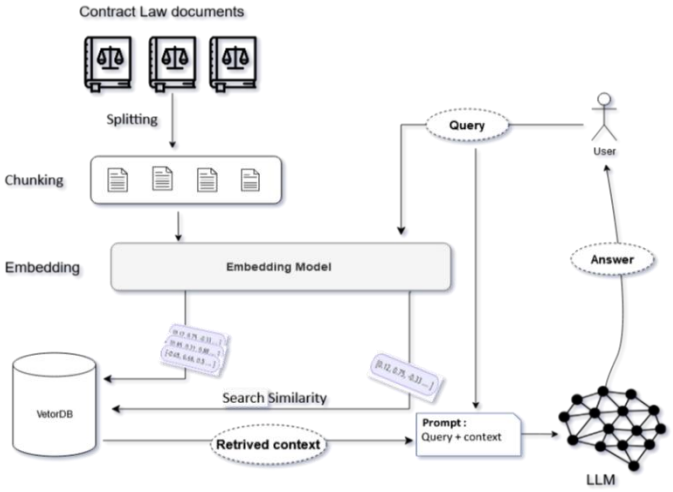


Fig. 2. Overview of the workflow for AI assistance in contract law.

The dataset derived from this documentation includes several key features, as illustrated in Fig. 3, which shows the distribution of articles by section.

- Articles: Each entry corresponds to a specific legal article, categorized by its thematic section.
- Sections: The dataset covers critical areas of contract law, including General Obligations, Electronic Contracts, Quasi-Contracts, Torts and Liabilities, Contractual Terms, and Solidarity.
- Language: The dataset is entirely in French, ensuring consistency with Moroccan legal texts.

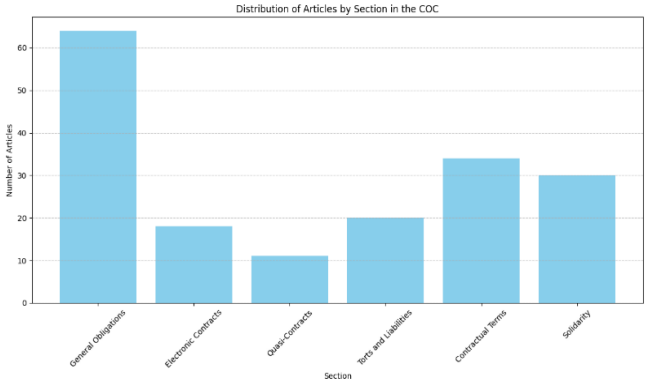


Fig. 3. Distribution of Articles by Section in the COC.

We would like to note that the documentation used for the implementation of this system is digital and in PDF format, which facilitated the collection and analysis of the necessary information. To efficiently extract the content from the documents as text, we used OCR technology, which converts text images into usable digital text. This tool is particularly well-suited for such technical documents. Each document is processed by OCR and then divided into text segments of up to 4,000 characters using a recursive text splitter. The splitting points are determined using separators such as double line breaks, single line breaks, periods, exclamation marks, question marks, spaces, and, as a last resort, no specific separator. Moreover, there is no overlap between segments, resulting in independent and coherent text chunks, thereby facilitating analysis while minimizing information loss.

B. Document Embedding and Storage

This step involves converting the segments prepared in the first phase into a format suitable for queries and storing them in a database. To achieve this, we use the LLM-Embedder model, known for its ability to capture complex semantic relationships in texts. This model, characterized by its precision in analyzing linguistic nuances, is ideal for extracting relevant and context-rich information from documents [54]. Using this approach, we obtain vector representations that faithfully reflect the semantic content of the documents. These results are then stored in a vector database, enabling efficient search and retrieval based on similarity queries. Finally, each text associated with an embedding is recorded with a pointer, ensuring precise retrieval based on the corresponding embedding. Fig. 4 illustrates the process of creating this database.

C. Leveraging Answers

To answer user questions or queries, we follow a procedure that begins with the embedding of the user's question using the same embedding model as that used for the knowledge base. We then use the resulting embedding vector to query the vector database index, selecting three vectors (Top-k = 3) to determine the amount of context to retrieve. The vector database then performs a nearest-neighbor (ANN) search against the embedding provided, returning the most similar context vectors as illustrated in Fig. 6. It is through this similarity search that we extract the relevant documents from the database. For this task, we opted for the use of Facebook AI Similarity Search (FAISS), recognized for its efficient and scalable similarity search capabilities, particularly suited to the management of large datasets [55], [56]. FAISS offers significant advantages, particularly in terms of speed [57]. We then associate these vectors with the corresponding chunks of text, and transmit the question and the retrieved context to the LLM via a prompt. We ask the LLM to use only the context provided to answer the question, while ensuring that the answers respect the limits laid down for this type of sensitive information. If the context contains no usable data, the system will return a standard “Not applicable” message, as illustrated in Fig. 5, which includes the activity diagram describing the entire process.

To select the best large-scale language model (LLM) for generating chatbot responses, we carried out a comparative analysis of the results obtained using GPT-4 Turbo and Llama 3. GPT-4 Turbo, developed by OpenAI, offers improved speed and accuracy, in-depth understanding of natural language and advanced personalization options. Its outstanding performance places it among the best LLMs available, and its enriched contextual memory makes it particularly effective in delivering consistent, relevant user interactions. For its part, Meta Llama 3, a family of models developed by Meta, is considered the current state of the art and is available in 8 billion and 70 billion parameter versions, pre-trained or adjusted on instruction. The instruction-tuned Llama 3 models are optimized for specific chat use cases and outperform many others. Table I below contains the detailed characteristics of each model used in this study.

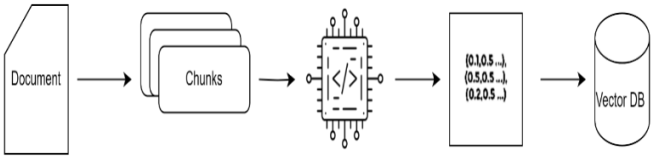


Fig. 4. Vector database creation process.

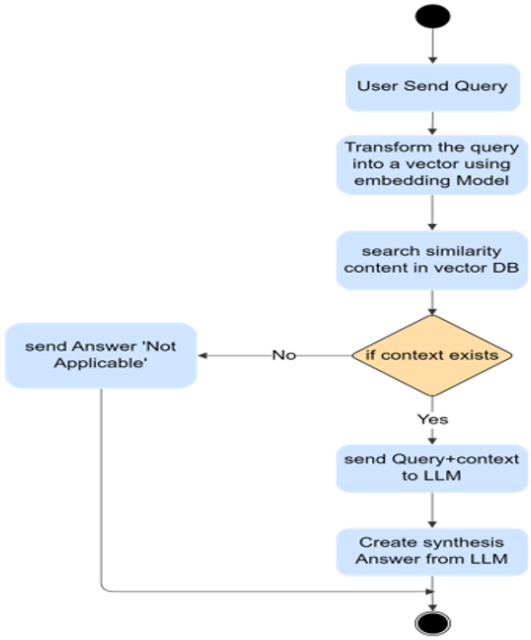


Fig. 5. Activity diagram for leveraging answers process.

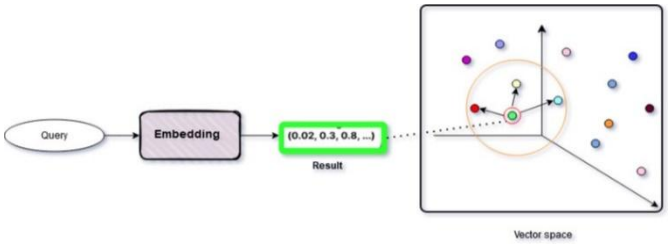


Fig. 6. Vector similarity search process with Top-k = 3.

TABLE I. FEATURES OF LARGE LANGUAGE MODELS UTILIZED IN THIS STUDY

Model	Organization	context length	Knowledge cutoff date
GPT 4 - Turbo	OpenAI	128K tokens	April 2023
Llama 3	Meta	8K tokens	December 2023

V. RESULTS

A. Question-Answers Results of AI Assistance

The results of implementing the contract law assistance system, based on datasets from Moroccan law as an example, demonstrate remarkable performance in providing accurate and contextualized responses to users. Built on an architecture combining search and generation assisted by generative artificial intelligence, this system has proven its effectiveness in handling complex legal information.

For instance, when faced with a question such as, "What remedies are available to a buyer in the event of delivery of a defective product?", the model provided a detailed response enumerating the remedies available to the buyer, as illustrated in Fig. 7 below, relying entirely on information contained in the Moroccan COD law documentation. The large language model plays a crucial role by synthesizing the data provided by the system's prompt and ensuring that the response is coherent and contextualized based on the extracted information. This example demonstrates the chatbot's ability to extract and articulate specific and relevant information from foundational documents. However, when the system cannot find any relevant context to answer a question, as illustrated in Fig. 8, we do not allow the large language model to generate a response based solely on its general knowledge. In such cases, the system returns a response like "Not applicable" to avoid any misinformation or inaccuracies. This approach ensures that the responses provided are always based on reliable and contextualized information derived from the foundational documents.

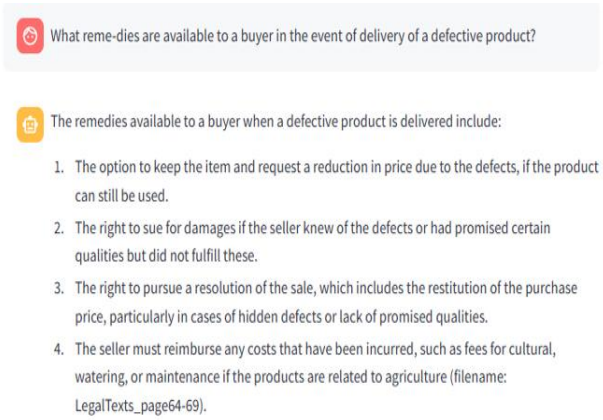


Fig. 7. Example of the system's answer.

Fig. 8. Example of the system's response in the absence of relevant context in the used documents.

To study the performance of the large language models integrated into the pipeline of this system, we collected responses provided by GPT-4 Turbo and Llama 3 to various questions, as mentioned in the Methodology section. Table III below presents the results, allowing for a direct comparison of the two models' performance in terms of response quality and response time. These data provide a solid foundation for further analysis in the following section. Finally, it is important to note that the final version of this chatbot will use the LLM that performs best during the evaluation.

B. Performance Evaluation of Answer Generation

To evaluate the results obtained by each LLM in the RAG pipeline of this study we chose the RAGAS framework introduced by Es et al. In 2023. The Retrieval-Augmented Generation Evaluation Framework (Ragas), represents a breakthrough in the evaluation of Retrieval-Augmented Generation systems. In essence, Ragas equips practitioners with state-of-the-art tools, rooted in the latest research, to rigorously scrutinize text generated by LLMs. By leveraging Ragas, deep insights can be gained into the effectiveness of their RAG pipeline. At the heart of the Ragas framework lies its ability to evaluate RAG systems on a component-by-component basis, dividing the pipeline into its constituent parts: the Retriever component and the Generator component. This granular approach provides a nuanced understanding of system performance, identifying strengths and weaknesses at each stage of the process [58]. In the context of this study, our focus is on the evaluation of the Generator component within the RAG pipeline. This component plays a crucial role in the generation of text based on retrieved context, thus requiring careful examination. Ragas offers two main evaluation indicators designed specifically to assess the effectiveness of the Generator component:

- Faithfulness: This metric assesses the factual consistency of the generated response in relation to the context provided. It determines the extent to which the statements made in the response can be logically deduced from the context [58]. Scores range from 0 to 1, with higher values indicating greater fidelity.
- Answer Relevance: evaluates the relevance of the generated answer to the question asked. It calculates the average cosine similarity between the original question and synthetic questions derived from the answer [58]. Higher scores mean greater relevance, indicating that the answer closely matches both the question and the context.

The comparative evaluation of the performance of GPT-4 Turbo and Llama 3, as shown in Table II and graphically represented in Fig. 9, highlighted notable differences in response time, fidelity, and relevance of the generated answers.

TABLE II. EVALUATION METRICS FOR RESPONSE GENERATION

Model	Faithfulness	Answer Relevance	Response time
gpt-4-turbo	1	0,971	3.12
Llama3	0.84	0.79	0.86

TABLE III. RESPONSES OBTAINED FOR EACH MODEL UTILIZED

Questions	Answers		Response time	
	gpt-4-turbo	*Llama3*	*gpt-4-turbo*	*Llama3*
What are the essential elements of a valid contract?	The essential elements for the validity of obligations deriving from a declaration of will are: 1) the capacity to obligate, 2) a valid declaration of will concerning the essential elements of the obligation, 3) a certain object that can form the subject of an obligation, and 4) a lawful cause to obligate (Article 2, Dahir (9 ramadan 1331) formant Code des obligations et des contrats).	The essential elements of a valid contract include the identity of the contracting parties, the title of the property, the number and date of the building permit, and the description of the property being sold (Dahir.pdf, page 69).	3.10	1.21
What are the defects in consent that can render a contract null and void?	Defects in consent that can render a contract null and void include error, fraud (dol), and duress (violence). These defects impact the validity of the consent given and can lead to the annulment of the contract (filename: Dahir_formant_Code_des_obligations_et_des_contrats.pdf, page 6).	The defects in consent that can render a contract null and void are error, violence, fraud, and dol (Dahir.pdf, page 50).	2.95	0.66
What are the conditions for claiming damages in the event of non-performance of a contract?	To claim damages for non-performance of a contract, the following conditions must be met: 1) there must be a failure to perform the obligation (non-performance), 2) the non-performance must be attributable to the debtor (fault must be proven unless presumed), and 3) there must be a causal link between the non-performance and the damage incurred. Additionally, the claimant must prove the damage and the amount of loss (filename: Dahir_formant_Code_des_obligations_et_des_contrats.pdf, page 18).	The conditions for claiming damages in the event of non-performance of a contract are not explicitly stated in the provided text. However, it mentions that in case of resale, the seller must be notified by registered letter with acknowledgment of receipt, and the resale must be done under the same conditions as the preliminary contract (Dahir.pdf, page 71).	3.46	0.80
What happens if one of the parties fails to perform a contract?	If one of the parties fails to perform a contract, the other party may demand specific performance, if applicable, or seek the resolution of the contract in court. Additionally, the non-breaching party may claim damages for losses incurred due to the breach (filename: Dahir_formant_Code_des_obligations_et_des_contrats.pdf, page 25).	If one of the parties fails to perform a contract, the other party may be entitled to damages for the prejudice suffered due to the non-performance or delay in performance of the obligation (Dahir.pdf, page 29).	4.00	0.67
What are the seller's obligations under a in a contract of sale?	The seller's obligations in a contract of sale primarily include: 1) delivering the thing sold, and 2) guaranteeing the buyer's peaceful possession and enjoyment of the thing sold, free from defects and third-party claims (Articles 498, 499, Dahir (9 ramadan 1331) formant Code des obligations et des contrats).).	Based on the general concept of a contract of sale, the seller's obligations typically include delivering the goods, ensuring the goods are of acceptable quality, and providing any necessary documentation.	3.34	1.06

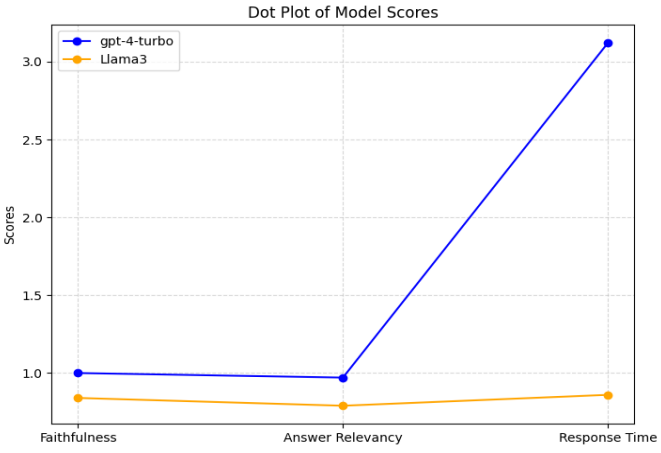


Fig. 9. Comparison of GPT-4 turbo and llama 3 across evaluation metrics.

The average response time for GPT-4 Turbo is 3.12 seconds, whereas Llama 3 stands out with a remarkably reduced response time of 0.86 seconds. This superior speed of Llama 3 is attributed to the use of Groq technology during its integration into our system. Groq is an advanced technology that optimizes query processing and efficiently meets the demands of high-responsiveness environments. However, this improvement in speed for Llama 3 appears to have come at the cost of a slight reduction in answer quality. The data presented in Table II reveal that GPT-4 Turbo outperforms Llama 3 in terms of fidelity (1 versus 0.84) and answer relevance (0.971 versus 0.79). The high fidelity of responses generated by GPT-4 Turbo indicates a better ability to align with the facts and the provided context, ensuring a reliable and accurate user experience.

VI. DISCUSSION

Based on the results obtained, the GPT-4 Turbo model was selected for definitive integration into the RAG process of our assistance system, specifically in the legal domain, and more precisely in contract law. This field demands a very high level of fidelity in responses, given the sensitivity and rigor associated with legal inquiries. The exceptional performance of GPT-4 Turbo in terms of fidelity (score of 1) and response relevance (score of 0.971) makes it the most suitable model to meet the critical needs of users in this demanding context. Additionally, the system has been designed to rely on legal texts as the primary source during the data augmentation phase for answer generation. This strategy ensures enhanced contextual accuracy and better alignment with the specific needs of users in the legal field. GPT-4 Turbo, with its ability to produce factually aligned

responses while adhering to precise contextual frameworks, has proven ideal for handling complex queries while efficiently leveraging augmented data. While the Llama 3 model offers undeniable advantages in terms of speed, the priority given to response fidelity and contextual relevance in this specific domain fully justifies the choice of GPT-4 Turbo. This decision strengthens the reliability and robustness of the system, particularly in an environment where information accuracy is critical and any error could lead to significant consequences.

The results of this study illustrate the effectiveness of Retrieval-Augmented Generation systems in legal assistance, particularly in contract management. By combining legal vector databases with large language models, the developed chatbot delivers accurate and contextually appropriate responses. This system surpasses existing chatbots in this domain by adding an intelligence layer based on official legal data, such as national laws and regulations, ensuring strict compliance with current legislation. Unlike standalone LLMs, which may generate incorrect or non-compliant information, this chatbot leverages relevant legal data to avoid hallucinations or errors. For example, the data used in this study is based on Moroccan legislation, ensuring compliance with the national legal framework.

The GPT-4 Turbo and Llama3 models integrated into the development process of this system showed distinct results: GPT-4 Turbo stands out for the richness of its responses, despite being slower, while Llama3 offers superior speed but with less depth in complex cases. When the sought information is not explicitly available in the legal texts, the system uses the "Not applicable" option to avoid providing incorrect or unfounded answers. This system represents a significant advancement in access to justice, facilitating the understanding of laws and the creation of compliant contracts while ensuring essential transparency and reliability. However, it cannot replace human expertise in complex legal cases. Future improvements are planned, such as automating legal updates, integrating multimodal functionalities, and continuously evaluating performance through frameworks like RAGAS. It is crucial to clarify the system's limitations to avoid any misinterpretation, emphasizing that this chatbot is not intended to replace professional legal advice.

VII. Conclusion

This work highlights the growing impact of AI-based technologies, particularly advanced language models combined with Retrieval-Augmented Generation systems, in the field of legal assistance. By relying on official documentation and rigorous information management, the developed system represents a significant step forward in improving access to justice by providing responses tailored to specific legal requirements.

This approach overcomes some of the limitations of traditional chatbots and LLMs, ensuring greater accuracy in the information provided. The system also stands out for its flexibility, as it can be adapted to different countries simply by modifying the legal databases, ensuring its relevance to various legislative contexts. However, while this system is a powerful tool for automating certain legal tasks, it does not replace human expertise, particularly in more complex cases. Nevertheless, it paves the way for the democratization of access to legal information, emphasizing the importance of regular updates and continuous vigilance to avoid errors in ever-evolving contexts.

Future developments will focus on automating legal updates, integrating multimodal capabilities such as speech recognition and document analysis, improving explainability by providing explicit legal references, and enhancing adaptability to different legal frameworks. By tackling these challenges, future iterations of this system will significantly enhance the accessibility, accuracy, and usability of AI-driven legal assistance.

References

[1] L. TARANENKO and N. CHUDYK-BILOUSOVA, “The Role of Legal Service for Contractual Work Organization in Social and Medical Spheres,” University Scientific Notes, 2021, doi: 10.37491/unz.80.9.

[2] B. Alarie, A. Niblett, and A. H. Yoon, “How artificial intelligence will affect the practice of law,” 2018. doi: 10.3138/utlj.2017-0052.

[3] S. B. Shedthi, V. Shetty, R. Chadaga, R. Bhat, B. Preethi, and P. K. Kini, “Implementation of Chatbot that Predicts an Illness Dynamically using Machine Learning Techniques,” International Journal of Engineering, Transactions B: Applications, vol. 37, no. 2, pp. 312–322, Feb. 2024, doi: 10.5829/IJE.2024.37.02B.08.

[4] F. Firdaus, R. A. Rajagede, A. Sari, S. Hanifah, and D. A. Perwitasari, “Digital Assistant for Pharmacists Using Indonesian Language Based on Rules and Artificial Intelligence,” International Journal of Engineering, vol. 37, no. 9, pp. 1746–1754, 2024, doi: 10.5829/ije.2024.37.09c.04.

[5] S. Perez-Soler, S. Juarez-Puerta, E. Guerra, and J. De Lara, “Choosing a Chatbot Development Tool,” IEEE Softw, vol. 38, no. 4, pp. 94–103, 2021, doi: 10.1109/MS.2020.3030198.

[6] R. P. Karchi, S. M. Hatture, T. S. Tushar, and B. N. Prathibha, AI-Enabled Sustainable Development: An Intelligent Interactive Quotes Chatbot System Utilizing IoT and ML, vol. 1939 CCIS. 2023. doi: 10.1007/978-3-031-47055-4_17.

[7] A. Savanur, M. Niranjanamurthy, M. P. Amulya, and P. Dayananda, “Application of Chatbot for consumer perspective using Artificial Intelligence,” in Proceedings of the 6th International Conference on Communication and Electronics Systems, ICCES 2021, 2021. doi: 10.1109/ICCES51350.2021.9488990.

[8] S. Meshram, N. Naik, M. Vr, T. More, and S. Kharche, “Conversational AI: Chatbots,” in 2021 International Conference on Intelligent Technologies, CONIT 2021, 2021. doi: 10.1109/CONIT51480.2021.9498508.

[9] W. Sanjaya, Calvin, R. Muhammad, Meiliana, and M. Fajar, “Systematic Literature Review on Implementation of Chatbots for Commerce Use,” in Procedia Computer Science, 2023, pp. 432–438. doi: 10.1016/j.procs.2023.10.543.

[10] T. Jindal, L. N. U. Ishika, P. Sharma, and G. Kaur, Chatbots benefications towards the education sector. 2023. doi: 10.4018/978-1-6684-8671-9.ch006.

[11] M. W. Ashfaque, S. Tharewal, T. Malche, S. I. Malikb, and C. N. Kayte, “Analysis Of Different Trends In Chatbot Designing And Development: A Review,” in ECS Transactions, 2022, pp. 7215–7227. doi: 10.1149/10701.7215ecst.

[12] R. Negi and R. Katarya, “Emerging Trends in Chatbot Development : A Recent Survey of Design, Development and Deployment,” in 2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023, 2023. doi: 10.1109/ICCCNT56998.2023.10307280.

[13] H. Naveed et al., “A Comprehensive Overview of Large Language Models,” Jul. 2023, [Online]. Available: http://arxiv.org/abs/2307.06435

[14] T. B. Brown et al., “Language models are few-shot learners,” in Advances in Neural Information Processing Systems, 2020.

[15] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, “Large Language Models Struggle to Learn Long-Tail Knowledge,” in Proceedings of Machine Learning Research, 2023.

[16] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval Augmentation Reduces Hallucination in Conversation," Apr. 2021, [Online]. Available: http://arxiv.org/abs/2104.07567

[17] Z. Levonian et al., "Retrieval-augmented Generation to Improve Math Question-Answering: Trade-offs Between Groundedness and Human Preference," Oct. 2023, [Online]. Available: http://arxiv.org/abs/2310.03184

[18] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," May 2020, [Online]. Available: http://arxiv.org/abs/2005.11401

[19] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking Large Language Models in Retrieval-Augmented Generation," 2024. [Online]. Available: www.aaai.org

[20] W. S. Nsaif, H. M. Salih, H. H. Saleh, and B. Talib, "Conversational Agents: An Exploration into Chatbot Evolution, Architecture, and Important Techniques," in Eurasia Proceedings of Science, Technology, Engineering and Mathematics, 2024, pp. 246–262. doi: 10.55549/epstem.1518795.

[21] N. Biswas, S. Biswas, and S. Maity, Analysis of chatbots: History, use case, and classification. 2024. doi: 10.4018/979-8-3693-1830-0.ch004.

[22] K. B. Prakash, A. J. S. Kumar, and G. R. Kanagachidambaresan, Chatbot. 2021. doi: 10.1007/978-3-030-57077-4_9.

[23] P. Kandpal, K. Jasnani, R. Raut, and S. Bhorge, "Contextual chatbot for healthcare purposes (using deep learning)," in Proceedings of the World Conference on Smart Trends in Systems, Security and Sustainability, WS4 2020, 2020, pp. 625–634. doi: 10.1109/WorldS450073.2020.9210351.

[24] G. K. Ahirwar, Chatterbot: Technologies, tools and applications, vol. 913. 2020. doi: 10.1007/978-981-15-6844-2_14.

[25] C. Ionuț-Alexandru, Experimental Results Regarding the Efficiency of Business Activities Through the Use of Chatbots, vol. 276. 2022. doi: 10.1007/978-981-16-8866-9_27.

[26] P. Suta, X. Lan, B. Wu, P. Mongkolnam, and J. H. Chan, "An overview of machine learning in chatbots," International Journal of Mechanical Engineering and Robotics Research, vol. 9, no. 4, pp. 502–510, 2020, doi: 10.18178/ijmerr.9.4.502-510.

[27] M. Ahmed, H. U. Khan, and E. U. Munir, "Conversational AI: An Explication of Few-Shot Learning Problem in Transformers-Based Chatbot Systems," IEEE Trans Comput Soc Syst, vol. 11, no. 2, pp. 1888–1906, 2024, doi: 10.1109/TCSS.2023.3281492.

[28] A. Chizhik and Y. Zherebtsova, "Challenges of building an intelligent chatbot," in CEUR Workshop Proceedings, 2021, pp. 277–287.

[29] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," May 2020, [Online]. Available: http://arxiv.org/abs/2005.11401

[30] G. Izacard et al., "Atlas: Few-shot Learning with Retrieval Augmented Language Models," Aug. 2022, [Online]. Available: http://arxiv.org/abs/2208.03299

[31] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Of Wikipedia: Knowledge-powered conversational agents," in 7th International Conference on Learning Representations, ICLR 2019, 2019.

[32] Y. Mao et al., "Generation-augmented retrieval for open-domain question answering," in ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference, 2021. doi: 10.18653/v1/2021.acl-long.316.

[33] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.703.

[34] W. Yu, "Retrieval-augmented Generation across Heterogeneous Knowledge," in NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Student Research Workshop, 2022. doi: 10.18653/v1/2022.naacl-srw.7.

[35] S. S. Sonawane, P. N. Mahalle, and A. S. Ghotkar, Information Retrieval, vol. 104. 2022. doi: 10.1007/978-981-16-9995-5_4.

[36] E. Tzoukermann, J. L. Klavans, and T. Strzalkowski, Information Retrieval, vol. 9780199276. 2012. doi: 10.1093/oxfordhb/9780199276349.013.0029.

[37] M. Erritali, "Information retrieval: Textual indexing using an oriented object database," Indonesian Journal of Electrical Engineering and Computer Science, vol. 2, no. 1, pp. 205–214, 2016, doi: 10.11591/ijeecs.v2.i1.pp205-214.

[38] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2019.

[39] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," in EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

[40] B. Li, P. Yang, Y. Sun, Z. Hu, and M. Yi, "Advances and challenges in artificial intelligence text generation," Frontiers of Information Technology and Electronic Engineering, vol. 25, no. 1, pp. 64–83, 2024, doi: 10.1631/FITEE.2300410.

[41] D. Hijam, S. Gottipati, and S. Fardeen, "Telugu Text Generation with LSTM," in 2024 3rd International Conference on Smart Technologies and Systems for Next Generation Computing, ICSTSN 2024, 2024. doi: 10.1109/ICSTSN61422.2024.10670921.

[42] D. Weber-Wulff, S. Bjelobaba, T. Foltýnek, J. Guerrero-Dib, and L. Waddington, "Testing of Detection Tools for AI-Generated Text."

[43] J. Casas, M. O. Tricot, O. Abou Khaled, E. Mugellini, and P. Cudré-Mauroux, "Trends & methods in chatbot evaluation," in ICMI 2020 Companion - Companion Publication of the 2020 International Conference on Multimodal Interaction, 2020. doi: 10.1145/3395035.3425319.

[44] J. Ng, E. Haller, and A. Murray, "The ethical chatbot: A viable solution to socio-legal issues," Alternative Law Journal, vol. 47, no. 4, 2022, doi: 10.1177/1037969X221113598.

[45] M. Wyawahare, S. Roy, and S. Zanwar, "Generative vs Intent-based Chatbot for Judicial Advice," in 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation, IATMSI 2024, 2024. doi: 10.1109/IATMSI60426.2024.10502550.

[46] P. Giampieri, "AI-Powered Contracts: a Critical Analysis," International Journal for the Semiotics of Law, 2024, doi: 10.1007/s11196-024-10137-z.

[47] A. Kumar, P. Joshi, A. Saini, A. Kumari, C. Chaudhary, and K. Joshi, Smart Chatbot for Guidance About Children's Legal Rights, vol. 681. 2023. doi: 10.1007/978-981-99-1909-3_35.

[48] I. Qasse, S. Mishra, and M. Hamdaqa, "IContractBot: A Chatbot for Smart Contracts' Specification and Code Generation," in Proceedings - 2021 IEEE/ACM 3rd International Workshop on Bots in Software Engineering, BotSE 2021, 2021. doi: 10.1109/BotSE52550.2021.00015.

[49] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," May 2020, [Online]. Available: http://arxiv.org/abs/2005.11401

[50] B. Peng et al., "Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback," Feb. 2023, [Online]. Available: http://arxiv.org/abs/2302.12813

[51] S. Borgeaud et al., "Improving Language Models by Retrieving from Trillions of Tokens," in Proceedings of the 39th International Conference on Machine Learning, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., in Proceedings of Machine Learning Research, vol. 162. PMLR, May 2022, pp. 2206–2240. [Online]. Available: https://proceedings.mlr.press/v162/borgeaud22a.html

[52] G. Izacard et al., "Atlas: Few-shot Learning with Retrieval Augmented Language Models." [Online]. Available: https://github.com/

[53] W. Lex, "Code des obligations et des contrats, notamment les articles 77, 79, 80, 81, 84, 264 et 435, (Dahir du 12 août 1913 (9 ramadan 1331))," 1913.

[54] P. Zhang, S. Xiao, Z. Liu, Z. Dou, and J.-Y. Nie, "Retrieve Anything To Augment Large Language Models," Oct. 2023, [Online]. Available: http://arxiv.org/abs/2310.07554

[55] C. and S. D. Danopoulos Dimitrios and Kachris, "Approximate Similarity Search with FAISS Framework Using FPGAs on the Cloud," in Embedded Computer Systems: Architectures, Modeling, and Simulation, M. and J. M. Pnevmatikatos Dionisios N. and Pelcat, Ed., Cham: Springer International Publishing, 2019, pp. 373–386.

[56] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," Feb. 2017, [Online]. Available: http://arxiv.org/abs/1702.08734

[57] C. Mu, B. Yang, and Z. Yan, "An Empirical Comparison of FAISS and FENSHSES for Nearest Neighbor Search in Hamming Space," Jun. 2019, [Online]. Available: http://arxiv.org/abs/1906.10095

[58] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," Sep. 2023, [Online]. Available: http://arxiv.org/abs/2309.15217.

Fourth Party Logistics Routing Optimization Problem Based on Conditional Value-at-Risk Under Uncertain Environment

Guihua Bo[1]*, Qiang Liu[2], Huiyuan Shi[3], Xin Liu[4], Chen Yang[5], Liyan Wang[6]
College of Information and Control Engineering, Liaoning Petrochemical University, Fushun, Liaoning, China[1, 2, 3, 4, 5, 6]
State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang, China[3]

***Abstract*—In order to improve the level of logistics service and considering the impact of uncertainties such as bad weather and highway collapse on fourth party logistics routing optimization problem, this paper adopts Conditional Value-at-Risk (CVaR) to measure the tardiness risk, which is caused by the uncertainties, and proposes a nonlinear programming mathematical model with minimized CVaR. Furthermore, the proposed model is compared with the VaR model, and an improved Q-learning algorithm is designed to solve two models with different node sizes. The experimental results indicate that the proposed model can reflect the mean value of tardiness risk caused by time uncertainty in transportation tasks and better compensate for the shortcomings of the VaR model in measuring tardiness risk. Comparative analysis also shows that the effectiveness of the proposed improved Q-learning algorithm.**

***Keywords*—Logistics service; routing optimization; tardiness risk; conditional value-at-risk; improved Q-learning algorithm**

I. Introduction

With the deepening of economic globalization and the intensification of competition in the logistics market, service-oriented manufacturing enterprises are pursuing more hierarchical and integrated logistics services, so the problem of third-party logistics (3PL) is becoming increasingly prominent. For example, the Toyota listed on its website that Toyota outsources its logistics services to the 3PL to focus on their core product business, on 10th May 2021 [1]. In order to obtain more orders, competition among 3PL enterprises is becoming more fierce. There is a lack of resource sharing among various 3PL service providers, which makes it difficult to accurately grasp logistics information and meet the current rapidly growing logistics demand. As a result, the traditional 3PL model cannot adapt to the pace of the times and restricts the progress of logistics globalization. Instead of relying on the 3PL providers, some large manufacturers (e.g. Haier [2] and Hisense) have recently cooperated with Cainiao Logistics (the largest fourth party logistics [4PL] provider in China) in supply chain solutions. 4PL [3] is necessary to integrate and effectively connect resources so as to achieve complementary advantages. 4PL supplier is an integrator of the supply chain that integrates and manages different resources, capabilities, and technologies of a company and complementary service providers, and performs a detailed analysis of the entire supply chain system or industry logistics system where enterprise customers are located. Thus, it provides a comprehensive solution for the design, construction, and operation of the supply chain. Many enterprises have used 4PL to complete their own logistics tasks. For example, Cainiao Logistics, the largest Chinese 4PL provider founded by the Alibaba Group, incorporates over thirty 3PL providers to serve Taobao.com and Tmall.com, two largest online markets in China [4].

With the continuous development of 4PL, many experts and scholars have begun to focus on and study various aspects of 4PL. They are committed to in-depth exploration of issues related to risk management [4], network design [5-6], combinatorial auctions [7], contract design [8-9], and routing optimization [10] in 4PL, which have proposed a series of theoretical and practical achievements.

The Fourth Party Logistics Routing Optimization Problem (4PLROP) is one of the most important problems and a hot topic about 4PL. As the creator of transportation plans, 4PL needs to optimize and design distribution routes. Based on factors such as transportation time, transportation capacity, reputation indicators, and throughput of 3PLs, it selects and allocates 3PLs that provide distribution services to achieve path optimization and select satisfactory transportation plans for enterprises.

However, in the actual delivery process, due to unpredictable reasons such as transportation, bad weather, and human error operations, all of them may cause the transit time and transportation time to be not fixed but random during the transportation process, which may lead to the risk of delivery tardiness. The impact of this randomness makes 4PL suppliers to be unable to provide timely delivery plans that satisfy customers, it may lead to additional costs and a decrease in customer satisfaction. This situation may even affect the reputation of 4PL enterprises, leading to customer churn, and having adverse effects on their long-term development. Therefore, when solving 4PLROP, the influence of the risk caused by the uncertainty cannot be ignored.

This article adopts CVaR to describe and measure the average risk of tardiness induced by multiple factors in the real delivery process of 3PL providers in a 4PL operation. In addition, a mathematical model is proposed with CVaR minimization as the objective function and delivery cost as the constraint, and the suggested CVaR model is compared to the VaR model [11]. Then, an improved Q-learning algorithm is proposed to solve examples with different scales of the two models, and the effectiveness of the proposed CVaR model is verified. Finally, the proposed algorithm is compared with GA

embedded with Dijkstra [12] and the results verify the feasibility of improved Q-learning algorithm for solving this problem.

The principal contributions of this article are as follows:

1) In the context of uncertain environment, a novel 4PL route optimization problem that considers the risk of delays has been investigated;

2) The CVaR is employed to characterize risk, and a nonlinear programming model is established with constraint on delivery costs, aiming to minimize the risk as the objective;

3) An improved Q-learning algorithm is proposed to solve the model presented. Through this approach, 4PL can better adapt to the ever-changing market environment, ensuring the stability and efficiency of the supply chain.

The structure of this paper is arranged as follow: Section II gives the establishment of mathematical models and the transformations of CVaR model. Section III introduces the overall design of the proposed algorithm. Section IV performs some numerical computations. Section V finally concludes this paper.

II. Literature Review

The current research on 4PLROP can be broadly divided into problem structure, solution approach, and distribution factors. Huang et al. [13] studied 4PLROP from single point to single point and single task, and established a mathematical model based on nonlinear integer programming and multiple graphs. Li et al. [14] studied the routing optimization problem of multi-point to multi-point 4PL systems with reliability constraints. Tao et al. [15] established a mixed integer programming model for 4PLROP from the perspective of cost discount. Hong et al. [16] studied a multi-objective transportation optimization model according to queuing theory, considering the option of 3PL providers, routes, as well as transportation methods. They used a priority based stochastic enhanced elite genetic algorithm (GA) to solve the infeasible solutions in 4PLROP. According to the prospect theory of customer psychological behavior and customer service level, Huang et al. [17] developed a nonlinear integer programming model for the design of 4PL network and offered an approximation linear approach to convert the model to an equivalent linear model so as to demonstrate the efficacy of the proposed method. Yue et al. [18] designed a particle swarm optimization with adaptive inertia weight to solve the proposed mathematical model. Lu et al. [19] designed a combination of ant colony optimization algorithm and improved grey wolf optimization algorithm to solve the 4PLROP. Huang et al. [20] studied the risk management of outsourcing logistics under the principal-agent framework from the perspective of product quality.

For the 4PLROP, most studies are based on the determination of delivery time, which assumes that the delivery cost and delivery time used in the transportation process are fixed quantities, such as references [14,15]. However, in the actual delivery process, due to unpredictable reasons such as transportation, bad weather, and human error operations, all of them may cause the transit time and transportation time to be not fixed but random during the transportation process, which may lead to the risk of delivery tardiness. The impact of this randomness makes 4PL suppliers to be unable to provide timely delivery plans that satisfy customers, it may lead to additional costs and a decrease in customer satisfaction. This situation may even affect the reputation of 4PL enterprises, leading to customer churn, and having adverse effects on their long-term development. Therefore, when solving 4PLROP, the influence of the risk caused by the uncertainty cannot be ignored.

The important question is how to define, measure, and control the risk to improve their logistics service quality, which is in the best interest of 4PL and creates a win-win for both parties. This is the focus of our paper. Value-at-Risk (VaR) was used to measure the time risk in Reference[13,21], the VaR model only considered the possible tardiness time that will not exceed VaR with the confidence level, but it did not take into account the extreme events (when the amount of tardiness time exceeds the VaR value), in which the tardiness risk mean value should be considered.

Conditional Value-at-Risk (CVaR) is a risk measurement tool proposed by Rockafellar et al. [22] on the basis of VaR, which considers the tardiness risk mean value. It is mainly used in combinatorial optimization, setting risk limits, resource allocation, and financial supervision and credit risk measurement of various financial regulators on relevant enterprises and institutions. In recent years, it has been widely used in inventory management optimization [23], supply chain [24], selection of fourth party logistics suppliers, network design and other fields for risk measurement and optimization [25].

In summary, this paper adopts a new risk measure tool, CVaR, to measure the delay risk, sets up a stochastic programming mathematical model, and designs an improved Q-learning algorithm. The aim is to help 4PL to provide an optimal supply chain distribution solution and improve the level of logistics service.

III. Mathematical Model

A. Problem Description

In this section, the 4PLROP with consideration of delay risk is described, and the notations used throughout the paper are introduced.

A manufacturing company (such as Haier) wants to invest in designing a distribution route to deliver its products and services from plant to customer through DCs and 3PL providers to reduce costs and improve customer satisfaction. As a result, it employs a 4PL provider to offer a comprehensive supply chain solution. Specifically, manufacturing companies are investors. A 4PL provider needs to help investors integrate 3PL providers, select the number and location of DCs, and complete the distribution of product from plants to customer.

The 4PLROP requires not only the selection of the route from the plant to the customer, but also the determination of 3PL providers which provide the delivery service. That increases the difficulty of solving 4PLROP.

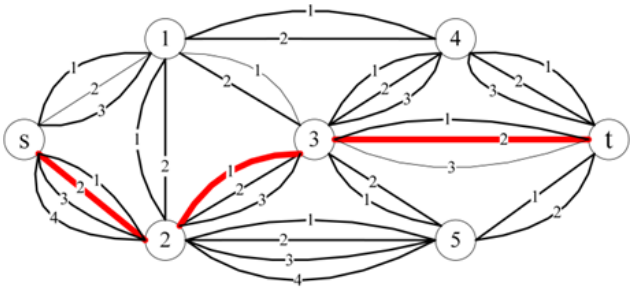


Fig. 1. Multiple graph for the 7-node problem.

A multiple graph, shown in Fig. 1, is used to describe the potential distribution network and demonstrate 4PLROP, where **V** represents the node cities, and **E** represents the edges. Among all nodes, indicates the supply city, indicates the target city, and others indicate the transit cities. All nodes have attributes such as time, cost, carrying capacity and reputation. In addition, there may be several edges between any two nodes because multiple 3PL suppliers may offer delivery services for any two cities. And each edge represents a 3PL, and the numeral on the edge is the serial number of 3PL.

Transportation time management is especially critical for 4PL providers. Traffic congestion, adverse weather, a surge in holiday order volumes, and node transfers can all cause uncertainty in time, leading to the risk of tardiness. Delayed delivery can directly affect customer interests, reduce the reputation of 4PL companies, and lead to customer loss, indirectly affecting the long-term development of 4PL supplier enterprises. Therefore, in order to improve customer satisfaction, 4PL suppliers need to consider the tardiness risk caused by time uncertainty. If the cost is within customer's budget, the less risk of tardiness, the better. Therefore, 4PL suppliers should monitor the tardiness risk mean value that may occur during transportation in real-time. Based on the above considerations, CVaR is introduced to quantify the average level of tardiness risk in delivery path, and a model with minimized CVaR and delivery cost as the constraint is developed.

TABLE I. THE DEFINITION OF PARAMETERS AND VARIABLES

Variables	Definition
r_{ij}	The quantity of 3PLs that offer transportation services between node cities i and j (i.e. the quantity of edges connecting two nodes)
C_{ijk}	The transportation cost required for the k-th 3PL supplier between node cities i and j
T_{ijk}	Random transportation time required for the k-th 3PL supplier between node cities i and j
$C_j^{'}$	Transfer cost required when passing through node city j
$T_j^{'}$	Random transit time required when passing through node city j
R	A path containing a set of nodes and edges, i.e. $R=\{\upsilon_s,2,\upsilon_2,1,\upsilon_3,2,\upsilon_t\}$ can be used to represent the red path in Fig. 1.

In order to set up the mathematical model, the definition of the parameters and variables are listed in Table I, the decision variables are defined as follows:

$$x_{ijk}(R)=\begin{cases}1 & \text{The } k\text{-th edge between nodes } i \\ & \text{and } j \text{ belongs to path } R \\ 0 & \text{others}\end{cases} \tag{1}$$

$$y_j(R)=\begin{cases}1 & \text{Node } j \text{ belongs to path } R \\ 0 & \text{others}\end{cases} \tag{2}$$

where $x_{ijk}(R)$ is used to determine whether the 3PL supplier provides a delivery task between cities i and j, and $y_j(R)$ indicates whether node city j provides a transfer task.

B. Mathematical Model Based on CVaR Criterion

The VaR model [21] only considered the possible tardiness time that will not exceed VaR with the confidence level β, but it did not take into account the extreme events (remaining $1-\beta$ when the amount of tardiness time exceeds the VaR value), the tardiness risk mean value is generated. Therefore, this paper adopts CVaR to improve the objective function of the model at the confidence level β, as shown in (3), the tardiness risk mean value generated when the tardiness time exceeds VaR, i.e. minimizing CVaR, is calculated to determine the delivery path with the lowest average tardiness risk. Therefore, the following mathematical model is established:

$$\min \quad CVaR_{\beta}(\Delta T) \tag{3}$$

$$\text{s.t.} \quad \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{r_{ij}} C_{ijk}x_{ijk}(R)+\sum_{j=1}^{n} C_j^{'}y_j(R)\le C_0 \tag{4}$$

$$\Delta T=\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{r_{ij}}\tilde{T}_{ijk}x_{ijk}(R)+\sum_{j=1}^{n}\tilde{T}_j x_{ijk}(R)\right)-T_0 \tag{5}$$

$$R=\left\{\upsilon_s,\cdots,\upsilon_i,k,\upsilon_j,\cdots,\upsilon_k\right\}\in G \tag{6}$$

$$x_{ijk}(R),y_j(R)\in\{0,1\} \tag{7}$$

Eq. (3) represents the objective function, which minimizes CVaR, where indicates the level of confidence, indicating the customer's level of risk aversion. The delivery cost constraint shown in (4) is the highest cost which is acceptable to the investor. Eq. (5) states the delayed time that the total delivery time needed on the route exceeds the due date, which is a random variable. Eq. (6) represents the constraint on the path, ensuring that it is a legitimate connecting path from the origin city to the target city. And then the constraints on decision variables are represented by (7).

C. Transformation of Mathematical Models

Theorem 1. The linear combination of a finite number of independent normal distribution random variables still obeys normal distribution. For a set of random variables $X_1,X_2,...,X_k,...,X_n$, if $X_1\sim N(\mu_1,\ \delta_1^2)$, $X_2\sim N(\mu_2,\ \delta_2^2),...,\ X_k\sim N(\mu_k,\ \delta_k^2),...,\ X_n\sim N(\mu_n,\ \delta_n^2)$, there is $\sum_{k=1}^{n}X_k\sim N(\sum_{k=1}^{n}\mu_k,\sum_{k=1}^{n}\delta_k^2)$.

Theorem 2. When the random variable follows the normal distribution, then $CVaR(\Delta T) = E(\Delta T) + c_1(\beta) \times STD(\Delta T)$, where $E(\Delta T)$ refers to the expectation of random variable, $STD(\Delta T)$ refers to the standard deviation of random variable, $c_1(\beta) = \varphi(\phi^{-1}(\beta)) \times (1-\beta)^{-1}$, $\phi^{-1}(\square)$ refers to the inverse function of standard normal distribution function, and $\varphi(\square)$ refers to the probability density of standard normal distribution [26].

This article assumes the random variable $\tilde{T}_{ijk} \sim N(\mu_{ijk}, \delta_{ijk}^2)$ and $\tilde{T}_j \sim N(\mu_j, \delta_j^2)$ in (5), so the objective function (3) can be converted to (8). Consequently, the proposed CVaR model includes (8), (4), (5), (6) and (7).

$$\min\Big(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k}^{r_{ij}}\mu_{ijk}x_{ijk}(R) + \sum_{j=1}^{n}\mu_j y_j - T_0\Big) + c_1(\beta) \times \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k}^{r_{ij}}\delta_{ijk}^2 x_{ijk}^2(R) + \sum_{j=1}^{n}\delta_j^2 y_j^2(R)} \tag{8}$$

IV. Algorithm Design

At present, the solution of the 4PLROP mainly includes the branch and bound, cut plane method, and other accurate solution algorithms that can only solve the optimization problem meeting specific conditions, as well as GA [27], particle swarm optimization algorithm [18] and ant colony algorithm [19], harmony search algorithm [13] and other simulation natural processes to form intelligent optimization algorithms. In solving 4PLROP, intelligent optimization algorithms can be roughly separated into two types, one approach is using repair strategies to repair illegal roads and obtain legitimate paths, which may result in the loss of good solution information during the repair process; another method is to utilize intelligent algorithms to construct simple graphs, followed by exact algorithms to discover the shortest path across the simple graph, which will waste a lot of storage space and computational time. Therefore, the Q-learning algorithm is used to solve the 4PL path problem in this article, directly resolving the proposed mathematical model on multiple graphs by setting state action pairs and reward values.

The Q-learning algorithm is proposed by Watkins [28] which involves the interaction between intelligent agents and the environment, constantly trying and learning, and ultimately obtaining one or more excellent behavior strategies. In a typical Q-learning algorithm, an intelligent agent decides to execute an action on the basis of the current state as well as past empirical knowledge. After executing the action, the agent transits to the following state according to a certain state transition strategy and obtains a return value. The Q-learning algorithm establishes a Q table based on the agent's state space and action space, which stores the corresponding Q values of all state action pairs. At the same time, the intelligent agent can be punished or rewarded by designing a reward function. When the action selected by the intelligent agent has an advantage in the environment, the state action pair can receive positive rewards from the environment, and the corresponding Q value of the state action pair will continue to increase. When the action selected by the intelligent agent is at a disadvantage in the environment, its state action pair will receive negative rewards from the environment, and its corresponding Q value will continue to decrease. The Q function is used to obtain the anticipated return value of a specific state action, which is updated in each training round and gradually approaches the optimal value. As shown in (9), the Q function is viewed as past-experienced knowledge that the agent has acquired, which is constantly updated and improved.

$$Q^*(s,a) \leftarrow Q(s,a) + \alpha\left[R(s,a,s') + \gamma \max Q(s',a') - Q(s,a)\right] \tag{9}$$

where $\gamma \in (0,1)$ is the discount factor, representing the degree of impact of the next state's Q value on the corresponding Q value of the current state, that is, the importance of immediate and future benefits. The larger the γ , the greater the weight given to prior experiences. The smaller the γ , the more emphasis is placed on immediate benefits R. When $\gamma = 0$,it means only focusing on current interests and not considering future interests. When $\gamma = 1$, it indicates a focus on past experience and future interests. Parameter α is the learning rate, which represents the learning speed of the agent throughout the entire learning process. Its range is $\alpha \in (0, 1)$, and it determines the degree that the new information obtained by the agent in the environment covers the old experience. The larger the α , the less effective it is to retain the previous training. The smaller the α , the more the effect of previous training will be retained.

A. Action State Settings

This article applies Q-learning algorithm to the 4PLROP, treating the selection of actions as a 3PL supplier selection problem, and treating node cities as different states. When the current node is s , the 3PL suppliers related to the node cities can be represented by action spaces $A = \{a_1, a_2, ..., a_k, ..., a_K\}, k = 1, 2, ..., K$, with every state action pair matching a Q value.

Taking 7 nodes as an example, as shown in Fig. 1, the initial node can be regarded as the initial state s . The nodes connecting the initial node are Transport Node 1 and Transport Node 2, respectively. There are three 3PL suppliers that can be selected from the initial node to Transport Node 1, with serial numbers 1, 2, and 3. Actions can be set to a_1, a_2, a_3 , respectively. There are four 3PL suppliers that can be selected from the initial node to Transport Node 2, with serial numbers 1, 2, 3, and 4, respectively, The action sequence numbers can be set to a_4, a_5, a_6, a_7 , and there are 7 corresponding actions that can be selected in the initial state. They are set to jump to the next state, namely transfer node 1, when the initial node selection action is a_1, a_2, a_3 , and also set to jump to the next state, namely transfer node 2, when the initial node selection action is a_4, a_5, a_6, a_7 . Other node action sets can be set by analogy. The selected the action and its corresponding next state for 7-node problem is shown in Table II.

TABLE II. 7-NODE PROBLEM ACTIONS AND CORRESPONDING NEXT STATES SETTINGS

Selected Action	The Corresponding Next State
$A=\{a_1,a_2,a_3\}$	Transit Node City 1
$A=\{a_4,a_5,\cdots,a_9\}$	Transit Node City 2
$A=\{a_{10},a_{11},a_{14},a_{15},a_{16}\}$	Transit Node City 3
$A=\{a_{12},a_{13},a_{21},a_{22},a_{23}\}$	Transit Node City 4
$A=\{a_{17},a_{18},a_{19},a_{20},a_{24},a_{25}\}$	Transit Node City 5
$A=\{a_{26},a_{27},\cdots,a_{33}\}$	Transit Node City t

B. Exploration Strategy

- When the intelligent agent interacts with the environment for learning, while selecting the known action with the maximum reward value, it is also necessary to ensure that more experience can be learned in the unknown environment, laying the foundation for obtaining more cumulative rewards. Therefore, it is necessary to set appropriate exploration strategies to achieve the optimal training effect. The exploration strategy adopted in this article is ε-greedy strategy.
- The mathematical description of the ε-greedy strategy is as follows:

$$\pi(a,s)=\begin{cases}\arg\max Q(s,a) & 1-\varepsilon \\ a_{random} & \varepsilon\end{cases} \tag{10}$$

- For (10), it can be understood as a certain probability ε. Randomly select the actions that can be selected in the current state, with 1-ε probability selects the action corresponding to the maximum Q value in the current action.

C. Construction of Reward Function

Due to the fact that the Q-learning algorithm is on the basis of the Markov Decision Process model, a more computationally efficient discrete reward and penalty function is adopted. Eq. (11) indicates that when there is a connection between two cities and it is not the endpoint, the reward is 1. When there is no connection between two cities, the reward is -1. When there is a connection between two cities and it is the endpoint, the reward is 100.

$$r(s,a)=\begin{cases}1 & \text{i. j has a connection point j is not a termination node} \\ -1 & \text{i. j is not connected} \\ 100 & \text{i. j is connected and j is the termination node}\end{cases} \tag{11}$$

Considering that the size of the reward is related to the mean and variance in the objective function, while also meeting certain cost constraints, the reward value function (12) can be constructed to ensure that the agent does not violate the constraints and obtains the optimal delivery path for the objective value. As shown in (13), ω_1 is related to the delivery cost related to the selected 3PL provider and transportation node, and the smaller the delivery cost, the larger the reward value obtained; ω_2, shown in (14), is related to the mean of the random time related to the selected 3PL provider and transit node. The smaller the mean of the random time, the greater the reward value obtained; as shown in (15), ω_3 is related to the variance of the random time related to the selected 3PL provider and transit node. While the variance of the random time is smaller, the reward value obtained is larger, where k_1 and k_2 are the weighting coefficients of the reward function.

$$r=\omega_1 r(s,a)+\omega_2 r(s,a)+\omega_3 r(s,a) \tag{12}$$

$$\omega_1=\frac{k_1}{C_{ijk}+C_j} \tag{13}$$

$$\omega_2=\frac{k_2}{\mu_{ijk}+\mu_j} \tag{14}$$

$$\omega_3=\frac{1-k_1-k_2}{\delta_{ijk}^2+\delta_j^2} \tag{15}$$

D. Model Training

The agent is trained by designing a Q table, in which each row represents all the states that the agent can choose, each column represents the actions that the agent can perform in the corresponding state, each state represents different city nodes in multiple graphs, and each action represents different 3PL suppliers in multiple graphs. Initially, set all states in the Q table to 0, and then calculate the reward values obtained by executing different actions (selecting different suppliers) based on the reward matrix established by the reward function. Use (9) to update the values of each element in the Q table. Treat each iteration as a training session for the agent. For each training session, the agent attempts to reach the destination node from the initial node, and after each action, updates the elements in the Q table.

E. Improved Q-learning Algorithm Process and Steps

When the improved Q-learning algorithm is used to solve 4PLROP, first initialize the elements in the matrix using a reward function based on existing data. Due to the existence of several various 3PL providers across two node cities, there is one state corresponding to multiple. Consequently, it is necessary to set the corresponding actions for each state, and then train and update the matrix Q through the setting of matrix R and related parameters. Finally, the optimal path planning can be obtained based on the Q table. The specific steps for solving the proposed model using the improved Q-learning algorithm are as follows:

Step 1: Preprocess the known factors of the problem.

Step 2: Import known information into Matlab.

Step 3: Initialize the parameters γ, α and Q table, set the initial and final states, and generate a reward matrix using (12) based on existing data.

Step 4: Initialize the state to the initial node.

Step 5: Utilize ε-greedy strategy selection action (the optional 3PL supplier corresponding to the state).

Step 6: Execute the action a (select a 3PL supplier from the current node), transfer to a new state s' (next node city), and update the Q table based on the reward matrix R and related parameter settings.

Step 7: Determine whether s' is in a terminated state. If not, proceed to step 5. Else, proceed to step 8.

Step 8: Determine whether the training frequency has been reached. If not, continue with step 4, otherwise continue with step 9.

Step 9: After training, output the Q table.

Step 10: Output the optimal delivery plan based on the Q table.

V. Experimental Results and Analysis

The improved Q-learning algorithm is used to solve different scale examples in this section to analyse the algorithm performance and the model effectiveness. The proposed CVaR model effectiveness is verified by comparing and analyzing the solution results of the two models. The algorithm is implemented using software MATLAB and runs in the Intel (R) Core (TM) i7-2600 @ 3.40GHz environment.

A. Parameter Testing Analysis

By conducting extensive experimental simulations to test parameters, the method is to observe the impact of certain parameters change on the solution results while keeping other parameters constant. The experimental results demonstrate that the optimal parameters are $\gamma = 0.8$, $\alpha = 0.9$, $episode = 100$, $\varepsilon = 0.95$.

On the basis of the data in Table III, through repeated experiments on the weighting coefficients k_1 and k_2 in reward functions of different scales, the improved Q-learning algorithm performs best.

B. Model Performance Analysis

Verify the effectiveness of the proposed CVaR model and improved Q-learning algorithm by solving several different scale examples. First, the 7-node problem is used as an instance, and the solution results obtained using this algorithm are carefully analyzed. Then, to demonstrate the validity of the model, VaR and CVaR values are solved for four examples of different sizes with 7, 15, 30 and 50 nodes, and the results are analyzed.

The solution results of the VaR model and CVaR model for the 7-node problem with different values are shown in TABLE IV, where β denotes the confidence level, i.e., the risk attitude of the customer, T_0 denotes the customer's latest acceptable delivery time, C_0 is the customer's latest acceptable delivery cost. The value of VaR is the optimal solution obtained from the VaR model, the value of CVaR is the optimal solution obtained from the CVaR model, the Best Path is the distribution path that corresponds to the optimal solution, and Best Rate is the probability that the total number of runs obtains the best solution when the algorithm is used to solve. At this time, the total number of runs is 100, Time/s indicates the time for the algorithm to run once.

TABLE III. Solutions and Parameter Settings for Different Instances

Number ofNodes	k_1	k_2	*Episode*	*CVaR*	Best Path	Time/s
7	0.6	0.3	100	32.0456	$R=\{\upsilon_s,2,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$	0.9s
15	0.1	0.7	200	3.8184	$R=\{\upsilon_s,1,\upsilon_2,2,\upsilon_6,3,\upsilon_{13},2,\upsilon_t\}$	1s
30	0.37	0.357	200	13.6412	$R=\{\upsilon_s,1,\upsilon_4,2,\upsilon_8,1,\upsilon_{12},1,$ $\upsilon_{15},2,\upsilon_{18},4,\upsilon_{21},1,\upsilon_{25},1,\upsilon_t\}$	1.5s
50	0.2	0.5	500	8.3652	$R=\{\upsilon_s,3,\upsilon_{39},1,\upsilon_{28},1,\upsilon_{29},1,$ $\upsilon_{10},1,\upsilon_{42},3,\upsilon_{37},3,\upsilon_t\}$	9.2s

TABLE IV. Solution Results of 7-node Problem when T_0=80 and C_0=73.

β	T_0	C_0	*VaR*	*CVaR*	Best Path	Best Rate	Time/s
0.9	80	73	10.9730	22.0456	$R=\{\upsilon_s,2,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$	1	0.9s
0.95	80	73	19.4699	29.2428	$R=\{\upsilon_s,2,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$	1	0.9s
0.99	80	73	35.4087	43.3341	$R=\{\upsilon_s,2,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$	1	0.9s

From the data in the Table IV, it can be seen that the best path corresponding to the best VaR value and CVaR value is the same. When the confidence level is 0.9 and the VaR value that meets the cost constraint is 10.9730, it means that the 4PL supplier has a 90% probability of ensuring that the delay amount will not exceed 10.9730. The CVaR value that satisfies the cost constraint is 22.0456, indicating that the tardiness risk mean value when the delivery task's delay exceeds the VaR is 22.0456, The related delivery cost is 73, and the best delivery path is $R=\{\upsilon_s,2,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$, which refers to the selection of transit cities 2 and 3 for transportation from the source node city s to the destination node city t, and the numbers of the 3PL supplier chosen between every two cities are 2, 2, and 1. When the confidence level is 0.95 and the VaR value that satisfies the cost constraint is 19.4699, it means that the 4PL supplier has a 95% probability of ensuring that the delay amount will not exceed 19.4699. The CVaR value that satisfies the cost constraint is 29.2428, indicating that the tardiness risk mean value when the delivery task's delay exceeds the VaR is 29.2428. The associated delivery cost is 73, and the best

delivery path is $R=\{\upsilon_s,2,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$, which refers to the selection of transit cities 2 and 3 for transportation from the source node city s to the destination node city t, and the 3PL supplier numbers selected between each two cities are 2, 2, and 1. When the confidence level is 0.99 and the VaR value that satisfies the time constraint is 35.4087, it means that the 4PL supplier has a 99% probability of ensuring that the delay amount will not exceed 35.4087. The CVaR value that satisfies the cost constraint is 43.3341, indicating that the tardiness risk mean value when the delivery task's delay exceeds VaR is 43.3341. The associated delivery cost is 73, and the best delivery path is $R=\{\upsilon_s,2,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$, which refers to the selection of transit cities 2 and 3 for transportation from the source node city s to the destination node city t, and the 3PL supplier numbers selected between each two cities are 2, 2, and 1. The above data indicates that when other constraints are constant, as the confidence level grows, the best delivery path will not change. However, customers will face higher tardiness risks, and the corresponding VaR is often smaller than CVaR. This also verifies that the VaR model can only evaluate the probability of risk occurrence, while the CVaR model can effectively measure tail risk and estimate the tardiness risk mean value faced by delivery tasks in extreme situations. Compared to VaR models, it can better reflect potential tardiness risks.

The statistics in Table V show that the 7-node problem's solution is provided for various combinations of confidence level, delivery time, and delivery cost. We know from analyzing these data that while the confidence level and time constraints are consistent, as the delivery cost increases, the corresponding VaR and CVaR values will decrease, indicating that the delay risk faced by the delivery path will be reduced. When the delivery cost and time constraints remain unchanged and the confidence level increases, the corresponding VaR and CVaR values will increase, indicating an increase in the risk of tardiness faced by the delivery path. When the confidence level and delivery cost are constant, as the time constraint increases, the corresponding VaR and CVaR values will increase, and thus the delay risk faced by the distribution path will rise. In addition, by comparing the VaR value and the CVaR value, it can be seen that when other conditions are the same, the CVaR value obtained is always greater than the VaR value, which also verifies that CVaR is more able to reflect the potential value at risk than VaR. Therefore, when using the VaR model to measure tardiness risk failure, 4PL suppliers can use the CVaR model to make up for the shortcomings of the VaR model. Combined with the risk tolerance of customers, they can comprehensively consider the risk level and expected tardiness risk of the distribution scheme, monitoring of potential tardiness risks in real time, providing customers with the delivery path with the minimum tardiness risk mean value at a given confidence level, and estimating the tardiness risk generated when extreme events occur, making reasonable delivery service decisions, thereby improving customer satisfaction.

Tables VI to VIII provide the solution results of the VaR model and CVaR model for the 15 node problem with time constraints, cost constraints, 30 node problem with time constraints, and 50 node problem with time constraints and cost constraints, respectively. When using the Q-learning algorithm to solve the proposed model, the solution time for small and medium-sized examples is about 1 second, and for large-scale problems, the solution time does not exceed 10 seconds. This fully demonstrates that the Q-learning algorithm has a high solution speed and high stability. When solving a 15 node problem, the training number is 200, and the optimal solution rate of the algorithm is as high as 98%, that is, the algorithm runs 100 times, 98 times can obtain the optimal solution, and when solving the 50 node problem, the best rate also reaches 95%, that is, the algorithm runs 100 times, and 95 times can obtain the optimal solution.

C. The Influence of Confidence Level

To investigate the influence of confidence level β on the 4PLROP, we provide three values of β with four different cases, which is customer's degree of risk appetite, shown in Tables V to VIII.

TABLE V. SOLUTION FOR 7-NODE PROBLEMS UNDER DIFFERENT CONFIDENCE LEVELS, TIME CONSTRAINTS, AND COST CONSTRAINTS

β	T_0	C_0	*Episode*	*VaR*	*CVaR*	Best Path	Time/s
0.9	70	73	100	20.9730	32.0456	$R=\{\upsilon_s,2,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$	0.9s
		75	100	17.2239	28.0198	$R=\{\upsilon_s,3,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$	0.9s
	80	73	100	10.9730	22.0456	$R=\{\upsilon_s,2,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$	0.9s
		75	100	7.2239	18.0198	$R=\{\upsilon_s,3,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$	0.9s
0.95	70	73	100	29.4699	39.2428	$R=\{\upsilon_s,2,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$	0.9s
		75	100	25.5084	35.0371	$R=\{\upsilon_s,3,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$	0.9s
	80	73	100	19.4699	29.2428	$R=\{\upsilon_s,2,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$	0.9s
		75	100	15.5084	25.0371	$R=\{\upsilon_s,3,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$	0.9s
0.99	70	73	100	45.4087	53.3341	$R=\{\upsilon_s,2,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$	0.9s
		75	100	41.0489	48.7762	$R=\{\upsilon_s,3,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$	0.9s
	80	73	100	35.4087	43.3341	$R=\{\upsilon_s,2,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$	0.9s
		75	100	31.0489	38.7762	$R=\{\upsilon_s,3,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$	0.9s

TABLE VI. SOLUTION FOR THE 15-NODE PROBLEM WHEN T_0=70, C_0=115

β	*Eposide*	*VaR*	*CVaR*	Best Path	Best Rate	Time/s
0.9	200	0.5969	3.7728	$R=\{\upsilon_s,1,\upsilon_2,2,\upsilon_6,3,\upsilon_{13},2,\upsilon_t\}$	0.98	1s
0.95	200	3.0340	5.8371	$R=\{\upsilon_s,1,\upsilon_2,2,\upsilon_6,3,\upsilon_{13},2,\upsilon_t\}$	0.98	1s
0.99	200	7.3406	9.4295	$R=\{\upsilon_s,2,\upsilon_2,2,\upsilon_6,3,\upsilon_{13},2,\upsilon_t\}$	0.98	1s

TABLE VII. SOLUTION FOR THE 30-NODE PROBLEM WHEN T_0=115, C_0=190

β	*Eposide*	*VaR*	*CVaR*	Best Path	Best Rate	Time/s
0.9	200	10.5009	13.6412	$R=\{\upsilon_s,1,\upsilon_4,2,\upsilon_8,1,\upsilon_{12},1,$ $\upsilon_{15},2,\upsilon_{18},4,\upsilon_{21},1,\upsilon_{25},1,\upsilon_t\}$	0.97	1.5s
0.95	200	12.9107	15.6825	$R=\{\upsilon_s,1,\upsilon_4,2,\upsilon_8,1,\upsilon_{12},1,$ $\upsilon_{15},2,\upsilon_{18},4,\upsilon_{21},1,\upsilon_{25},1,\upsilon_t\}$	0.97	1.5s
0.99	200	17.4312	19.679	$R=\{\upsilon_s,1,\upsilon_4,2,\upsilon_8,1,\upsilon_{12},1,$ $\upsilon_{15},2,\upsilon_{18},4,\upsilon_{21},1,\upsilon_{25},1,\upsilon_t\}$	0.97	1.5s

TABLE VIII. SOLUTION FOR THE 50-NODE PROBLEM WHEN T_0=150, C_0=165

β	*Eposide*	*VaR*	*CVaR*	Best Path	Best Rate	Time/s
0.9	500	4.4900	8.3652	$R=\{\upsilon_s,3,\upsilon_{39},1,\upsilon_{28},1,\upsilon_{29},1,$ $\upsilon_{10},1,v_{42},3,\upsilon_{37},3,\upsilon_t\}$	0.95	1.5s
0.95	500	7.4637	10.4637	$R=\{\upsilon_s,3,\upsilon_{39},1,\upsilon_{28},1,\upsilon_{29},1,$ $\upsilon_{10},1,\upsilon_{42},3,\upsilon_{37},3,\upsilon_t\}$	0.95	1.5s
0.99	500	13.042	15.8157	$R=\{\upsilon_s,3,\upsilon_{39},1,\upsilon_{28},1,\upsilon_{29},1,$ $\upsilon_{10},1,\upsilon_{42},3,\upsilon_{37},3,\upsilon_t\}$	0.95	1.5s

The result in these four tables shows that the CVaR, the tardiness risk, increases with the confidence level β increasing. Because with a smaller β, the 3PL can make more effort and the tardiness risk is smaller. 4PL can select the proper delivery solution for the investor to control the tardiness risk according to the customer's degree of risk aversion.

D. The Influence of Cost and Due Date

To investigate the influence of cost C_0 and due date T_0 on the 4PLROP, we provide two values of C_0 and T_0 with 7-node problem, shown in Table V.

The result in Table V shows that the CVaR, the tardiness risk, decreases with the C_0 and T_0 increasing when β in fixed. Because with a fixed β, if the budget or time is enough, the 3PL can make more effort and the tardiness risk is smaller. 4PL can select the proper delivery solution for the investor to control the tardiness risk according to the customer's budget and due date.

E. Algorithm Comparison

This section uses the improved Q-learning and GA embedded with Dijkstra algorithm to solve three different scale examples, and the comparative data is shown in Table IX.

TABLE IX. COMPARISON OF RESULTS OF DIFFERENT ALGORITHMS

Number of Nodes	Algorithm	*CVaR*	Best Path	Best Rate	Time/s
7	Improved Q-learning	22.0456	$R=\{\upsilon_s,2,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$	1	0.9s
	GA embedded with Dijkstra	22.0456	$R=\{\upsilon_s,2,\upsilon_2,2,\upsilon_3,1,\upsilon_t\}$	0.95	19.4s
15	Improved Q-learning	3.7728	$R=\{\upsilon_s,1,\upsilon_2,2,\upsilon_6,3,\upsilon_{13},2,\upsilon_t\}$	0.98	1s
	GA embedded with Dijkstra	3.7728	$R=\{\upsilon_s,1,\upsilon_2,2,\upsilon_6,3,\upsilon_{13},2,\upsilon_t\}$	0.94	24.5s
30	Improved Q-learning	13.641	$R=\{\upsilon_s,1,\upsilon_4,2,\upsilon_8,1,\upsilon_{12},1,$ $\upsilon_{15},2,\upsilon_{18},4,\upsilon_{21},1,\upsilon_{25},1,\upsilon_t\}$	0.95	1.5s
	GA embedded with Dijkstra	13.641	$R=\{\upsilon_s,1,\upsilon_4,2,\upsilon_8,1,\upsilon_{12},1,$ $\upsilon_{15},2,\upsilon_{18},4,\upsilon_{21},1,\upsilon_{25},1,\upsilon_t\}$	0.9	28.5s

It can be seen from the Table IX that both methods can find the optimal solution when solving small-scale problems with 7 nodes. However, the latter performs poorly in solving speed because that GA embedded with Dijkstra algorithm is used to generate the simple graph based on the multi-graph shown in Fig. 1 and the Dijkstra algorithm is used to generate the shortest path on the generated simple graph, but the shortest path may be not met the constraints. Thus, it needs a lot of time to find the feasible solution of the problem. However, the improved Q-learning algorithm directly solves the problem on

the multi-graph shown in Fig. 1, which saves much computational time. Furthermore, as the solution size increases, the improved Q-learning algorithm exhibits higher solving speed and quality.

F. Discussion

In summary, the results show that when the delivery costs and time constraints remain unchanged, the higher the confidence level, the higher the corresponding values of VaR and CVaR, which means that the risk of delivery tardiness faced by customers will increase. When other conditions such as confidence level, time, and cost constraints are the same, the obtained CVaR value is always greater than the VaR value. The proposed CVaR model can reflect the average delay risk of delayed delivery exceeding the VaR value due to various factors, better compensating for the shortcomings of the VaR model in measuring tardiness risk, and real-time monitoring of potential tardiness risks that may occur during the delivery process. 4PL can adjust the customer's aversion to risk, use this model to calculate the tardiness risk mean value and provide a reliable delivery plan.

VI. Conclusion

This article fully considers the tardiness risk caused by the uncertainty of transit time and transportation time in the actual delivery process in complex and uncertain environments. A risk measurement tool CVaR is introduced to measure and control the risk, and a mathematical model with CVaR minimization as the optimization objective and distribution cost as the constraint is established. Meanwhile, the proposed algorithm is compared with GA embedded with Dijkstra. The results demonstrate that the proposed model is effective for the 4PLROP and improved Q-learning algorithm can solve the large-scale 4PL path problem rapidly and with excellent stability. 4PL can adjust the customer's aversion to risk, use this model to calculate the tardiness risk mean value and provide a reliable delivery plan. Customers can obtain multiple schemes according to their risk preferences and take corresponding measures. This article provides scientific decision making basis and efficient and safe distribution plans for the 4PL, which can improve the level of logistics service.

Meanwhile, the stochastic variables' probability distributions may follow other distributions, such as the exponential distribution, uniform distribution, etc. Therefore, our research can be extended to a robust 4PLROP considering delay risk or multiple risks.

Acknowledgment

This research was funded by the National Natural Science Foundation of China, grant number 62203202; Natural Science Fund Project of Liaoning Province, grant number 2022-BS-295; Youth Project of the Educational Department of Liaoning Province, Grant number LJKQZ20222432.

References

[1] Y. X. Zhang, Z. M. Gao, M. Huang, S. C. Jiang, M. Q. Yin, and S. C. Fang, "Multi-period distribution network design with boundedly rational customers for the service-oriented manufacturing supply chain: a 4PL perspective,"Int. J. Prod. Res., vol. 62, pp. 7412-7431, 2022.

[2] E. Lee, "Alibaba's cainiao logistics conffrms first financing at $7.7B valuation," The technode, March 15. 2016. https://technode.com/2016/03/15/alibaba-cainiaofunding/.

[3] F. Q. Lu, W. D. Chen, W. J. Feng, and H. L. Bi, "4PL routing problem using hybrid beetle swarm optimization," Soft Comput., vol. 27, pp. 17011, 2023.

[4] M. Huang, J. Tu, X. Chao, and D. Jin, "Quality risk in logistics outsourcing: A fourth party logistics perspective," Eur. J. Oper. Res., vol. 276, pp. 855-879, 2019.

[5] M. Huang, L. W. Dong, H. B. Kuang, Z. Z. Jiang, L. H. Lee, X.W. Wang, "Supply chain network design considering customer psychological behavior−a 4PL perspective," Comput. Ind. Eng. pp. 159, 2021.

[6] M. Q. Yin, M. Huang, X. H. Qian, D. Z. Wang, X. W. Wang, L. H. Lee, "Fourth-party logistics network design with service time constraint under stochastic demand," J. Intell. Manuf. vol. 34, pp. 1203‐1227, 2023.

[7] F. Q. Lu, H. L. Bi, W. J. Feng, Y. L. Hu, S. X. Wang, and X. Zhang, "A two-stage auction mechanism for 3pl supplier selection under risk aversion," Sustainability, vol. 13, pp. 9745, 2021.

[8] H. Y. Wang, M. Huang, H. F. Wang, and Y. J. Zhou, "Fourth party logistics service quality management with logistics audit," J. Indus. Manag. Optim., Vol. 19, pp. 7105-7129, 2023.

[9] H. Y. Wang, M. Huang, H. F. Wang, X. H. Feng, and Y. J. Zhou, "Contract design for the fourth party logistics considering tardiness risk," Int. J. Indus. Eng. Comput., vol. 13, pp. 13-30, 2022.

[10] F. Q. Lu, W. J. Feng, M. Y. Gao, H. L. Bi, and S. X. Wang, "The fourth-party logistics routing problem using ant colony system-improved grey wolf optimization," J. Adv. Transp., vol. 2022, pp. 9864064, 2022.

[11] J. H. Wu, "The relationship between port logistics and international trade based on VAR model," J. Coastal Res., pp. 601-604, 2020.

[12] X.F. Lyu, Y. C. Song, C. Z. He, Q. Lei, and W. F. Guo, "Approach to integrated scheduling problems considering optimal number of automated guided vehicles and conflict-free routing in flexible manufacturing systems," IEEE Access, vol. 7, pp. 74909-74924, 2019

[13] G. Bo, M. Huang, "Model and Solution of Routing Optimization Problem in the Fourth Party Logistics with Tardiness Risk,"ComplexSyst. Complex. Sci, vol. 15, no. 03, pp. 66‐74, 2018.

[14] J. Li, Y. Liu, Y. Zhang, and S. Xu, "Algorithms for routing optimization in multipoint to multipoint 4PL system," Discr. Dyn. Nat. Soc., vol. 2015, pp. 426947, 2015.

[15] Y. Tao, E. P. Chew, L. H. Lee, and Y. Shi, "A column generation approach for the route planning problem in fourth party logistics," J. Oper. Res. Soc., vol. 68, pp. 165-181, 2017.

[16] W. Hong, Z. Xu, W. Liu, L. Wu, and X. Pu, "Queuing theory-based optimization research on the multi-objective transportation problem of fourth party logistics," Proc. Inst. Mech. Eng. B J. Eng. Manuf., vol. 235, pp. 1327-1337, 2021.

[17] M. Huang, L. Dong, H. Kuang, Z. Z. Jiang, L. H. Lee, and X. Wang, "Supply chain network design considering customer psychological behavior-a 4PL perspective," Comp. Ind. Eng., vol. 159, pp. 107484, 2021.

[18] D. Yue, M. Huang, M. Yin, "PSO algorithm for the fourth party logistics network design considering multi-customer behavior under stochastic demand," In 2017 29th Chinese Control And Decision Conference, Chongqing, China, 01 May 2017.

[19] F. Lu, W. Feng, M. Gao, H. Bi, and S. Wang, "The fourth-party logistics routing problem using ant colony system-improved grey wolf optimization," J. Adv. Transp., vol. 2020, pp. 1-15, 2020.

[20] M. Huang, L. Dong, H. Kuang, and X. Wang, "Reliable fourth party logistics location-routing problem under the risk of disruptions," IEEE Access, vol. 9, pp. 84857-84870, 2021.

[21] X. Liu, G. Bo, "Q-learning algorithm for fourth party logistics route optimization considering tardiness risk," Proceedings of the 2022 International Conference on Cyber-Physical Social Intelligence, 2022.

[22] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," J. risk, vol. 2, pp. 21-42, 2000.

[23] W. Xue, L. Ma, and H. Shen, "Optimal inventory and hedging decisions with CVaR consideration," Int. J. Prod. Econ., vol. 162, pp. 70-82, 2015.

[24] V. Dixit, P. Verma, and M. K. Tiwari, "Assessment of pre and post-disaster supply chain resilience based on network structural parameters with CVaR as a risk measure," Int. J. Prod. Econ., vol. 227, pp. 107655, 2020.

[25] F. Ding, M. Liu, S. M. Hsiang, P. Hu, Y. Zhang, and K. Jiang, "Duration and labor resource optimization for construction projects-a conditional-value-at-risk-based analysis,"Buildings, vol. 14, pp.1-20, 2024.

[26] R. T. Rockafellar and S. Uryasev, "Conditional Value-at-Risk for general loss distributions," J. Bank. Fin., vol. 26, pp. 1443-1471, 2002.

[27] J. Li, Y. Liu, and Z. Hu, "Routing optimization of fourth party logistics with reliability constraints based on Messy GA," J. Ind. Eng. Manag., vol. 7, pp. 1097-1111, 2014.

[28] C. J. C. H. Watkins, "Learning from delayed rewards," Ph.D., University of Cambridge, Cambridgeshire, May 1989.

Optimized Dynamic Graph-Based Framework for Skin Lesion Classification in Dermoscopic Images

J. Deepa, P. Madhavan
Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

***Abstract*—Early and accurate classification of skin lesions is critical for effective skin cancer diagnosis and treatment. However, the visual similarity of lesions in their early stages often leads to misdiagnoses and delayed interventions. This lack of transparency makes it challenging for dermatologists to interpret with validate decisions made by such methods, reducing their trust in the system. To overcome these complications, Skin Lesions Classification in Dermoscopic Images using Optimized Dynamic Graph Convolutional Recurrent Imputation Network (SLCDI-DGCRIN-RBBMOA) is proposed. The input image is pre-processed utilizing Confidence Partitioning Sampling Filtering (CPSF) to remove noise, resize, and enhance image quality. By using the Hybrid Dual Attention-guided Efficient Transformer and UNet 3+ (HDAETUNet3+) it segment ROI region of the preprocessed dermoscopic images. Finally, segmented images are fed to Dynamic Graph Convolutional Recurrent Imputation Network (DGCRIN) for classifying skin lesion as actinic keratosis, dermatofibroma, basal cell carcinoma, squamous cell carcinoma, benign keratosis, vascular lesion, melanocytic nevus, and melanoma. Generally, DGCRIN does not express any adaption of optimization strategies for determining optimal parameters to exact skin lesion classification. Hence, Red Billed Blue Magpie Optimization Algorithm (RBBMOA) is proposed to enhance DGCRIN that can exactly classify type of skin lesion. The proposed SLCDI-DGCRIN-RBBMOA technique attains 26.36%, 20.69% and 30.29% higher accuracy, 19.12%, 28.32%, and 27.84% higher precision, 12.04%, 13.45% and 22.80% higher recall and 20.47%, 16.34%, and 20.50% higher specificity compared with existing methods such as a deep learning method dependent on explainable artificial intelligence for skin lesion classification (DNN-EAI-SLC), multiclass skin lesion classification utilizing deep learning networks optimal information fusion (MSLC-CNN-OIF), and classification of skin cancer from dermoscopic images utilizing deep neural network architectures (CSC-DI-DCNN) respectively.**

***Keywords*—Confidence partitioning sampling filtering; dynamic graph convolutional recurrent imputation network; ISIC-2019 skin disease dataset; red billed blue magpie optimization algorithm; hybrid dual attention-guided efficient transformer and UNet 3+**

I. INTRODUCTION

Skin cancer is considered one of the most dangerous cancer types due to its high prevalence and potential for metastasis if not detected early. Melanoma, in particular, has a high mortality rate and can spread rapidly to other organs [1]. Statistics show that skin cancer is more common than other cancers, with rising incidence rates globally. Comparative studies also indicate that, although treatable when caught early, skin cancer has a higher likelihood of fatal outcomes compared to many other cancer types [2]. A skin lesion refers to abnormal appearance otherwise growth of skin analyzed to surrounding region [3]. Lesions can vary in color, type, shape, texture, position, distribution, they are categorized to classification prearranged hierarchically [4]. The first two major classifications of this hierarchy are melanocytic, non-melanocytic lesions [5]. The classification as melanocytic or non-melanocytic depends on presence or absence of melanocytes, melanin pigment in lesion [6]. Melanocytic lesions possess eight global features that assist in detailed categorization of pigmented skin lesions, along with fourteen local features provide more specific information about each lesion [7]. Hemoglobin causes non-melanocytic lesions to seem purple, red, blue, or black, while keratin causes them to appear orange or yellow [8]. These lesions can be either cancerous otherwise non-cancerous. Dermoscopic is commonly utilized skin imaging methods, aimed at enhancing diagnostic accuracy, reducing mortality from skin cancer [9]. This non-invasive method captures magnified with well-illuminated image of skin, allowing for a clearer examination of lesion area [10]. This method improves doctors' diagnostic ability and is typically used to detect skin cancer in its premature stages [11]. Dermatologists typically utilize visual examination to evaluate dermoscopic images, known as biomedical images [12]. It takes a lot of time, is labor-intensive, and is subject to operator bias. The reason for this is normal moles and skin infections can sometimes be so similar that a precise diagnosis might be challenging [13]. Several computer-aided diagnostic systems created to aid dermatologists identify skin cancer [14]. These methods not only get over the aforementioned problems but also increase the diagnosis system's impartiality, accuracy, and efficiency [15]. Deep learning approaches have established encouraging results, important promise in this area for data analysis and image processing [16]. Owing to its widespread use, distinctive characteristics in several complicated fields, likes object detection, classification, identification, and recognition, deep learning applied extensively [17]. The deep learning method gives method greater depth; changes input utilizing different functions enable hierarchical data representation across multiple levels of abstraction [18]. Because more sophisticated models are used, deep learning can quickly and effectively learn more difficult issues. The significant component of typical computer aided diagnostic systems is deep learning approaches like CNN and image processing techniques [19]. However, as processing cycle behind method learning with feature encoding is poorly implicit, utility of such computer-aided diagnostic systems by dermatologists, patients is still

questionable. Without a logical justification, the model prevents dermatologists from making informed decisions [20].

The problem involves in several existing methods, skin cancer at premature stage owing to challenges in analyzing dermoscopic images of skin lesions, which are often subtle and require a high level of expertise. Traditional visual inspection by dermatologists is time-consuming, subjective, and prone to error, making it hard to achieve consistent and reliable results. There is a need for automated systems that can assist in the accurate, efficient, and objective classification of skin lesions to enhance diagnostic accuracy, ultimately lessen skin cancer mortality.

This paper intends to overcome these issues to improve skin lesion detection methods, it is crucial to address the significant challenges that skin lesion. Enhancing the accuracy of identifying and classifying skin lesion detection is vital for facilitating timely and effective responses to these incidents. The suggested method attempts to improve detection accuracy and reliability by exploiting optimization using Red Billed Blue Magpie Optimization Algorithm.

The novelty of SLCDI-DGCRIN-RBBMOA method lies in its use of Skin Lesions Classification of Dermoscopic Images using Optimized Dynamic Graph Convolutional Recurrent Imputation Network. CPSF is used for preprocessing dermoscopic images, effectively removing noise and resizing them to improve image quality. The ROI region is segmented utilizing HDAETUNet3+, which improves the accuracy of lesion detection. The main innovation of the approach lies in the use of the DGCRIN for classifying a diverse range of skin lesions. The RBBMOA strategy significantly enhances accuracy, precision, recall, specificity, F1-score while also reducing error rate compared to existing methods, making it more suitable for classification on Skin Lesions in the future.

Major contribution of this investigate work is brief as below,

- To propose SLCDI-DGCRIN-RBBMOA framework, which employs an Optimized Dynamic Graph Convolutional Recurrent Imputation Network to enhance the Skin Lesions Classification of Dermoscopic Images process.
- Here, CPSF improves data integrity by efficiently for remove noise, resize and improve image quality within the dataset.
- To segment the ROI region using the Hybrid Dual Attention-guided Efficient Transformer and UNet 3+ (HDAETUNet3+) process and to classify Skin Lesion using DGCRIN, thereby improving classification accuracy.
- RBBMOA introduces an optimization approach to improve the weight parameters of the DGCRIN classifier.
- Performance comparison to analyze the efficiency of the SLCDI-DGCRIN-RBBMOA approach in comparison to well-known DGCRIN in context of Skin Lesion.

The proposed model addresses the limitations of previous approaches by enhancing transparency, accuracy, and optimization in skin lesion classification. To build trust in AI-based decisions, it incorporates Hybrid Dual Attention-guided Efficient Transformer and UNet 3+ (HDAETUNet3+), enabling precise segmentation and improved interpretability. Confidence Partitioning Sampling Filtering (CPSF) enhances image quality by eliminating noise, facilitating more accurate early-stage lesion classification and reducing diagnostic errors. The model leverages Transformer-based feature extraction and multi-scale segmentation to refine region of interest (ROI) identification, outperforming conventional deep learning methods. Furthermore, the Dynamic Graph Convolutional Recurrent Imputation Network (DGCRIN) captures spatial relationships within dermoscopic images, offering a structured and adaptive classification approach. Unlike previous graph-based models that lack effective parameter tuning, the Red Billed Blue Magpie Optimization Algorithm (RBBMOA) optimizes DGCRIN, significantly improving classification performance.

Remaining part of the paper is arranged as follows: Literature review is presented in Section II, Methodology employed is discussed in Section III, and result with discussion is described in Section IV and conclusion in Section V.

II. Literature Review

A. Related Work

Several investigate works are presented in literatures based on the Skin Lesion Classification utilizing Deep Learning. Table I presents various advantages and disadvantages of the existing Skin Lesion Segmentation and Classification model. In 2022, Nigar, N., et al., [21] have presented a Deep Learning approach based on Explainable Artificial Intelligence for skin lesion classification. A skin lesion classification system based on Explainable Artificial Intelligence is suggested to enhance the accuracy of skin lesion detection, aiding dermatologists in making more informed and rational diagnoses, particularly in the early stages of skin cancer. The system accurately identifies eight types of skin lesions: dermatofibroma, squamous cell carcinoma, benign keratosis, melanocytic nevus, vascular lesion, actinic keratosis, basal cell carcinoma, and melanoma. It attains high accuracy and low precision.

In 2024, Khan, M.A., et al., [22] have presented multiclass skin lesion classification utilizing deep learning networks optimal information fusion. A computerized method for multiclass skin lesion classification, leveraging a fusion of optimal deep learning model features is developed. The collection of data used in is unbalanced, thus mathematical operations are performed to address this problem through data augmentation. The augmented dataset is used to refine and train two pre-trained deep learning models, DarkNet-19 and MobileNet-V2. After training, features extracted from the average pooling layer are optimized using a hybrid firefly optimization technique. The selected features are then fused using both the threshold-based and serial approaches, and classified using machine learning classifiers. It attains higher recall and low specificity.

In 2023, SM, J., et al., [23] have presented classification of skin cancer from dermoscopic images using deep neural network architectures. A deep convolutional neural network (DCNN) model is developed to accurately classify melanoma and non-melanoma skin cancer. The datasets, sourced from various challenges, have issues like class imbalance and varying image resolutions. To address these, EfficientNet with transfer learning is used to capture more complex patterns by adjusting the network's depth, width, and resolution. The dataset is augmented, and metadata is incorporated to improve classification performance. Additionally, the EfficientNet model is optimized with the Ranger optimizer, reducing the need for extensive hyperparameter tuning. It provides high F1-score and high computational time in 2023, Alsahafi, Y.S., et al., [24] have presented Skin-Net, a novel deep residual network for skin lesions classification. It utilizes multilevel feature extraction and cross-channel correlation, along with outlier detection. In this suggested paper, the Residual Deep Convolutional Neural Network is designed with multiple convolutional filters for multi-layer feature extraction and cross-channel correlation, using sliding dot product filters instead of sliding filters along the horizontal axis. To address the problem of imbalanced datasets, the method transforms the dataset from image-label pairs to image-weight vectors. It has been tested and assessed on complex and demanding datasets and demonstrates superior performance compared to existing deep convolutional networks in multiclass skin lesion classification. It attains higher detection accuracy and low kappa coefficient.

TABLE I. STRENGTHS AND LIMITATIONS OF THE CURRENT MODELS FOR CLASSIFYING SKIN LESIONS

Authors Name	Methods	Advantage	Limitation
Nigar, N., et.al, [21]	Convolutional neural network, deep neural network	It achieves higher accuracy	It provides low precision
Khan, M.A., et.al, [22]	CNN, DarkNet-19, MobileNet-V2	It attains higher recall	It attains low specificity
SM, J., et.al, [23]	Deep neural network, DCNN, EfficientNet, DenseNet	It provides higher F1-score	It provides high computational time
Alsahafi, Y.S., et.al, [24]	Residual Deep CNN, DNN, Probabilistic neural network	It provides high detection accuracy	It provides low kappa coefficient
Raghavendra, P.V., et.al, [25]	Deep convolutional neural network, ResNet50, VGG-16, MobileNetV2, and DenseNet121	It attains higher RoC	It attain slow precision
Rezaee, K. et.al, [26]	Convolutional neural network, ResNet-50, and ResNet-101	It provides high specificity	It provides high loss function
Thanka, M.R., et.al, [27]	Generative adversarial network, VGG16 and XGBoost	It attains low error rate	It attains high sensitivity

In 2023, Raghavendra, P.V., et al., [25] have presented Deep Learning Based Skin Lesion Multi-class Classification with Global Average Pooling Improvement. The model is trained and tested on the dataset, which includes seven distinct classes of skin lesions. During the preprocessing stage, the black hat filtering technique is applied to remove artifacts, along with resampling techniques to address class imbalance. The performance of the proposed model is evaluated by comparing it with several transfer learning models, including ResNet50, VGG-16, MobileNetV2, and DenseNet121. It attains high RoC and low precision.

In 2024, Rezaee, K. et al., [26] have presented self-attention transformer based deep learning framework for skin lesions classification in smart healthcare. This approach fuses global and local features through cross-fusion to generate fine-grained features. The branches of the parallel systems are merged using feature-fusion architecture, creating a pattern that identifies the characteristics of various skin lesions. Additionally, the paper introduces an optimized and lightweight version of the CNN architecture, optResNet-18, designed to effectively discriminate between skin cancer lesions. It attains higher specificity and high loss function.

In 2023, Thanka, M.R., et al., [27] have presented hybrid technique for melanoma classification utilizing ensemble ML methods and deep transfer learning. A hybrid approach combining a pre-trained convolutional neural network and machine learning classifiers is employed for feature extraction and classification, enhancing the model's accuracy. VGG16 is used for feature extraction, while XGBoost serves as the classifier. This combination leverages the strengths of deep learning for feature extraction and the power of machine learning for efficient classification, leading to improved performance in skin lesion classification. It provides low error rate and low sensitivity.

B. Research Gap

Current skin lesion classification models suffer from a lack of transparency, reducing trust in AI-driven diagnoses. Early-stage lesion detection is challenging due to visual similarities. Existing methods struggle with accurate segmentation and feature extraction, while graph-based networks lack optimization. The proposed SLCDI-DGCRIN-RBBMOA improves classification through enhanced segmentation and advanced optimization strategies.

III. PROPOSED METHODOLOGY

The SLCDI-DGCRIN-RBBMOA methodology, which aims to classification of skin lesion, begins by input dermoscopic images are collected from ISIC-2019 skin disease dataset. The proposed block diagram of SLCDI-DGCRIN-RBBMOA is represented in Fig. 1. Input images are pre-processed with filtering to remove noise, improve image quality, followed by segmentation process, and are then fed into a classification process. The Dynamic Graph Convolutional Recurrent Imputation Network is then used to categorize the skin lesion as actinic keratosis, Dermatofibroma, Basal cell carcinoma, Squamous cell carcinoma, Melanocytic nevus, benign keratosis, Melanoma, vascular lesion. To improve classification accuracy, Red-Billed Blue Magpie Optimization Algorithm is utilized to enhance DGCRIN parameters. The overall system aims to detect, mitigate skin lesion, ensuring the integrity and reliability of image.

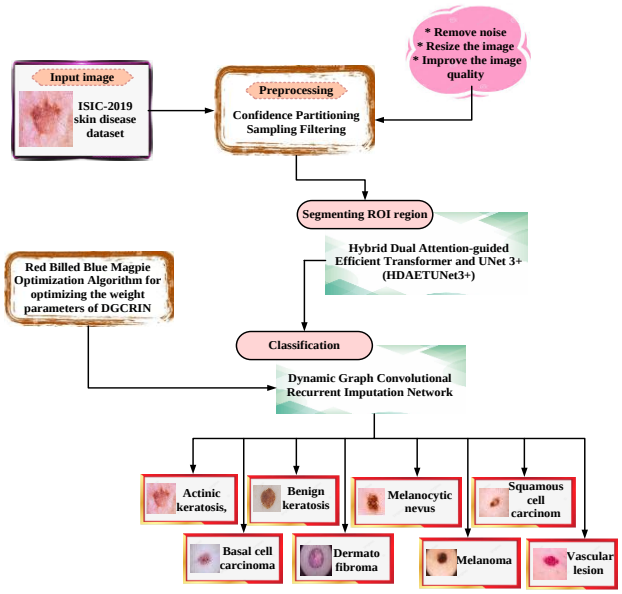


Fig. 1. Block diagram of SLCDI-DGCRIN-RBBMOA method.

A. Image Acquisition

The input data are gathered from ISIC-2019 Skin Disease Dataset [28]. This dataset comprises of 25,331 dermoscopic images, classified into eight types such as actinic keratosis, dermatofibroma, melanoma, benign keratosis, vascular lesion, basal cell carcinoma, melanocytic nevus, squamous cell carcinoma. The generated dataset is randomly splitted as 80% training, 10% testing, and 10% validation.

B. Pre-processing utilizing Confidence Partitioning Sampling Filtering

The image pre-processing utilizing CPSF [29] to eliminate noise, resize the image, and enhance image quality. CPSF methods were selected as a pre-processing technique for skin lesion analysis because they enhance image contrast, remove noise, and resize image, thereby improving the accuracy and clarity of lesion boundaries. This technique effectively reduces noise while preserving important features, enhancing the overall quality of segmentation. As a result, CPSF leads to more robust segmentation and better performance in subsequent analysis, providing more reliable results in dermoscopic images by reducing artifacts and noise. CPSF improves overall quality of images, enabling subsequent models, such as segmentation networks, to focus more effectively on relevant features and generate more accurate predictions. It also facilitates image resizing without sacrificing important details, which is particularly useful when working with datasets that require standardized input dimensions. Additionally, CPSF is a versatile tool that applied to different image types beyond dermoscopic images, making it valuable for medical image preprocessing as well as other computer vision applications. Then, the CPSF is given in Eq. (1),

$$w.k \int_{B_{q(y)}}^{\alpha} q(y)dx = 1 - \alpha \tag{1}$$

where, $q(y)$ represents the dimensional space, w denotes the hyper parameter, k represents the input data values, represents the filtering process, dx denotes the steady-state phase. Then, it is calculated as given in Eq. (2),

$$\omega_h = \frac{q(\hat{Y}_h)}{\sum_{h=1}^{H} q(\hat{Y}_h)} \tag{2}$$

Here, $q(\hat{Y}_h)$ represents the partitioning filtering of a distribution, ω_h denotes the bounded subspace, and H represents the sampling interval. It evaluates the correlations between features to predict quality images using variables that are strongly related.

Let qc represents the impulse function, y_t denotes the noise model, $\vec{y}_{t,c}$ represents the weighted grid samples, then image resize c is calculated using Eq. (3),

$$q(y_t | \hat{y}_{t-1,c}) = qc(y_t - \vec{y}_{t,c}) \tag{3}$$

It helps maintain the variation and structure of the original image, ensuring that the imputed values do not alter overall trends. In this step the denoise images are calculated as in Eq. (4),

$$\bar{y}_t = (\omega_t)^K \hat{y}_t \tag{4}$$

In the above equation, $\bar{y}_t$ represents the probability region, ω_t denotes the CPSF framework, $\hat{y}_t$ represents the denoise image. Then, the improved image quality is predicted as given in Eq. (5).

$$\vec{Y}_t = \left[\vec{y}_{t,1}, \vec{y}_{t,2}, ..., \vec{y}_{t,C_{t-1}}\right]^K \tag{5}$$

Here C_{t-1} represents the estimation of the prior CPSF, $\vec{y}_{t,1}$ represents the improved image quality, $\vec{y}_{t,2}$ denotes the local inference, $\vec{Y}_t$ denotes improved image quality. Here, by using CPSF method it remove noise, resize and improve image quality. After that, the pre-processed images undergoes segmentation phase.

C. Segmentation Using Hybrid Dual Attention-Guided Efficient Transformer and UNet 3+

Segmentation using HDAETUNet3+ [30, 31] method is to segment ROI region from dermoscopic images. The DAET and U-Net3+ is selected for its synergistic benefits in image segmentation. Dual attention mechanisms in DAET are highly effective at capturing long-range dependencies and fine-grained details, which are essential for precise segmentation. Meanwhile, Hierarchical structure and skip connections of U-Net3+ efficiently propagate contextual information across different scales, enhancing localization and boundary delineation. It enhances methods ability to exactly segment skin lesions, even with complex boundaries, by capturing both local and global features. This improves lesion classification and results in a more computationally efficient and robust method for skin lesion recognition in dermoscopic images.

Dual Attention-guided Efficient Transformer, which integrates spatial and channel attention, enables method to emphasis on the most applicable image regions and key features, resulting in accurate, robust segmentation of skin lesions. The architecture efficiently scales to learn large image sizes or long sequences, which is essential for tasks such as high-resolution medical image segmentation. By leveraging the attention mechanism, the model can adapt to a wide range of input data, including varying image sizes and resolutions, making it versatile for deployment across different applications. Additionally, the attention mechanism provides greater transparency into methods decision-making process, which is particularly valuable in high-stakes domains like medical imaging. Attention maps highlight areas the model prioritizes, offering insights into its focus during predictions and improving interpretability. It is given in Eq. (6).

$$R(P,K,U) = soft\max\left(\frac{PK^{S}}{\sqrt{c_k}}\right)U \tag{6}$$

Here, the term c is the embedding dimension, U is the value vector of the image. Conventional self-attention may produce redundant context matrix, effectual way to calculate self-attention mechanism is provided in Eq. (7).

$$F(P,K,U) = \rho_p(P)\left(\rho_k(K)^{S}U\right) \tag{7}$$

Here, ρ_p and ρ_k denotes normalization functions for queries, keys. These are softmax normalization functions. Hence, efficient attention first normalizes the keys and queries and then multiplies keys with values. The transpose attention is shown in Eq. (8).

$$MLP(Y) = FC\left(GELU\left(DW-Conv\left(FC(Y)\right)\right)\right) \tag{8}$$

In Eq. (8), MLP represents the activation function, FC is the fully connected layer, DW is the depth wise convolution. UNet 3+ improves lesion segmentation by leveraging nested skip pathways to capture features at multiple scales, which enhances its ability to identify lesions of different sizes and shapes. It ensures that intermediate layers actively contribute to the learning process, accelerating convergence and reducing the risk of overfitting, particularly when working with limited medical image data. Its capacity to extract fine-grained details while preserving global context makes the model more robust to noise, artifacts commonly found in dermoscopic images, lead to more reliable segmentation in real-world clinical scenarios. It is given in Eq. (9),

$$A_{j,i} = \frac{\exp(N_j \cdot M_i)}{\sum_{j=1}^{m}\exp(N_j \cdot M_i)} \tag{9}$$

where, $A_{j,i}$ is measures impact of i^{th} location on j^{th} location, m is the number of pixel values. The architecture diagram of HDAETUNet3+ is represented in Fig. 2.

The HDAETUNet3+ architecture is designed for dermoscopic image segmentation, focusing on accurately segmenting regions of interest in dermoscopic images. It features a dual-transformer block structure, with skip connections and self-supervised contrastive learning to improve feature representation. The encoder extracts features progressively through patch merging and dual-transformer blocks, while the decoder upsamples and refines the segmentation mask. The network also incorporates full-scale intra- and inter-skip connections and ground truth supervision to guide the learning process. This sophisticated design aims to provide precise and reliable segmentation of ROIs in dermoscopic images, supporting the diagnosis of skin conditions. Then the each position of the image is given in Eq. (10).

$$GSA(N,M,W)_q = \sum_{p=1}^{h\times w}\left(W_p B_{q,p}\right) \tag{10}$$

Here $GSA(N,M,W)_q$ are features to generate the output image Finally, HDAETUNet3+ is used to segment the ROI region. Segmentation output is given into classification phase.

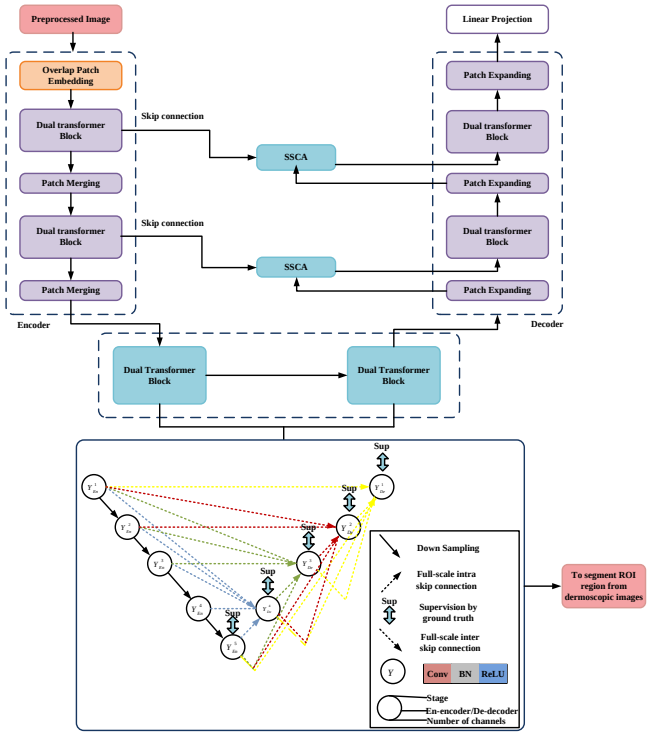


Fig. 2. Architecture of HDAETUNet3+.

D. Classification utilizing Dynamic Graph Convolutional Recurrent Imputation Network

The classification using DGCRIN is categorizing skin lesion such as actinic keratosis, Dermatofibroma, Basal cell carcinoma, benign keratosis, vascular lesion, Squamous cell carcinoma Melanocytic nevus, Melanoma [32]. The architecture diagram of DGCRIN is represented in Fig. 3. The DGCRIN captures the evolving relationships between image points over time, enabling it to adapt to changes in the image structure. The recurrent component adds temporal context, helping the model learn sequential patterns and enhance prediction accuracy, particularly when dealing with incomplete image. By incorporating imputation techniques, the model can effectively manage missing, leading to more robust and

accurate classification results. DGCRIN utilizes Graph Convolutional Networks to capture spatial dependencies within graph-structured image, making it particularly effective for datasets where the relationships between nodes are crucial for understanding missing data patterns. By incorporating both spatial and temporal information, DGCRIN can learn richer, more comprehensive representations of the image. Its ability to adapt to both spatial and temporal complexities enables it to handle intricate image patterns, making it especially well-suited for imputation tasks where the connections between missing values are complex and non-linear. The architecture diagram illustrates the DGCRIN model, where the input to the process is a set of segmented images. The model consists of three main components: a masked GRU, a DGCGRU, and a graph generator. The graph generator dynamically creates a graph at every time step to denote the geographical correlations of the road network, using both historical data and the current imputed data. The DGCGRU effectively captures the spatiotemporal relationships in the data by integrating the dynamic graph with a static graph and replacing the fully connected layers in a conventional GRU with a dynamic graph convolution operation. Additionally, a masked GRU is employed to independently analyze the masking matrix and identify patterns in the missing data. The information is then combined by a fusion layer using a temporal decay mechanism, and data inference is performed by a fully connected layer. This iterative process helps the model achieve high accuracy in classifying skin lesions. Thus, network layers are part of the recurrent imputation network model calculated as given in Eq. (11),

$$G_{E,t} = \tanh\left(V_E\left(G_{A,t}\Theta\tilde{G}_{m,t}\right) + y_E\right) \tag{11}$$

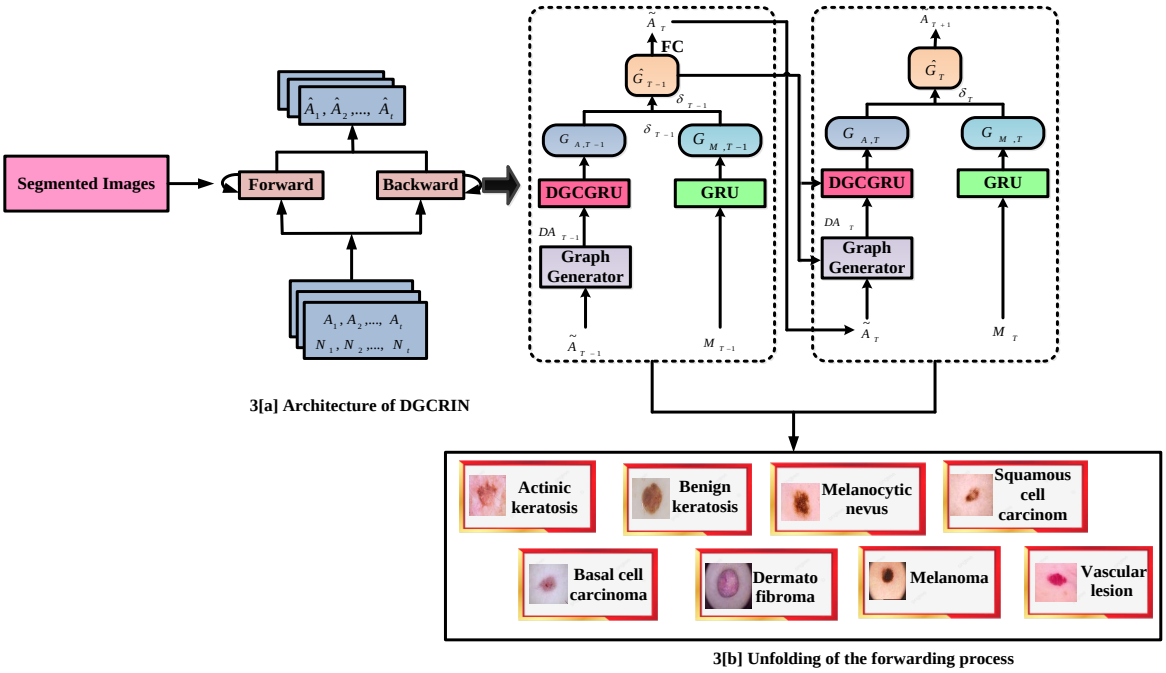


Fig. 3. (a) Architecture diagram of DGCRIN and (b) Unfolding of the forward process.

where, indicates the observation of each features, represents the significance of observation, indicates the hidden state, indicates the exponential function, indicates the sigmoid multiplication, indicates the constantly updated value, and represents the squared random initialization value. To classify the actinic keratosis, Basal cell carcinoma can be expressed as given in Eq. (12).

$$\lambda_t = \frac{1}{f^{\max}\left(0, V_\lambda \delta_t + y_\lambda\right)} \tag{12}$$

Here λ_t represents the efficient reconstruction value, δ_t represents the learnable variable, V_λ represents the forward along backward feature matrices, y_λ represents the activation functions and $f^{\max}$ depicts at each time step. To classify the benign keratosis, dermatofibroma, and melanocytic nevusis, the calculation is performed as given in Eq. (13).

$$A_{T+1} = V_A \hat{G}_t + y_A \tag{13}$$

In Eq. (13), the term B_{T+1} represent the estimated value, V_A denotes the scoring vector, $\hat{G}_t$ is the dynamic adjacency matrix is regularized by the activation function and y_A indicates the learnable parameter. To classify the Melanoma, Squamous cell carcinoma, vascular lesion is calculated as given in Eq. (14),

$$A_{T+1} = A_{T+1}\Theta N_{T+1} + \tilde{A}_{T+1}\Theta\left(1 - N_{T+1}\right) \tag{14}$$

where, A_{T+1} represents the prediction value, $\tilde{A}_{T+1}$ represents the image consumption, Θ indicates the sigmoid multiplication operator, and N_{T+1} denotes the SCC. Finally, the DGCRIN classified the skin lesion likes SCC, AK, BCC, Dermatofibroma, Melanoma, vascular lesion, Melanocytic nevus, benign keratosis. The RBBMOA is utilized to optimize

DGCRIN optimal parameter y_a and δ_t. The RBBMOA is used for tuning DGCRIN weight, bias parameter.

E. Optimization using Red Billed Blue Magpie Optimization Algorithm

The weights parameter y_a and δ_t of proposed DGCRIN is optimized utilizing the proposed Red Billed Blue Magpie Optimization Algorithm (RBBMOA) [33].This optimizer works efficiently and quickly, converging to the optimal weight parameters in a shorter time than other optimization techniques. It optimizes network parameters, leading to improved model performance and higher classification accuracy for skin lesion detection. The RBBMOA refines the standard algorithm by introducing a more effective, reducing the risk of getting trapped in local minima. It also quicker convergence and improved solution accuracy. Furthermore, RBBMOA adjusts dynamically to complex problem environments, enhancing its efficiency across various optimization tasks. As foraging, it uses mix of ground-walking, jumping, searching along branches to find food. It also stores food for later consumption. To protect its cached items from theft by other an also there wise birds; it hides food in secure locations such as tree hollows, forks, and rock crevices. Overall, it is versatile predator, employing various strategies to acquire and store food. It also displays social behavior and cooperation when hunting. The flowchart illustrating the proposed RBBMOA approach is presented in Fig. 4.

1) Stepwise procedures for RBBMOA: The step by step technique is defined to obtain optimal value of DGCRIN dependent on RBBMOA. At first, RBBMOA makes evenly allocating populace to enhance parameter of DGCRIN.

Step 1: Initialization

The initialization phase of RBBMOA applicant solutions is created smoothly within the constraints of a given problem, requiring updates after iteration. It is given in Eq. (15),

$$Y = \begin{bmatrix} y_{1,1} & \cdots & y_{1i} & y_{1,\dim-1} & y_{1,\dim} \\ y_{2,1} & \cdots & y_{2,i} & \cdots & y_{2,\dim} \\ \cdots & \cdots & y_{j,i} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ y_{m-1,1} & \cdots & y_{m-1,i} & \cdots & y_{m-1,\dim} \\ y_{m,1} & \cdots & y_{m,i} & y_{m,\dim-1} & y_{m,\dim} \end{bmatrix} \quad (15)$$

where, Y signifies location of search agent, $\dim$ denotes dimension of solving problem, m represents population size.

Step 2: Random Generation

Input parameters are generated randomly. Optimal fitness values are preferred depending upon explicit hyper parameter situation.

Step 3: Fitness function

A random solution is generated utilizing initialized assessments using factor optimization value; it is evaluated for optimizing weight parameter y_a and δ_t of skin lesion. It is given in Eq. (16),

$$Fitness\ Function = optimize\ (y_a\ \&\ \delta_t) \quad (16)$$

where, y_a is utilized to increasing accuracy, δ_t is utilized to decreasing error rate.

Step 4: Search for food

To enhance efficiency, red-billed blue magpies typically hunt in small groups. It uses various techniques, including walking, jumping on ground, scouring trees for food resources. Billed blue magpies use adaptable hunting strategies that depend on environmental conditions, resources at hand, ensure adequate food supply as shown in Eq. (17).

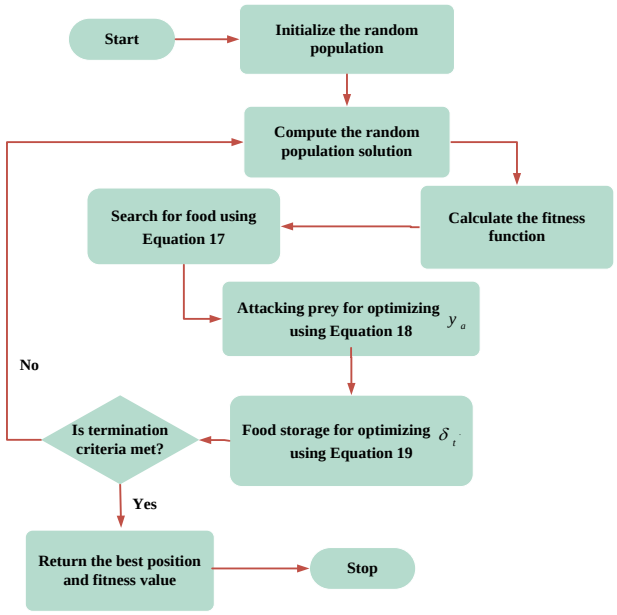


Fig. 4. Flowchart of RBBMOA for optimizing DGCRIN parameter.

$$Y^j(l+1) = Y^j(l) + \left(\frac{1}{q} \times \sum_{n=1}^{q} Y^n(l) - Y^{sa}(l)\right) \times Rand_2 \quad (17)$$

where, $Y^j(l+1)$ represents the prey attackers, Y^j represents the normal distribution, l represents the hunting behaviour, q indicates red-billed blue magpies typically hunt within small groups, Y^n represents the tactics, $Y^{sa}(l)$ denotes randomly selected search agents for current iteration, $Randn_2$ denotes random number 2.

Step 5: Attacking prey for optimizing y_a

When pursuing prey, red-billed blue magpie demonstrates higher degree of hunting skill, teamwork. It uses fast picking, jumping, and flying to grab insects, among other strategies. Typically, small prey or plants are the main focus of small group operations. Red-billed blue magpies can work together to pursue larger prey, such as tiny animals or large insects, when they are in groups. It is expressed as Eq. (18).

$$Y^j(l+1) = Y^{food}(l) + DG \times y_a \left(\frac{1}{q} \times \sum_{n=1}^{q} Y^n(l) - Y^j(l) \right) \times Randn_1 \quad (18)$$

where, Y^{food} represents the position of the food, DG represents the diverse strategies, and $Randn_1$ denotes random number 1 and y_a is used to increase the accuracy.

Step 6: Food storage for optimizing δ_t

Red-billed blue magpies not only hunt and fight prey, but it also stockpile extra food in tree holes and other hidden places so they may eat it later, providing a consistent source of food even in times of famine. This procedure saves information about the solutions, making it easier for people to find the value that is globally optimal. It is given in Eq. (19).

$$Y^j(l+1) = \delta_t \begin{cases} Y^j(l) & if \ \ fitness^j_{old} > fitness^j_{new} \\ Y^j(l+1) & else \end{cases} \quad (19)$$

where, $fitness^j_{old}$ and $fitness^j_{new}$ denotes fitness values earlier than, following position modernize of j^{th} red-billed blue magpie, δ_t is used for decreasing the error rate.

Step 7: Termination

The weight parameter value of creator y_a and δ_t from DGCRIN is improved by RBBMOA; it repeat step 3 until it acquires its halting criteria $Y = Y + 1$. Then, the SLCDI-DGCRIN-RBBMOA methods effectively classify the skin lesion by higher accuracy and low error rate.

2) Complexity analysis: The RBBMOA algorithm is based on two key components: solution initialization with the primary approach functions, which include computing the fitness values and update the solutions. Let n represent the cout of search agents, T denote the maximal count of iterations, and $\dim$ refer to the problem's dimension. Finding the optimal location and updating the location of every solutions are included in the solution updating process, which has a computational difficulty of $O(T \times n) + O(2 \times T \times n \times \dim)$, while the solution initialization process has a computational complexity of $O(n)$. As a result, the overall computational complexity of the proposed RBBMOA algorithm is $O(n \times T \times (2 \times \dim + 1))$.

IV. Result and Discussion

The experimental results of the proposed SLCDI-DGCRIN-RBBMOA approach is implemented in Python utilizing PC along Intel Core i5, 16GB RAM, 3.2 GHz CPU, Windows7, examined utilizing various performance measures likes accuracy, specificity, recall, precision, F1-score, computation time, error rate. The SLCDI-DGCRIN-RBBMOA model is tested against several performance metrics. Attained result of SLCDI-DGCRIN-RBBMOA approach is compared with existing methods likes DNN-EAI-SLC, MSLC-CNN-OIF, CSC-DI-DCNN respectively. Fig. 5 depicts the skin lesion classification workflow across different dermoscopic images, from segmented images, raw input images and pre-processed. Every modality captures exclusive anatomical details, which give to exact skin lesion detection.

A. Performance Measures

Performance metric accuracy, recall, specificity, F1-score, precision, computational time and error rate are examined for performance matrices. To measure performance metrics, performance matrix is required. Next matrixes are necessary to measure performance metrics.

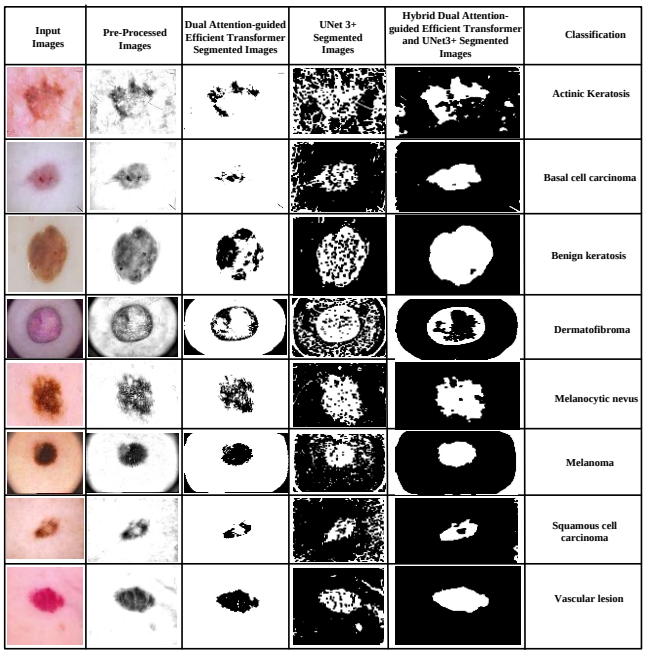


Fig. 5. Output of the proposed SLCDI-DGCRIN-RBBMOA method.

- True Positives: The number of actual positive cases that are accurately categorized as positive.
- True Negatives: The number of actual negative cases that are accurately categorized as negative.
- False Positives: The number of actual negative cases that are inaccurately classified as positive.
- False Negatives: The number of actual negative cases that are inaccurately classified as negative.

1) Accuracy: It is known as percentage of correctly identified cases among total instances and it is evaluated by Eq. (20),

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (20)$$

2) Precision: It assess the capacity of model for recognize positive instances accurately out of all predicted instances cases, which is estimated by Eq. (21),

$$\Pr ecision = \frac{TP}{(TP + FP)} \quad (21)$$

3) Recall: It is measured through separating total count of elements in positive class with count of real positives and determined by Eq. (22),

$$\text{Re}call = \frac{TP}{(TP + FN)} \tag{22}$$

4) Specificity: The percentage of true negatives that approach exactly recognizes is known as specificity, it is exhibits in Eq. (23),

$$Specificity = \frac{TN}{TN + FP} \tag{23}$$

5) F1-score: It represents ensemble mean of precision, recall. It is calculated using Eq. (24),

$$F1 - score = 2 * \frac{precision * recall}{precision + recall} \tag{24}$$

6) Error rate: It is a statistic used to express the prediction inaccuracy of the methodology depending on the actual approach. This is scaled in Eq. (25),

$$Error\ \ Rate = 100 - Accuracy \tag{25}$$

B. Performance Measures

Fig. 6-14 shows simulation result of SLCDI-DGCRIN-RBBMOA method. The performance measures are analyzed with existing DNN-EAI-SLC, MSLC-CNN-OIF, and CSC-DI-DCNN methods. The analysis of accuracy performance is portrayed in Fig. 6. The graph-based design's superior ability to capture local interactions between features enhances prediction accuracy and minimizes misclassifications. Furthermore, modern optimization techniques refine the models tuning, resulting in improved accuracy in identifying actual cases of skin lesion. Here, SLCDI-DGCRIN-RBBMOA method attains 21.51%, 12.38%, and 24.61% higher accuracy for actinic keratosis; 16.26%, 14.05%, 19.51% greater accuracy for Basal cell carcinoma; 26.21%, 20.65%, 22.31% greater accuracy for benign keratosis; 41.79%, 20.25%, 15.85% greater accuracy for Dermatofibroma; 25.21%, 30.65%, 20.45% greater accuracy for Melanocytic nevus; 10.56%, 27.56%, 19.67% higher accuracy for Melanoma; 18.78%, 30.45%, and 29.67% higher accuracy for Squamous cell carcinoma; 28.78% higher accuracy for vascular lesion; analyzed with existing techniques such as DNN-EAI-SLC, MSLC-CNN-OIF, and CSC-DI-DCNN respectively.

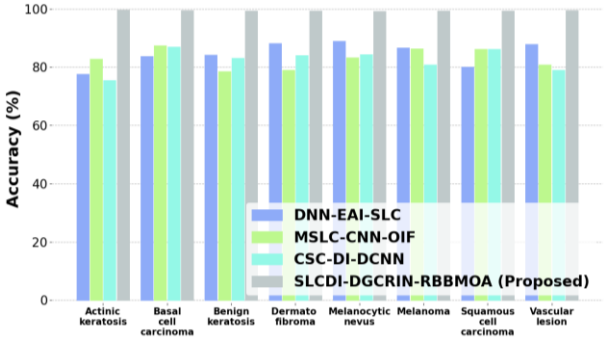


Fig. 6. Analysis of accuracy performance.

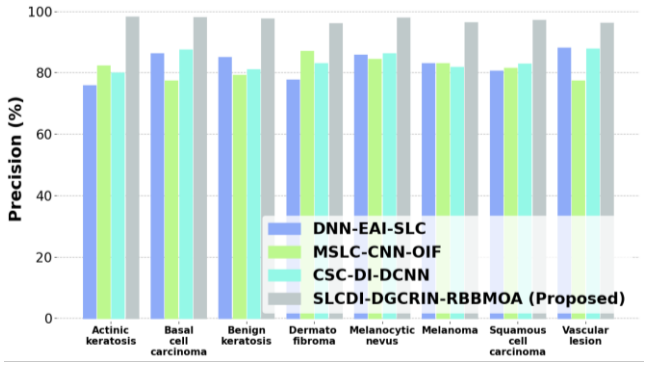


Fig. 7. Analysis of precision performance.

When compared to existing methods, the graph in Fig. 7 shows that SLCDI-DGCRIN-RBBMOA achieves higher precision across most skin lesion types due to model parameter optimization using the RBBMOA. This improvement in precision results from methods ability to better distinguish among different types of skin lesion. The enhanced accuracy is reflected in the methods overall performance, demonstrating its efficiency in managing and mitigating skin lesion. Here, SLCDI-DGCRIN-RBBMOA method attains 28%, 50%, and 21.51% higher Precision for actinic keratosis; 41.79%, 20.25%, 15.85% higher Precision for Basal cell carcinoma; 20.21%, 30.65%, and 28.31% higher precision for benign keratosis; 30.78%, 28.45%, and 29.56% higher precision for Dermatofibroma; 10.56%, 18.2%, and 16.46% higher precision for Melanocytic nevus; 29.59%, 30.89%, and 38.56% higher precision for Melanoma; 30.89%, 29.35%, and 29.89% higher precision for Squamous cell carcinoma; 29.50%, 26.40%, 30.67% greater precision for vascular lesion; analyzed with existing techniques likes DNN-EAI-SLC, MSLC-CNN-OIF, and CSC-DI-DCNN correspondingly.

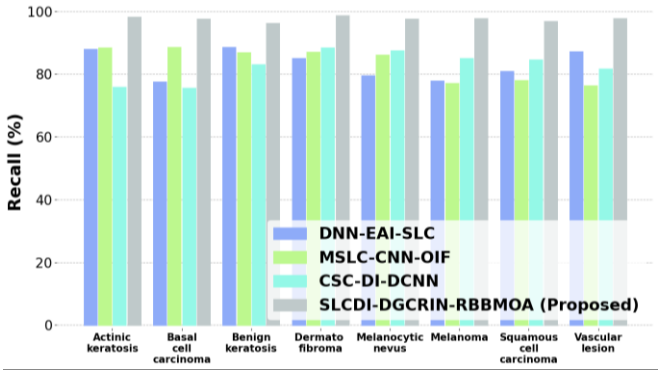


Fig. 8. Analysis of recall performance.

The model's high recall in SLCDI-DGCRIN-RBBMOA is attributed to its ability to effectively identify positive cases which is shown in fig. 8. By capturing intricate relationships between features, DGCRIN ensures that cases indicators are not overlooked. Additionally, RBBMOA optimizes the model's parameters to enhance sensitivity and increase the number of true positives identified. This combination strengthens the model's robustness against false negatives, reducing the number of missed examples and thereby improving recall. The SLCDI-DGCRIN-RBBMOA method attains 23.07%, 41.17%, and 24.67% higher recall for actinic keratosis; 43.47%,

25.31%, 16.47% higher recall for Basal cell carcinoma; 26.21%, 20.65%, and 22.31% higher recall for benign keratosis; 20.78%, 30.78%, and 12.56% higher recall for Dermatofibroma; 20.56%, 30.2%, and 28.46% higher recall for Melanocytic nevus; 30.29%, 35.89%, and 20.56% higher recall for Melanoma; 15.89%, 17.35%, and 29.46% higher recall for Squamous cell carcinoma; 30.26%, 20.46%, and 39.67% higher recall for vascular lesion; analyzed with existing techniques likes DNN-EAI-SLC, MSLC-CNN-OIF, and CSC-DI-DCNN.

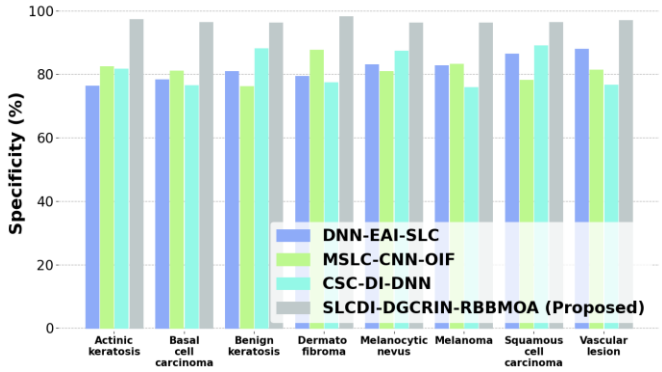


Fig. 9. Analysis of specificity performance.

Fig. 9 portrays the analysis of specificity performance. The specificity curve of a well-performing model will be lower at first and progressively higher as training goes on. The SLCDI-DGCRIN-RBBMOA method attains 17.75%, 24.55% and 12.66% high specificity for actinic keratosis; 21.01%, 14.08% and 17.33% high specificity for Basal cell carcinoma; 16.41%, 29.65%, and 24.31% higher specificity for Benign keratosis; 20.41%, 30.25%, and 14.83% higher specificity for Dermatofibroma; 29.41%, 30.48%, and 20.40% higher specificity for Melanocytic nevus; 17.01%, 29.08% and 30.67% higher specificity for Melanoma; 10.40%, 28.30% and 20.33% higher specificity for Squamous cell carcinoma; 29.59%, 20.9%, 25.31% greater specificity for Vascular lesion, analyzed with existing DNN-EAI-SLC, MSLC-CNN-OIF, and CSC-DI-DCNN models.

A model is considered to be comprehensive in recognizing all relevant cases and precise in forecasting positive cases if it's F1-score is higher. A high F1-score for the meningioma would propose that the system successfully detects critical circumstances without generating an excessive count of false alarms in context of skin lesion classification. Fig. 10 displays the analysis of F1-score performance. The SLCDI- DGCRIN-RBBMOA method attains 10.30%, 17.10% and 30.26% higher F1-score for the actinic keratosis; 28.02%, 11.56% and 13.67%.

F1-score for Basal cell carcinoma; 20.15%, 19.55%, and 12.21% high F1-score for Benign keratosis; 28.67%, 30.67%, and 23.67% high F1-score Dermatofibroma; 30.19%, 20.78%, and 20.56% higher F1-score for Melanocytic nevus; 29.68%, 10.56%, 26.67% greater F1-score for Melanoma; 29.78%, 10.56%, 30.56% greater F1-score for Squamous cell carcinoma; 20.67%, 17.49%, and 30.92% larger F1-score for Vascular lesion; compared with existing DNN-EAI-SLC, MSLC-CNN-OIF, and CSC-DI-DCNN models.

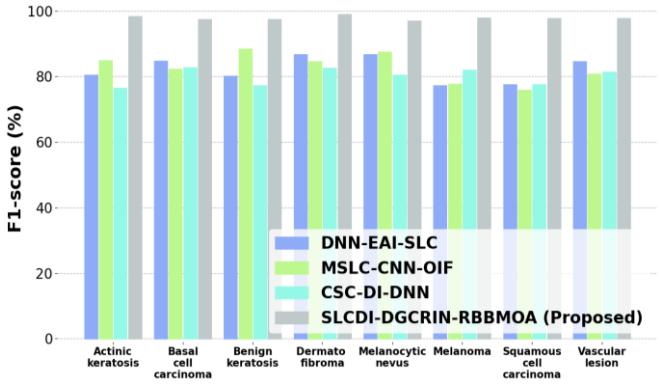


Fig. 10. Analysis of F1-score

Fig. 11 shows the analysis of computational time performance. It is faster due to optimized training methods, which include advanced optimization algorithms and effective hyperparameter tuning, reducing the number of training iterations required. Additionally, the spatio-temporal nature of DGCRIN allows for the parallel handling of spatial and temporal features, further enhancing computational efficiency. The proposed SLCDI-DGCRIN-RBBMOA method attains 15.01%, 13.44%, 14.27% lower computational time; analyzed with existing DNN-EAI-SLC, MSLC-CNN-OIF, and CSC-DI-DCNN methods.

A decreased error rate suggests improved model performance. Fig. 12 shows performance of error rate analysis.

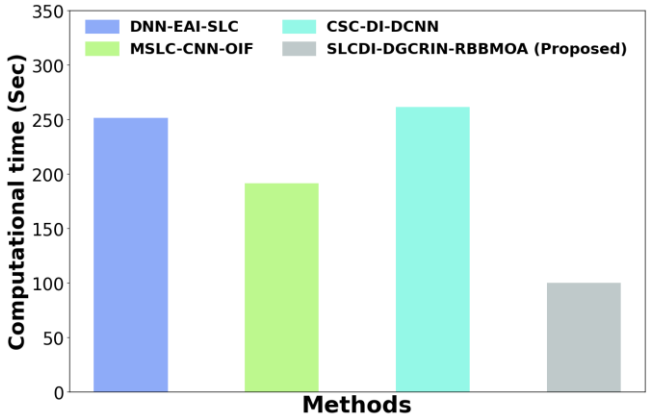


Fig. 11. Analysis of computational time performance.

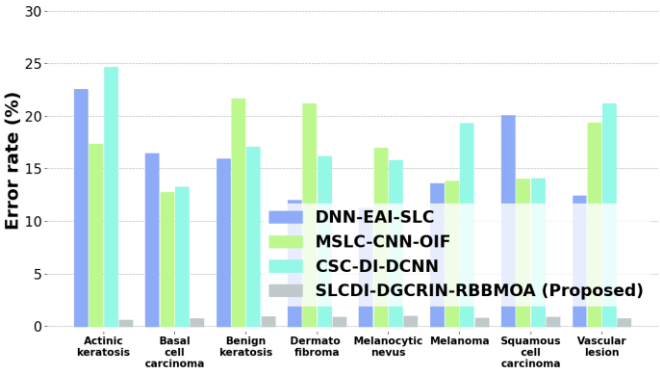


Fig. 12. Error rate analysis.

SLCDI-DGCRIN-RBBMOA has a significantly lower error rate than these existing methods, it indicates that the optimization techniques and advanced model architecture used in DGCRIN effectively reduce misclassifications, improving overall accuracy in identifying various skin lesion types. This reduction in error rates illustrates the new model's capacity to improve skin lesion classification reliability. Here, SLCDI-DGCRIN-RBBMOA method attains15.85%, 23.37%, and 22.04% lower Error rate for actinic keratosis; 25.97%, 20.57%, 19.23% lower Error rate for Basal cell carcinoma;24.78%, 22.67%, and 30.46% lower Error rate for benign keratosis; 30.79%, 10.29%, 25.59% lower error rate for Dermatofibroma; 20.45%, 25.89%, and 30.57% lower error rate for Melanocytic nevus; 19.67%, 30.16%, and 20.45% lower error rate for Melanoma; 20.67%, 30.15%, and 30.10% lower error rate for Squamous cell carcinoma; 15.79%, 20.56%, and 12.18% low error rate for vascular lesion; analyzed with existing approaches such as DNN-EAI-SLC, MSLC-CNN-OIF, and CSC-DI-DCNN correspondingly.

The Computational Complexity of the proposed SLCDI-DGCRIN-RBBMOA approach increases with increasing input size, indicating its scalability and appropriateness for bigger datasets. This graph in Fig. 13 highlights how much faster the SLCDI-DGCRIN-RBBMOA to the existing techniques likes DNN-EAI-SLC, MSLC-CNN- OIF and CSC-DI-DCNN respectively. Here, when the input size increases then the CPU operation decreases gradually. Fig. 14 depicts the confusion matrix.

C. Comparative Analysis of Proposed Approach

Table II presents a segmentation comparison with other methods depending upon accuracy, precision, recall, and F1-score. ResNet 50 shows an accuracy of 70.7%, with a strong recall of 86.6%, but a lower precision of 76.8%. DarkNet19 has higher accuracy but struggles with lower precision and recall. EfficientNet offers a well-balanced performance, achieving 79.5% accuracy, high precision, and moderate recall. Both the Dual Swin Transformer and Dual Vision Transformer perform well in recall; but both models have lower precision. The Dual Attention-guided Efficient Transformer shows good precision but lower recall. UNet3+ delivers strong results across all metrics, with an accuracy of 84.6%. The proposed Hybrid Dual Attention-guided Efficient Transformer and UNet 3+ model stands out with exceptional results, achieving 99.4% accuracy, 92.9% precision, 94.5% recall, and 95.4% F1-score, making it the most effective model in segmentation.

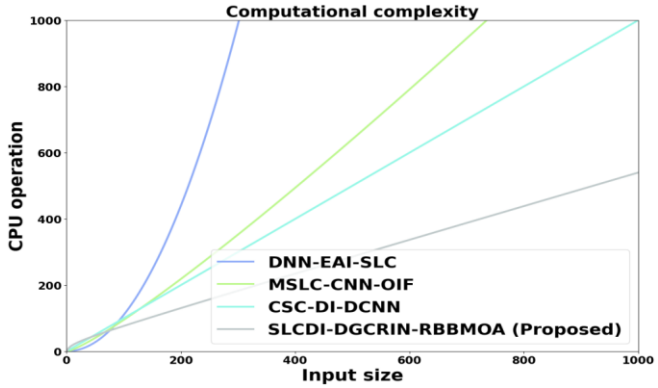


Fig. 13. Computational complexity performance analysis.

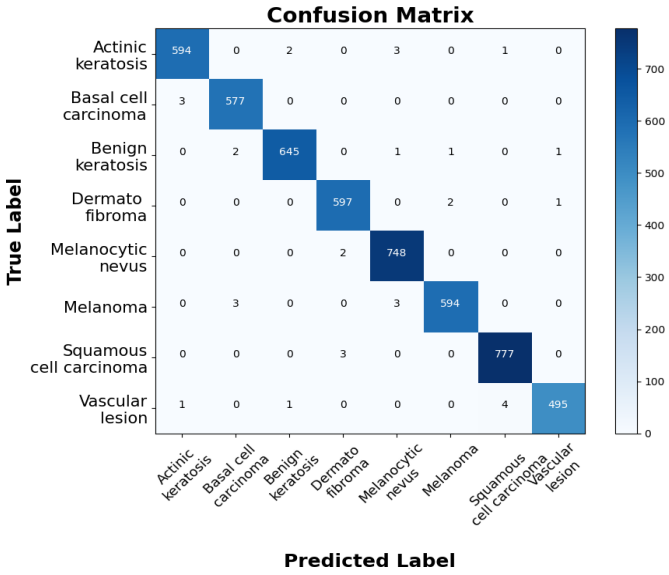


Fig. 14. Confusion matrix analysis.

TABLE II. SEGMENTATION COMPARISON WITH OTHER MODELS

Models	Accuracy	Precision	Recall	F1-score
ResNet 18 [21]	70.7%	76.8%	86.6%	80.3%
DarkNet19 [22]	75.4%	70.1%	67.3%	82.4%
EfficientNet [23]	79.5%	87.9%	72.6%	80.7%
Dual Swin Transformer [34]	76.4%	70.3%	74.9%	79.4%
Dual Vision Transformer [35]	80.2%	76.9%	85.2%	71.5%
Dual Attention-guided Efficient Transformer [30]	79.7%	80.1%	72.9%	67.8%
UNet3+ [31]	84.6%	79.9%	83.2%	80.6%
Hybrid Dual Attention-guided Efficient Transformer and UNet 3+ (Proposed) [30, 31]	99.4%	92.9%	94.5%	95.4%

TABLE III. COMPARATIVE ANALYSIS OF PROPOSED APPROACH

Methods	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)	Computational time (sec)	Error rate (%)
DNN-EAI-SLC [21]	77.47	75.69	87.80	76.22	80.37	250	22.52
MSLC-CNN-OIF [22]	82.70	82.20	88.26	82.30	84.82	190	17.29
CSC-DI-DCNN [23]	75.35	79.85	75.75	81.55	76.45	260	24.64
RDCNN-SLC-MFCC [24]	83.59	86.15	77.48	78.22	84.64	240	16.40
DCNN-SLMC-GAPI [25]	86.32	77.32	88.46	80.94	82.15	150	12.72
CNN-SLC-SH [26]	86.76	87.32	75.75	76.40	82.58	180	13.23
GAN-HMC [27]	84.08	85.20	88.45	80.75	80.01	230	15.91
SLCDI-DGCRIN-RBBMOA (proposed)	99.16	98.16	98.11	97.20	98.27	99	0.592

Table III provides a comparative analysis of SLCDI-DGCRIN-RBBMOA method alongside several other existing methods based on key performance metrics, comprising specificity, accuracy, precision, recall, F1-score, computational time, and error rate. Among the methods evaluated, the proposed SLCDI-DGCRIN-RBBMOA approach outperforms the others, achieving the highest accuracy and strong results across each metrics, like specificity, precision, recall, and F1-score. Additionally, it boasts the lowest error rate and computational time compared to the other methods. In contrast, approaches such as DNN-EAI-SLC, CSC-DI-DCNN, and RDCNN-SLC-MFCC exhibit lower accuracy and higher error rates, emphasizing the superior efficiency and effectiveness of the proposed model. This comparison emphasizes the exceptional performance with effectiveness of the SLCDI-DGCRIN-RBBMOA approach.

D. Discussion

The proposed model for Skin Lesions Classification in Dermoscopic Images employs an Optimized Dynamic Graph Convolutional Recurrent Imputation Network (DGCRIN) to enhance classification performance. The methodology begins with preprocessing, utilizing Confidence Partitioning Sampling Filtering (CPSF) to remove noise, resize images, and enhance quality. This aligns with [29], who demonstrated that CPSF significantly improves feature extraction in medical imaging by preserving essential details while eliminating distortions. Following preprocessing, the Hybrid Dual Attention-guided Efficient Transformer and UNet3+ (HDAETUNet3+) is applied for segmentation. The efficiency of hybrid transformer-based segmentation models in medical imaging has been well-documented. The study in [30] proposed Dae-former, a dual attention-guided efficient transformer, demonstrating superior segmentation accuracy for medical images, which supports the effectiveness of HDAETUNet3+ in identifying precise lesion boundaries. Additionally, the integration of UNet3+ [31], which employs full-scale connectivity, further enhances segmentation performance, ensuring robust ROI extraction from dermoscopic images. For classification, Dynamic Graph Convolutional Recurrent Imputation Network (DGCRIN) is utilized. Research on graph convolutional networks in handling complex spatial dependencies supports the use of DGCRIN. The research in [32] demonstrated the effectiveness of dynamic graph convolutional networks in managing spatiotemporal dependencies, which is crucial for accurately classifying skin lesions with varying patterns and structures. Despite its advantages, the proposed model presents some limitations. One challenge is the computational demand of deep learning and optimization algorithms. As noted by study [33], optimization algorithms like the Red-Billed Blue Magpie Optimizer (RBBMOA) enhance accuracy but may require extensive computational resources, which could be a concern for real-time medical applications. Additionally, the model's complexity and fine-tuning requirements may pose challenges for clinical integration, a concern also raised in prior studies on deep learning-based skin lesion classification [2].

V. CONCLUSION

The SLCDI-DGCRIN-RBBMOA model achieves significant advancements in skin lesion detection and classification. Utilizing dermoscopic images from the ISIC-2019 dataset, the model incorporates CPSF for noise reduction, resizing, and image enhancement. The HDAETUNet3+ effectively segments the ROI, while DGCRIN classifies lesions, and RBBMOA optimizes DGCRIN, enhancing classification accuracy. This approach demonstrates superior performance, achieving 21.51%, 12.38%, and 21.51% higher accuracy, along with 15.85%, 23.37%, and 22.04% lower error rates compared to DNN-EAI-SLC, MSLC-CNN-OIF, and CSC-DI-DCNN, respectively. However, challenges such as image quality variability and overlapping lesion features remain areas for further exploration. Future research could explore applying skin lesion classification techniques, with a focus on addressing disputes like data variability, noise in medical imaging, and the need for real-time diagnosis. Incorporating advanced feature extraction techniques could enhance lesion detection accuracy and improve classification performance. Additionally, optimizing deep learning models for domain-specific tasks could lead to more reliable skin cancer detection and prognosis. Moreover, incorporating multi-modal data, such as clinical metadata (e.g., patient history and genetic factors), alongside dermoscopic images could improve diagnostic precision. Lastly, future research could focus on edge and mobile deployment, adapting the model for lightweight, resource-efficient implementations, making it accessible in remote and resource-limited areas.

REFERENCES

[1] Hosny, K.M., Said, W., Elmezain, M. and Kassem, M.A., 2024. Explainable deep inherent learning for multi-classes skin lesion classification. Applied Soft Computing, 159, p.111624.

[2] Khater, T., Ansari, S., Mahmoud, S., Hussain, A. and Tawfik, H., 2023. Skin cancer classification using explainable artificial intelligence on pre-

extracted image features. *Intelligent Systems with Applications*, *20*, p.200275.

[3] Hosny, K.M., Said, W., Elmezain, M. and Kassem, M.A., 2024. Explainable deep inherent learning for multi-classes skin lesion classification. *Applied Soft Computing*, *159*, p.111624.

[4] Jasil, S.G. and Ulagamuthalvi, V., 2023. A hybrid CNN architecture for skin lesion classification using deep learning. *Soft Computing*, pp.1-10.

[5] Maqsood, S. and Damaševičius, R., 2023. Multiclass skin lesion localization and classification using deep learning based features fusion and selection framework for smart healthcare. *Neural networks*, *160*, pp.238-258.

[6] Sulthana, R., Chamola, V., Hussain, Z., Albalwy, F. and Hussain, A., 2024. A novel end-to-end deep convolutional neural network based skin lesion classification framework. *Expert Systems with Applications*, *246*, p.123056.

[7] Deng, X., 2024. LSNet: a deep learning based method for skin lesion classification using limited samples and transfer learning. *Multimedia Tools and Applications*, pp.1-21.

[8] Alenezi, F., Armghan, A. and Polat, K., 2023. Wavelet transform based deep residual neural network and ReLU based Extreme Learning Machine for skin lesion classification. *Expert Systems with Applications*, *213*, p.119064.

[9] Ajmal, M., Khan, M.A., Akram, T., Alqahtani, A., Alhaisoni, M., Armghan, A., Althubiti, S.A. and Alenezi, F., 2023. BF2SkNet: Best deep learning features fusion-assisted framework for multiclass skin lesion classification. *Neural Computing and Applications*, *35*(30), pp.22115-22131.

[10] Bozkurt, F., 2023. Skin lesion classification on dermatoscopic images using effective data augmentation and pre-trained deep learning approach. *Multimedia Tools and Applications*, *82*(12), pp.18985-19003.

[11] Dillshad, V., Khan, M.A., Nazir, M., Saidani, O., Alturki, N. and Kadry, S., 2023. D2LFS2Net: Multi-class skin lesion diagnosis using deep learning and variance-controlled Marine Predator optimisation: An application for precision medicine. *CAAI Transactions on Intelligence Technology*.

[12] Tsai, W.X., Li, Y.C. and Lin, C.H., 2023. Skin lesion classification based on multi-model ensemble with generated levels-of-detail images. *Biomedical Signal Processing and Control*, *85*, p.105068.

[13] Arora, G., Dubey, A.K. and Jaffery, Z.A., 2023. Multiple skin lesion classification using deep, ensemble, and shallow (DEnSha) neural networks approach. *International Journal of System Assurance Engineering and Management*, *14*(Suppl 1), pp.385-393.

[14] Akram, A., Rashid, J., Jaffar, M.A., Faheem, M. and Amin, R.U., 2023. Segmentation and classification of skin lesions using hybrid deep learning method in the Internet of Medical Things. *Skin Research and Technology*, *29*(11), p.e13524.

[15] Asiri, Y., Halawani, H.T., Algarni, A.D. and Alanazi, A.A., 2023. IoT enabled healthcare environment using intelligent deep learning enabled skin lesion diagnosis model. *Alexandria Engineering Journal*, *78*, pp.35-44.

[16] Khan, M.A., Akram, T., Zhang, Y.D., Alhaisoni, M., Al Hejaili, A., Shaban, K.A., Tariq, U. and Zayyan, M.H., 2023. SkinNet-ENDO: Multiclass skin lesion recognition using deep neural network and Entropy-Normal distribution optimization algorithm with ELM. *International Journal of Imaging Systems and Technology*, *33*(4), pp.1275-1292.

[17] Golnoori, F., Boroujeni, F.Z. and Monadjemi, A., 2023. Metaheuristic algorithm based hyper-parameters optimization for skin lesion classification. *Multimedia Tools and Applications*, *82*(17), pp.25677-25709.

[18] Yang, Y., Xie, F., Zhang, H., Wang, J., Liu, J., Zhang, Y. and Ding, H., 2023. Skin lesion classification based on two-modal images using a multi-scale fully-shared fusion network. *Computer Methods and Programs in Biomedicine*, *229*, p.107315.

[19] Kaur, R. and Kaur, N., 2024. Ti-FCNet: Triple fused convolutional neural network-based automated skin lesion classification. *Multimedia Tools and Applications*, *83*(11), pp.32525-32551.

[20] QasimGilani, S., Syed, T., Umair, M. and Marques, O., 2023.Skin cancer classification using deep spiking neural network. *Journal of Digital Imaging*, *36*(3), pp.1137-1147.

[21] Nigar, N., Umar, M., Shahzad, M.K., Islam, S. and Abalo, D., 2022. A deep learning approach based on explainable artificial intelligence for skin lesion classification. *IEEE Access*, *10*, pp.113715-113725.

[22] Khan, M.A., Hamza, A., Shabaz, M., Kadry, S., Rubab, S., Bilal, M.A., Akbar, M.N. and Kesavan, S.M., 2024. Multiclass skin lesion classification using deep learning networks optimal information fusion. *Discover Applied Sciences*, *6*(6), pp.1-13.

[23] SM, J., P, M., Aravindan, C. and Appavu, R., 2023. Classification of skin cancer from dermoscopic images using deep neural network architectures. *Multimedia Tools and Applications*, *82*(10), pp.15763-15778.

[24] Alsahafi, Y.S., Kassem, M.A. and Hosny, K.M., 2023. Skin-Net: a novel deep residual network for skin lesions classification using multilevel feature extraction and cross-channel correlation with detection of outlier. *Journal of Big Data*, *10*(1), p.105.

[25] Raghavendra, P.V., Charitha, C., Begum, K.G. and Prasath, V.B.S., 2023. Deep Learning–Based Skin Lesion Multi-class Classification with Global Average Pooling Improvement. *Journal of Digital Imaging*, *36*(5), pp.2227-2248.

[26] Rezaee, K. and Zadeh, H.G., 2024. Self-attention transformer unit-based deep learning framework for skin lesions classification in smart healthcare. *Discover Applied Sciences*, *6*(1), p.3.

[27] Thanka, M.R., Edwin, E.B., Ebenezer, V., Sagayam, K.M., Reddy, B.J., Günerhan, H. and Emadifar, H., 2023. A hybrid approach for melanoma classification using ensemble machine learning techniques with deep transfer learning. *Computer Methods and Programs in Biomedicine Update*, *3*, p.100103.

[28] https://www.kaggle.com/datasets/mdefajalam/isic-2019-skin-disease

[29] Qiang, X., Xue, R. and Zhu, Y., 2024. Confidence partitioning sampling filtering. *EURASIP Journal on Advances in Signal Processing*, *2024*(1), p.24.

[30] Azad, R., Arimond, R., Aghdam, E.K., Kazerouni, A. and Merhof, D., 2023, October. Dae-former: Dual attention-guided efficient transformer for medical image segmentation. In *International Workshop on PRedictive Intelligence InMEdicine* (pp. 83-95). Cham: Springer Nature Switzerland.

[31] Zhao, B., Tang, P., Luo, X., Li, L. and Bai, S., 2022. SiUNet3+-CD: A full-scale connected Siamese network for change detection of VHR images. *European Journal of Remote Sensing*, *55*(1), pp.232-250.

[32] Kong, X., Zhou, W., Shen, G., Zhang, W., Liu, N. and Yang, Y., 2023. Dynamic graph convolutional recurrent imputation network for spatiotemporal traffic missing data. *Knowledge-Based Systems*, *261*, p.110188.

[33] Fu, S., Li, K., Huang, H., Ma, C., Fan, Q. and Zhu, Y., 2024. Red-billed blue magpie optimizer: a novel metaheuristic algorithm for 2D/3D UAV path planning and engineering design problems. *Artificial Intelligence Review*, *57*(6), pp.1-89.

[34] Lin, A., Chen, B., Xu, J., Zhang, Z., Lu, G. and Zhang, D., 2022. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, *71*, pp.1-15.

[35] Yao, T., Li, Y., Pan, Y., Wang, Y., Zhang, X.P. and Mei, T., 2023. Dual vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, *45*(9), pp.10870-10882.

Optimized Wavelet Scattering Network and CNN for ECG Heartbeat Classification from MIT–BIH Arrhythmia Database

Mohamed Elmehdi AIT BOURKHA[1]*, Anas HATIM[2], Dounia NASIR[3], Said EL BEID[4]
Research Laboratory in Innovative and Sustainable Technologies (LaRTID),
National School of Applied Sciences (ENSA) of Marrakech, Cadi Ayyad University (UCA), 40000, Marrakech, Morocco[1, 2, 3]
Control and Computing for Smart Systems and Green Energy (CISIEV),
ENSA of Marrakech, UCA, 40000, Marrakech, Morocco[4]

***Abstract*—Early detection of cardiovascular diseases is vital, especially considering the alarming number of deaths worldwide caused by heart attacks, as highlighted by the world health organization. This emphasizes the urgent need to develop automated systems that can ensure timely and accurate identification of cardiovascular conditions, potentially saving countless lives. This paper presents a novel approach for heartbeats classification, aiming to enhance both accuracy and prediction speed in classification tasks. The model is based on two distinct types of features. First, morphological features that obtained by applying wavelet scattering network to each ECG heartbeat, and the maximum relevance minimum redundancy algorithm was also applied to reduce the computational cost. Second, dynamic features, which capture the duration of two pre R–R intervals and one post R–R interval within the analyzed heartbeat. The feature fusion technique is applied to combine both morphological and dynamic features, and employ a convolutional neural network for the classification of 15 different ECG heartbeat classes. Our proposed method demonstrates an overall accuracy of 98.50% when tested on the Massachusetts institute of Technology -Boston's Beth Israel hospital arrhythmia database. The results obtained from our approach highlight its superior performance compared to existing automated heartbeat classification models.**

Keywords*—Electrocardiogram (ECG); Convolutional Neural Network (CNN); Arrhythmia Rhythm (ARR); Maximum Relevance Minimum Redundancy (MRMR); Wavelet Scattering Network (WSN)

I. INTRODUCTION

The cardiac conduction system ensures an electrical impulse from pacemaker cells in the Sinoatrial (SA) node travels through atria and ventricles, causing a coordinated and timely muscle contraction [1]. Components include SA node, Atrioventricular (AV) node, bundle of His, bundle branches, and Purkinje network.

Einthoven pioneered visualizing heart electrical activity using vectors in an equilateral triangle [2]. Six standard leads: I, II, III, aV_R, aV_L, and aV_F provide a frontal view, while combining them gives a biplanar view of the 3D heart. ECG records instant heart electrical activity on the surface.

ECG signals are commonly employed for the diagnosis of Cardiovascular Diseases (CVD) [3]. Advances in digital tech and cost-effective miniaturized acquisition units have led to the digital acquisition and processing of ECG signals.

Studying Electrocardiogram (ECG) signals could greatly improve early detection of CVD, a major cause of mortality globally [4]. More than that coronary heart disease contributes to premature deaths, disabilities, and a cycle of poverty and ill health [5]. Manual CVD analysis is a time-consuming, an error-prone, especially with large datasets, and requires extensive training due to the signal's complexity, as referred in [6-7-8, 9]. Mistakes in ECG analysis can lead to incorrect diagnoses and treatment. hence, the automatic classification of arrhythmias in ECG signals can be very useful as it can not only offer an impartial diagnosis, but it also has the potential to reduce the workload of medical professionals. As a result, the identification and categorization of ECG hold substantial importance in this field [10], facilitating advancements in CVD research.

Cardiac ARR, termed abnormal cardiac rhythms, arise from irregular initiation or propagation of cardiac excitation waves [11]. They are categorized into ventricular such as Ventricular Premature Complex (VPC), couplets, and triplets, and into supraventricular such as Supraventricular Tachycardia (SVT) and Atrial Fibrillation (AF) types [12]. Ventricular Arrhythmia Rhythm (ARR) stemming from heart's lower chambers, heighten the risk of sudden cardiac death, causing around 450,000 annual US fatalities. Most deaths result from ventricular tachycardia progressing to fatal ventricular fibrillation [13], necessitating prompt defibrillation.

The rest of the paper begins with the related work in Section II, where automated arrhythmia detection methods are discussed in-depth. This is followed by the ECG heartbeat classification in Section III, which starts with a detailed description of the dataset. Next, the feature extraction section provides an in-depth explanation of feature extraction using WSN. The MRMR section then presents a comprehensive overview of dimensionality reduction and the optimization of scattering paths within the scattering network. The subsequent section describes the proposed CNN model for the classification task. Finally, the results in Section IV presents the various outcomes achieved in terms of evaluation metrics and optimization results, while the discussion in Section V

compares the findings of this research with state-of-the-art models. Finally, the paper is concluded in Section VI.

II. Related Works

In the last ten years, significant advancements have occurred in automatic ECG classification algorithms that employs classical machine learning models such as decision trees [18], linear discriminants [19], and logistic regression [20] for diagnosing cardiac arrhythmias [14-15-16, 17]. Techniques like Naïve Bayes, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) have also been utilized in this context [21-22, 23]. Artificial Neural Networks (ANN) have emerged also as a powerful tool [24-25, 26], capable of real-time arrhythmia detection through the recognition of intricate patterns and correlations in ECG signals. Other approaches combine feature extraction with machine learning, including time domain features [26], frequency domain features [27], and combinations of both [28]. Wavelet analysis has also proven effective [29-30].

Recently, deep learning has become a promising approach for analyzing ECG signals, outperforming traditional machine learning methods. Models like Convolutional Neural Networks (CNN) [31-32], Recurrent Neural Networks (RNN) [33], and Long-Short Term Memory (LSTM) neural networks [34] excel due to their automatic feature extraction from raw ECG data. GPU and TPU, integral to high-performance computing, have significantly bolstered deep learning in ECG analysis by efficiently processing extensive data.

Noteworthy datasets like PhysioNet [35] and PTB Diagnostic ECG database have further advanced deep learning in this field. These datasets encompass diverse ECG signals with varying abnormalities, facilitating improved learning and generalization of deep learning models. Moreover, deep learning has proven effective in vital tasks such as denoising [36], segmentation, and reconstruction of ECG signals. Recent advancements in ECG have enhanced the diagnosis and treatment of CVD, a leading global cause of mortality [37-38].

Past research has concentrated on categorizing ECG signals into broader groups like Normal Sinus Rhythm (NSR), ARR, and Congestive Heart Failure (CHF) [39-40]. Some researchers have also suggested interpreting ECG heartbeats based on the Association for the Advancement of Medical Instrumentation (AAMI) classes: non-ectopic beats (N), supraventricular ectopic beats (S), ventricular ectopic beats (V), fusion beats (F), and unknown beats (Q) [41-42]. Some other works used the annotations from the American Heart Association (AHA) that has proposed a set of 15 classes for arrhythmia classification based on the MIT–BIH arrhythmia Database.

Osowski et al. [43] proposed a method to classify ECG heartbeats from the MIT–BIH arrhythmia database with 13 classes. They combined features extracted using Higher-Order Statistics (HOS) and Hermite characterization of the QRS complex, feeding them into an SVM classifier.

Rodriguez et al. [44] describe their approach to creating a classification algorithm for a variety of 14 ECG heartbeat classes using the MIT–BIH arrhythmia database. Their reported outcomes demonstrate significant accuracy. Furthermore, they highlight the algorithm's integration potential with Personal Digital Assistants (PDA).

Chen et al. [45] proposed an innovative method for categorizing ECG beats. Their approach combines projected and dynamic features. The projected features involve a random projection matrix, normalized columns, and row-wise Discrete Cosine Transform (DCT). The dynamic features include 3 weighted R–R intervals. The SVM classifier is employed for the categorization of heartbeats into either 15 or 5 distinct classes.

Ihasanto et al. [46] present the Ensemble Multilayer Perceptron (MLP) method, streamlining ECG beat classification by integrating feature extraction and classification into one step. This eliminates the need for a separate preprocessing stage, potentially reducing computational requirements. The technique achieves over 97% accuracy, even with 10 ECG heartbeat classes.

Melgani et al. [47] demonstrated SVM effective generalization in classifying sets of 10 ECG beats. They introduced an innovative approach, combining Particle Swarm Optimization (PSO) with SVM, to enhance the performance.

Alqudah et al. [48] introduced an innovative deep learning technique aimed at arrhythmia classification through ECG analysis, utilizing iris spectrograms. Their method demonstrated remarkable recognition performance accuracy rates of 99.13%, 98.223%, and 97.494% for 13, 15, and 17 distinct categories, respectively. These results were obtained using a dataset comprising 744 ECG from 45 individuals.

Alqudah et al. [49] compared spectrogram representations in various CNN architectures using a dataset of 10,502 heartbeats from MIT–BIH arrhythmia database, covering 6 classes. They explored Log/Mel-Scale spectrograms, Bi-Spectrum, and third-order cumulant in models like AOCT-NET, Mobile-Net, Squeeze-Net, and Shuffle-Net.

Rajkumar et al. [50] devised an intelligent approach employing CNN for the automated classification of ECG signals. Their methodology obviates the need for manual feature extraction, potentially enhancing the efficiency of cardiac patient screening for cardiologists. Demonstrating its effectiveness, the CNN adeptly categorized seven distinct ARR classes sourced from the MIT–BIH database.

Shaker et al. [51] improved deep learning on the MIT–BIH arrhythmia dataset with GAN-based data augmentation. They used two CNN based approaches to avoid manual feature engineering.

Ramkumar et al. [52] used ECG data from MIT–BIH arrhythmia database to classify normal, atrial premature, and ventricular escape heartbeats. They employed wavelet transform for preprocessing, independent component analysis for feature extraction, and MLP for classification.

Arslan et al. [53] simultaneously trained an autoencoder and classifier, allowing the network to reduce overall error while accurately reconstructing the input and extracting key features useful for classification. Their work focused on classifying six types of ECG heartbeats, including normal beats, left and right bundle branch block beats, premature

ventricular contractions, atrial premature beats, and paced beats, using data from the MIT–BIH dataset. They achieved a classification accuracy of 99.99% by employing a convolutional autoencoder with an integrated classifier.

Vavekanand et al. [54] used deep CNN to classify ECG beats as either normal or abnormal. They applied transfer learning, first training a generic model on ECG data from the MIT–BIH, then fine-tuning the model for specific patients. They compared the performance of these fine-tuned models of individual models trained only on a single patient's data. Both approaches performed well, with individual classifiers achieving an average balanced accuracy of 94.6% on the test set, while the fine-tuned models had a slightly lower accuracy of 93.5%.

Zhou et al. [55] transformed ECG signals into different types of images using techniques like Recurrence Plot (RP), Gramian Angular Field (GAF), and Markov Transition Field (MTF), which they then fed into their classification model. To better retain important details, they developed a CNN based model with FCA for handling multiple types of ECG tasks. Their model achieved an accuracy of 99.6% when classifying five types of heartbeats using data from the MIT–BIH arrhythmia database.

Although the previous proposed approaches successfully classify ECG heartbeats, most state-of-the-art methods tend to focus on specific types of heartbeats rather than addressing all 15 types. When attempting to classify all heartbeat types, the proposed methods show lower performances compared to those works that limit their scope to just a few types. Additionally, many existing studies on heartbeat classification fail to analyze the computational complexity of their models, which makes them impractical for real-world applications. The importance of this new research lies in addressing the challenge of low classification performance in ECG heartbeat analysis, especially when dealing with a large number of classes. To overcome this issue, it proposes a novel approach to optimize the WSN and CNN, both of which have demonstrated effectiveness in classification tasks, particularly in arrhythmia detection. The main objective of this research paper is to develop an advanced heartbeat classification model that achieves high accuracy. Additionally, the proposed model is optimized to ensure its applicability in clinical settings.

Based on these observations, the contributions of this study can be summarized as follows:

- Develop a new method for detecting arrhythmias from ECG heartbeats with high accuracy.
- Optimize the Wavelet Scattering Network (WSN) for feature extraction to make the proposed approach suitable for clinical use.

III. ECG HEARTBEAT CLASSIFICATION

A. Data Acquisition

The research utilizes publicly available data from PhysioNet in .mat format [56], sourced from MIT–BIH arrhythmia database [57] a collection of ECG recordings.

Our research used data from the MIT–BIH arrhythmia database, including 48 labeled ECG signals taken over 30 minutes from 47 individuals. Lead II ECG signals are used in this study due to their sensitivity to heart rhythm, crucial for ARR detection. Sampling frequency was 360 Hz to capture detailed heart activity.

Our study aims to categorize ECG heartbeats into 15 distinct types as outlined in Table I. The sequential procedure outlined in Fig. 1 shows different steps of our proposed approach. Initially, to detect QRS complexes in each ECG signal, the Pan-Tompkins algorithm [58] is employed, which is a well-known method commonly used for QRS detection in ECG signals. This algorithm is tailored for accurate identification of QRS complexes in ECG signals.

By applying the Pan-Tompkins algorithm, the majority of R peaks in ECG signals were successfully detected, ensuring comprehensive coverage of the QRS complex. A window of 300 samples before and 500 samples after the R peak was used, resulting in 801 sample ECG heartbeats at 360 Hz sampling frequency. This approach captured P waves, QRS complexes, and T waves, enabling accurate arrhythmias identification. Fig. 2 presents normal ECG heartbeat detected by pan-Tompkins algorithm. Dynamic features like post R–R interval, pre R–R interval, pre-pre R–R interval, and the (post R–R)-(pre R–R) interval difference were computed, aiding arrhythmia diagnosis based on heart rate variability.

B. Data Preprocessing

In this research, the data was utilized in its original raw, unprocessed state without applying any data cleaning techniques. Afterwards, the dataset was split into two segments, one for training and the other for testing. The hold-out validation technique was employed, incorporating stratification to ensure an equal representation from each class. Specifically, 80% of samples from each class were allocated for training, while the remaining samples were designated for testing. The distribution of the training and testing data are illustrated in Table II.

TABLE I. ECG HEARTBEATS DISTRIBUTION

Symbols	Types of ECG Heartbeats	No. of Heartbeats
N	Normal beat	11000
L	Left bundle branch block beat	8059
R	Right bundle branch block beat	7235
A	Atrial premature beat	1769
E	Ventricular escape beat	106
V	Premature ventricular contraction	6178
/ (P)	Paced beat	7017
F	Fusion of ventricular and normal beat	801
Q	Unknown beat	33
!	Ventricular flutter wave	472
a	Aberrated atrial premature beat	150
e	Atrial escape beat	16
f	Fusion of paced and normal beat	982
j	Nodal (junctional)escape beat	229
x	Non-conducted P wave (Blocked APB)	193

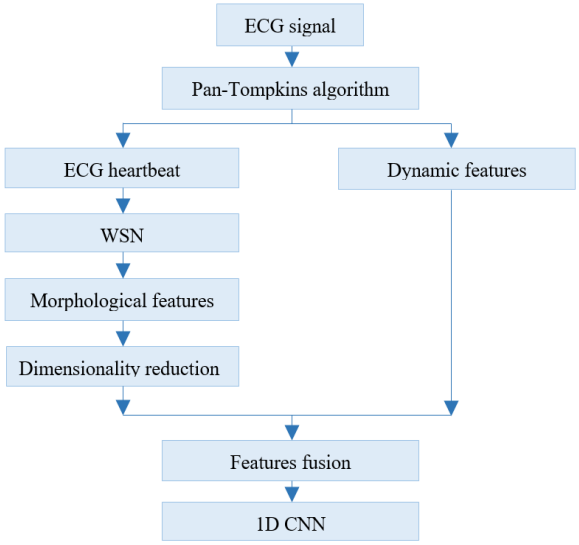


Fig. 1. Different steps of ECG heartbeat classification.

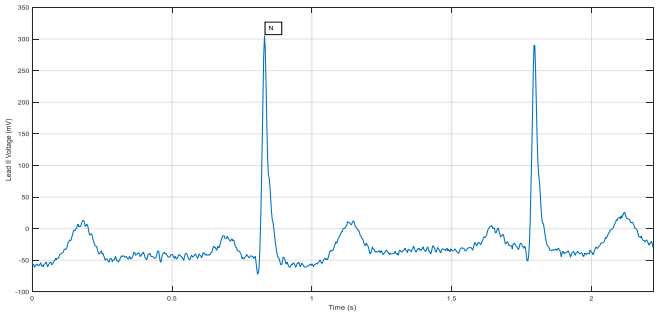


Fig. 2. Normal ECG heartbeat.

TABLE II. HEARTBEAT DISTRIBUTION FOR TRAINING AND TESTING

Symbols	Training Instances	Testing Instances
N	8800	2200
L	6448	1611
R	5788	1447
A	1416	353
E	85	21
V	4943	1235
/ (P)	5614	1403
F	641	160
Q	27	6
!	378	94
a	120	30
e	13	3
f	786	196
j	184	45
x	155	38

The dataset used for training shows an imbalance in class distribution, which lead to biased predictions and reduced model performance. To address this, the Synthetic Minority Oversampling Technique (SMOTE) [59] is employed.

C. Feature Extraction

Our research aimed to classify ECG heartbeats into 15 classes to detect ARR. Accurate predictions relied on extracting key features, split into time domain for capturing signal variations, and frequency domain for understanding spectral characteristics. Time domain features reflect changes, while frequency domain features reveal signal frequencies. The WSN was utilized as a powerful technique, for efficient feature extraction from ECG signals. This method capitalizes on wavelet transforms unique properties to capture local and global ECG waveform variations.

The WST is a mathematical method rooted in wavelets, allowing efficient signal analysis. Its strengths include translation and rotation invariances, suitable for image and audio analysis, stable features for denoising, and dimensionality reduction for enhanced accuracy [60]. This versatile technique finds use various domains such as audio, image, biomedical signals, speech, computer vision, and finance, excelling in classification and signal processing [61].

Fig. 3 illustrates the WSN with multiple layers, each applying a WST consisting of 3 stages. In the first stage, convolution uses a scale-specific wavelet ψ, to gauge similarity between the wavelet and the signal. The second stage introduces nonlinearity via modulus operations to retain signal magnitude and diminish redundant details like noise. The final stage involves low-pass filter convolution Φ , for dimensionality reduction enhancing signal representation.

The Gabor complex wavelet offers valuable traits beyond its time-frequency focus, including selective transmission of low-frequency components through its modulus acting as a low pass filter. By representing signal envelopes, it's adept at tasks like denoising and feature extraction [62].

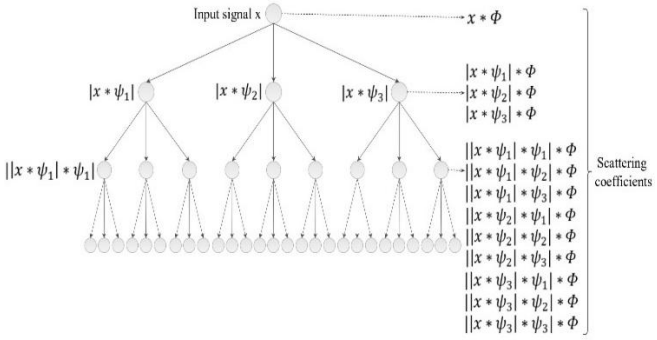


Fig. 3. Wavelet scattering network.

A Gabor wavelet in Eq. (1) is defined as the product of a Gaussian function and a complex exponential function.

$$\psi(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-t^2}{2\sigma^2}} e^{i\omega t} \quad (1)$$

Where t is the time, and σ is the standard deviation of the Gaussian function. In $\omega = 2\pi f$, f is the center frequency of ψ, and i is the imaginary unit. The envelope of the Gabor complex wavelet represents the low-pass filter, denoted as Φ.

$$\Phi(t) = |\psi(t)| \quad (2)$$

The scale parameter σ in Gabor wavelet analysis is denoted as the standard deviation of the Gaussian envelope. It shapes the wavelet window. Larger σ means a wider Gaussian envelope for a broader wavelet, while smaller σ leads to a more localized wavelet. In ECG analysis, Gabor wavelets suit the QRS complex detection due to its resemblance to the QRS waveform, making it suitable for arrhythmia detections.

Fig. 4 outlines the Gabor complex wavelet with its real and imaginary parts, with 0.5 second invariance scale. Fig. 5 displays frequency bands for different scaling functions used.

$$S_1 x(t) = |x * \psi_{\sigma_1}| * \Phi \tag{3}$$

In the initial stage, the signal undergoes convolution with a low pass filter Φ in Eq. (4), offering high time resolution but limited frequency accuracy. Progressing to the first order in Eq. (3), 29 paths using various Gabor wavelets capture fast variations, yet some high-frequency details are lost due to a final convolution with this filter Φ. The second order in Eq. (5) employs 10 paths with different wavelets, enhancing frequency resolution. The resulting coefficients of all stages are summarized in matrix S presented in Eq. (6), providing a comprehensive multi-scale description of signal variations.

$$S_0 x(t) = x(t) * \Phi \tag{4}$$

$$S_2 x(t) = ||x * \psi_{\sigma_1}| * \psi_{\sigma_2}| * \Phi \tag{5}$$

$$Sx(t) = \{S_0, S_1, \dots, S_n\} \tag{6}$$

The study used a second order WSN to maintain 99% of signal energy. This choice prevented data loss and excessive computation. Two filter banks were employed, $Q_1 = 8$ and $Q_2 = 1$, ensuring accurate and efficient signal analysis.

One ECG heartbeat represented as a feature matrix of 40x26. The training dataset grows to 145,250 instances using the SMOTE technique. To handle this, the matrix becomes 40x26x145,250. Down-sampling with low pass filter reduces scattering coefficients, creating 26–time windows for 40 paths. Each tensor entry represents a path and time window.

Fig. 6(a) outlines the scattering coefficients of the first filter bank. The scattergram depicts time on the x–axis and frequency on the y–axis, and their amplitudes on the z–axis.

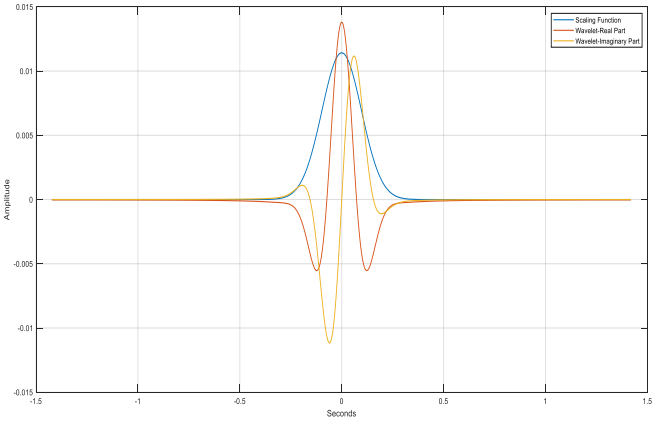


Fig. 4. Real part, imaginary part and low pass filter of Gabor wavelet.

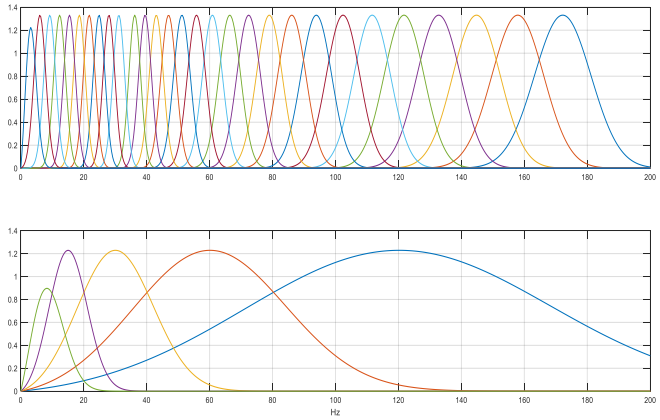
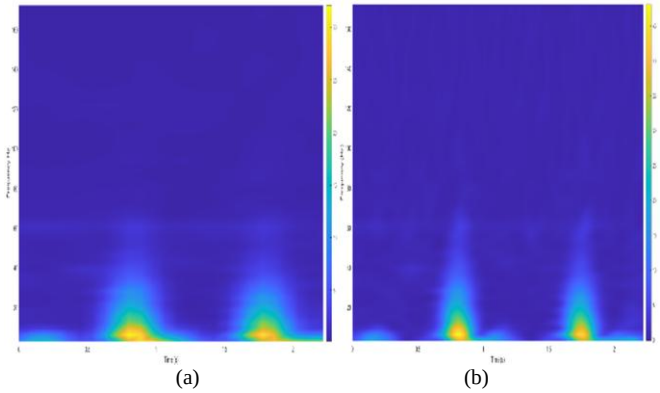


Fig. 5. Bandwidths of the first and the second filter banks.



(a) (b)

Fig. 6. Scattergram of the first filter bank. (a) Scattering coefficients, (b) Scalogram coefficients.

Fig. 6(b) displays the scalogram of the first filter bank, which effectively captures distinct high frequency details of the ECG signals. The dynamic features are oversampled 40 times, resulting in 40x4 tensor. After fusion, the tensor size becomes 40x30 for one ECG heartbeat.

D. Maximum Relevance Minimum Redundancy (MRMR)

To tackle high computational costs and enhance testing accuracy while avoiding overfitting, the MRMR feature selection technique is employed to reduce the dimensionality. The MRMR algorithm incorporates various parameters, among which are entropy, joint entropy, and mutual information.

Entropy is a measure of uncertainty in a random variable. It helps gauge information within it. To find entropy, the probabilities of outcomes was used in a sequence $\{X_1, X_2, \dots, X_n\}$, denoted as X. This indicates the information required for prediction.

$$H(X) = -\sum_{i=1}^{n} p(x_i).\, log\, p(x_i) \tag{7}$$

Joint entropy gauges the uncertainty within a group of random variables, reflecting their unpredictability when interconnected. It evaluates the information content between two variables X and Y, expressing their correlation.

$$H(X:Y) = -\sum_{x,y} p(x,y).\log(p(x,y)) \tag{8}$$

Where $p(x,y)$ represents the joint probability.

Mutual information serves as a metric to uncover the degree of knowledge sharing between two or more random variables. It gauges their similarity and correlation, shedding light on their interdependencies [64]. During mutual information calculation, the focus lies on shared information and the robustness of their relationship.

A mutual information of zero implies independence between variables, while a higher value indicates a strong relationship and valuable information exchange.

$$I(X;Y) = H(X) - H(X:Y) = \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x).p(y)} \quad (9)$$

Where $H(X:Y)$ is the uncertainty left about X after knowing Y.

Common dimensionality reduction methods like Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA) often focus solely on feature category relationships, overlooking mutual information between features and targets. In contrast, the MRMR algorithm considers both, resulting in enhanced feature selection, predictive accuracy, and interpretability, offering a superior approach to dimensionality reduction. The MRMR algorithm measures the mutual information between features and the class label. A higher value indicates strong correlation, making a feature significant for classification. So, high mutual information between a feature X and class label C implies its importance in classification.

$$V_s = \frac{1}{|S|} \sum_{x\ \varepsilon\ S} I(x;C) \quad (10)$$

The concept of maximal relevance V_s involves identifying features that have the highest mutual information, represented by max V_s, with the target class label C. And $|S|$ which is the number of subset features in S.

Minimal redundancy evaluates the shared information between two features. High shared information implies redundancy. If two features convey the same data, one can be chosen for selection, reducing dimensions. Lower redundancy means better feature selection. So, the aim is to find features with low shared data. Let's set S as subset features, and X and Y are the features. The redundancy computes as follows:

$$W_s = \frac{1}{|S|^2} \sum_{x,y\ \varepsilon\ S} I\ (x;y) \quad (11)$$

The objective is to identify features with minimum redundancies by minimizing the function W_s , where $|S|$ represents the number of subset features in S, and $I(X;Y)$ denotes the mutual information between features X and Y.

The MRMR algorithm aims to discover a subset of features that display high relevance and low redundancy. It considers both the interdependencies among features and their connections to the target variable. It aims to maximize a specific function to find features with the highest relevance V_s to the target variable, while minimizing redundancy W_s among themselves [63]. The main goal is to achieve a balance between feature importance and mutual information, resulting in the selection of an optimal set of features.

$$\max(\beta)\ , \beta = \frac{V_S}{W_S} \quad (12)$$

Our study focuses on scoring features in subset features S. These scores stem from computing the mutual information between the target and the feature in question. This value is then divided by the mean mutual information between the previously chosen feature and the current one.

$$Score_i(x) = \frac{I(x;C)}{\sum_{S\varepsilon\, i-1\ selected\ features} \frac{I(x;S)}{m}} \quad (13)$$

Where m is the number of features in the subset S.

The MRMR algorithm evaluates scores of the 40 scattering paths in the scattering network as illustrated in Fig. 7. These scores indicate feature significance in classifying 15 ECG heartbeats. Higher scores highlight critical roles while lower scores mean less important features that can be replaced without accuracy loss.

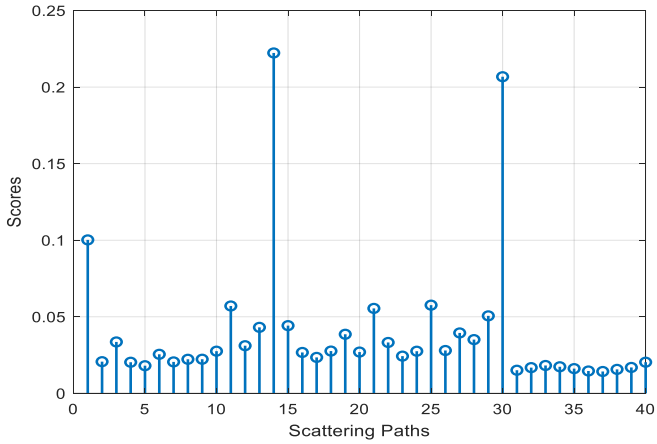


Fig. 7. Scores of different paths in wavelet scattering network.

E. Heartbeat Classification Using CNN

CNN is a specialized model designed specifically for processing 1-dimensional sequential data. Its key component is the convolutional layer, which employs filters to detect local patterns and dependencies within the input sequence. CNN chosen as a classifier to detect 15 heartbeat classes for efficiency. This minimizes computational costs compared to a 2D CNN. Our decision considers available resources and the need to process large data quickly.

In this research, a CNN with 14 hidden layers was used for classification as depicted in Table III. These layers include convolution, ReLU, normalization, global average pooling, fully connected, and SoftMax. They collectively differentiate the 15 ECG heartbeat classes. The convolutional layer involves sliding filters for computation as in Eq. (14).

$$y_t = b + \sum_{i=1}^{k} \omega_i . x_{t-i+1} \quad (14)$$

To ensure causality in the feature map, the output y_t is determined by the current input and the past samples at each time step t. This means that for a given feature vector x and a filter kernel ω, the mathematical formula for convolution can be expressed by taking into account both the current feature and previous inputs, meaning that a model only has access to past inputs to keep the causal nature of the process.

The ReLU function sets negatives to zero, leaving positives unaffected as presented in Eq. (15).

$$y_i = \max(0, x_i) \quad (15)$$

TABLE III. CNN LAYERS

Layer Types	Activations	Learnables	Total Learnables
Sequence Input	40	—	0
Convolution 1D Padding: [7,0] Stride: 1 No. of Filters: 16 Filter Size: 8	16	Weights: 8x40x16 Bias: 1x16	5136
ReLU	16	—	0
Layer Normalization	16	Offset: 16x1 Scale: 16x1	32
Convolution 1D Padding: [7,0] Stride: 1 No. of Filters: 32 Filter Size: 8	32	Weights: 8x16x32 Bias: 1x32	4128
ReLU	32	—	0
Layer Normalization	32	Offset: 32x1 Scale: 32x1	64
Convolution 1D Padding: [7,0] Stride: 1 No. of Filters: 64 Filter Size: 8	64	Weights: 8x32x64 Bias: 1x64	16448
ReLU	64	—	0
Layer Normalization	64	Offset: 64x1 Scale: 64x1	128
1D Global Average Pooling	64	—	0
Fully Connected Layers	15	Weights: 15x64 Bias: 15x1	975
SoftMax	15	—	0
Classification Output Loss Function: Cross-entropy	15	—	0

Normalization layers like batch normalization enhance neural network stability and convergence by standardizing prior layer outputs, ensuring suitable input ranges for downstream layers and promoting effective learning.

$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\delta^2 + \varepsilon}} \quad (16)$$

Where μ is the mean value of the features from one example, δ is the standard deviation, and the term $\varepsilon=10^{-5}$ used to prevent division by zero, ensures numerical stability. Finally, the output y_t is scaled and shifted as follows:

$$y_i = \gamma . \hat{x}_i + \beta \quad (17)$$

Where γ and β represent the scaling and shifting learnable parameters.

Global Average Pooling layers condense the spatial dimensions of the feature maps into a single value, which reduces the computational complexity of the model. This operation calculates the average of each feature map, resulting in a fixed-length vector.

$$Y = Softmax(\sum WX + b) \quad (18)$$

The SoftMax layer is applied to produce the probability distribution over the 15 ECG heartbeat classes.

$$y_i = \frac{\exp(z_i)}{\sum_{j=1}^{n} \exp(z_j)} \quad (19)$$

The configuration, architecture and settings for each layer of the proposed CNN model, analyzing an ECG heartbeat with the WSN, are outlined in Table III.

In our study, CNN is used for analysis. Training employed the Adam optimizer with a 0.01 initial learning rate. Hyperparameters were manually tuned, like shuffling data using a mini-batch size of 64, enhancing performance, ensuring diverse pattern exposure, and preventing overfitting.

F. Performance Metrics

The effectiveness of our proposed model was evaluated using multiple assessment metrics like accuracy, sensitivity, specificity, precision, F1 score, negative predictive value, false positive rate, false discovery rate, false negative rate, and the Matthew Correlation Coefficient (MCC).

$$\text{Accuracy (ACC)} = \frac{TP+TN}{TP+TN+FP+FN} \quad (20)$$

$$\text{Sensitivity (SEN)} = \frac{TP}{TP+FN} \quad (21)$$

$$\text{Specificity (SPE)} = \frac{TN}{TN+FP} \quad (22)$$

$$\text{Precision (PRE)} = \frac{TP}{TP+FP} \quad (23)$$

$$\text{F1 score (F1)} = \frac{2*precision*sensitivity}{precision+sensitivity} \quad (24)$$

$$\text{Negative Predictive Value (NPV)} = \frac{TN}{TN+FN} \quad (25)$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP+TN} \quad (26)$$

$$\text{False Discovery Rate (FDR)} = \frac{FP}{FP+TP} \quad (27)$$

$$\text{False Negative Rate (FNR)} = \frac{FN}{FN+TP} \quad (28)$$

$$\text{MCC} = \frac{(TP.TN)-(FP.FN)}{\sqrt{(TP+FP).(TP+FN).(TN+FP).(TN+FN)}} \quad (29)$$

The MATLAB Version R–2021b programming language was utilized to implement all algorithms on windows server. The system used for execution had an Intel(R) Core (TM), i5, CPU 6300U processor with a clock speed of 2.40 GHz. The RAM capacity was 12 GB, operated on a 64–bit architecture.

IV. RESULTS

The purpose of our research is to distinguish 15 types of ECG heartbeats. By integrating the MRMR algorithm, the computational challenges are reduced. Moreover, a CNN is used as a classifier to successfully identify and classify ECG heartbeats. The WSN generated a 40x26 feature matrix, and MRMR evaluated path importance by calculating scores.

The study employed dimensionality reduction through score sorting to discover optimal paths for classification as illustrated in Fig. 8. It started with the top 5 paths, used a CNN model for testing data accuracy, and gradually added 5 more paths at each step until all 40 were included. Following the

approach presented in Fig. 9, the testing phase is assessed accuracy at each step. Increasing scattering paths boosted the testing accuracy as outlined in Fig. 10.

The test results indicate a remarkable accuracy of 98.50% when utilizing $K = 20$ scattering paths for ECG heartbeat classification. Adding more paths beyond the first 20 does not lead to any noticeable improvement in the overall testing accuracy. By selecting the top 20 paths based on their scores, the matrix size is reduced by half, maintaining accuracy while reducing feature dimensions by 50%. These changes include modifying the sequence input layer to 20 activations and adjusting the first convolution layer, resulting a reduction of 9.51% in parameter count. The confusion matrix of 15 ECG heartbeats presented in Fig. 11 shows a 98.50% overall accuracy, although classes "Q" and "e" had lower accuracy due to limited training data, affecting the overall score.

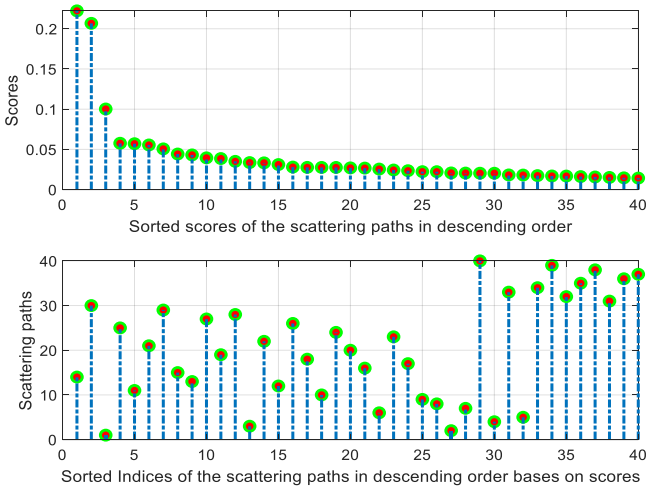


Fig. 8. Sorted scores of scattering paths in descending order.

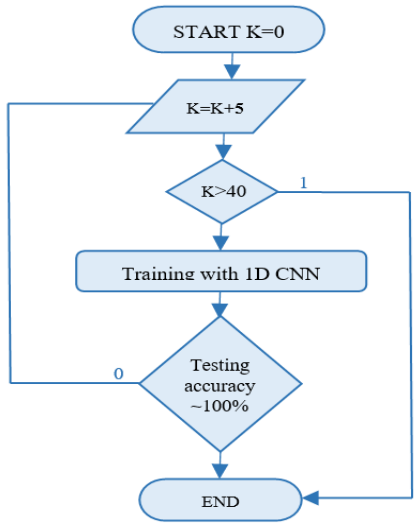


Fig. 9. Scattering paths selection algorithm.

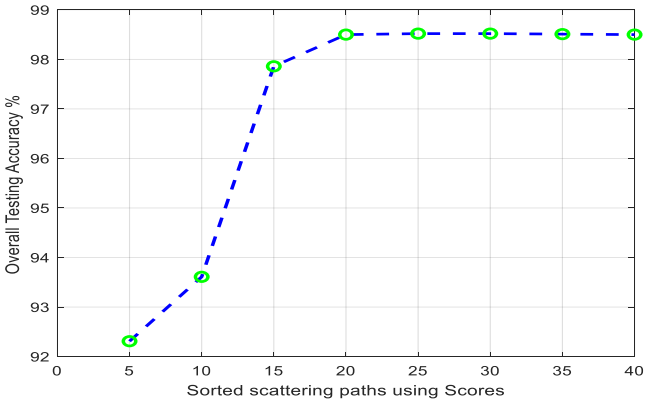


Fig. 10. Impact of adding scattering paths on testing accuracy.

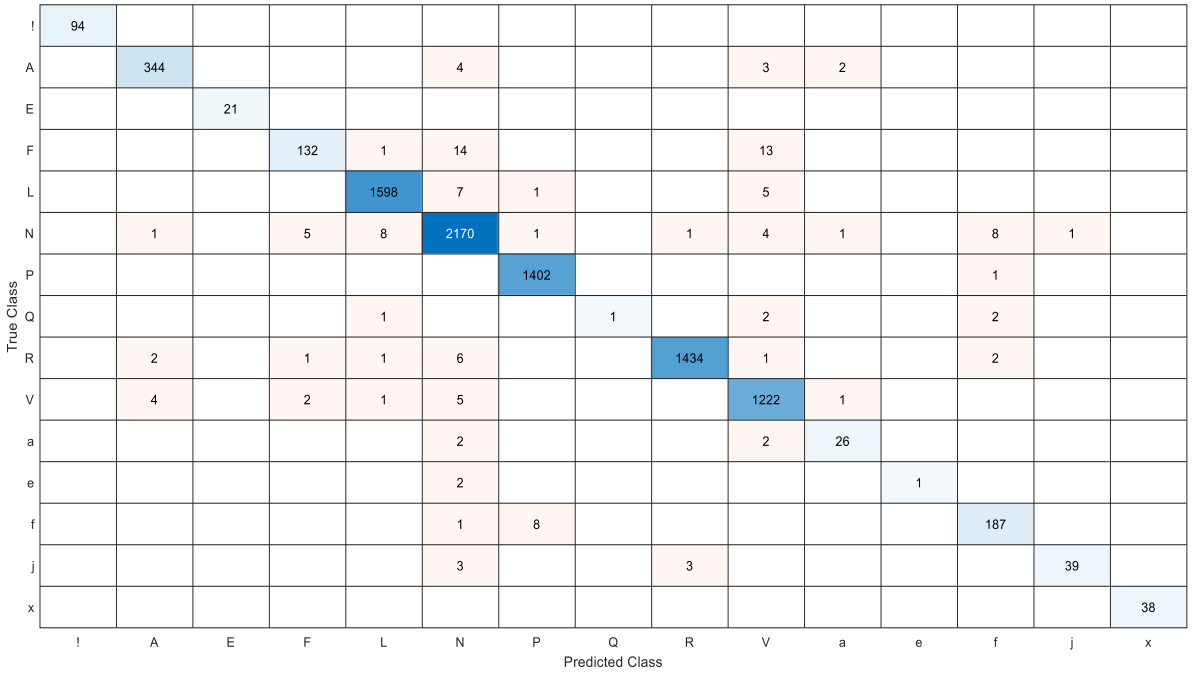


Fig. 11. Testing confusion matrix.

TABLE IV. EVALUATION METRICS OF TESTING DATA

Heartbeat Classes	PRE (%)	SEN (%)	SPE (%)	F1 (%)	NPV (%)	FPR (%)	FDR (%)	FNR (%)	MCC (%)
!	100	100	100	100	100	0	0	0	100
A	98.01	97.54	99.92	97.73	99.89	0.08	1.99	2.55	97.63
E	100	100	100	100	100	0	0	0	100
F	94.29	82.50	99.91	88.00	99.68	0.09	5.71	17.50	88.00
L	99.25	99.19	99.83	99.22	99.82	0.17	0.75	0.81	99.05
N	98.01	98.64	99.34	98.32	99.55	0.66	1.99	1.36	97.77
/	99.29	99.93	99.87	99.61	99.99	0.13	0.71	0.07	99.54
Q	100	16.67	100	28.57	99.94	0	0	83.33	40.81
R	99.72	99.10	99.95	99.41	99.82	0.05	0.28	0.90	99.30
V	97.60	98.95	99.61	98.27	99.83	0.39	2.40	1.05	98.00
A	86.67	86.67	99.95	86.67	99.95	0.05	13.33	13.33	86.62
e	100	33.33	100	50.00	99.98	0	0	66.57	57.73
F	93.50	95.41	99.85	99.44	99.90	0.15	6.50	4.59	94.32
J	97.50	86.67	99.99	91.76	99.93	0.01	2.50	13.33	91.89
x	100	100	100	100	100	0	0	0	100
Average	97.60	86.30	99.88	88.80	99.90	0.12	2.41	13.70	90.01

TABLE V. COMPLEXITY ANALYSIS OF THE PROPOSED MODEL

Execution time for feature extraction from one ECG heartbeat using WSN	14.8 ms
Prediction speed of the CNN	~3254 obs/s
Total number of learnable parameters of the CNN	24351
Memory usage after feature extraction for one ECG heartbeat	4160 bytes
Morphological feature dimensionality for one ECG heartbeat	20x26

Our innovative approach combines WSN with CNN for the classification of 15 ECG heartbeat classes. The results outlined in Table IV contain a remarkable performance metrics such as, a precision of 97.60% demonstrates precise positive predictions, a sensitivity of 86.30% signifies accurate identification of actual positives, a specificity of 99.88% reflects the model's prowess in correctly identifying negatives, a negative predictive value of 99.90% showcases minimal false negatives, a false positive rate of 0.12% indicates an extremely low occurrence of false positives, a false discovery rate of 2.41% demonstrates limited false positive predictions, a false negative rate of 13.70% portrays proficiency in classifying positive heartbeats, and a Matthew correlation coefficient of 90.01% signifies strong overall performance.

As outlined in Table V, our analysis confirms the efficiency and speed of our WSN based model using CNN for ECG heartbeat classification. The feature extraction process takes just 14.8 milliseconds per heartbeat, with a prediction speed of 3254 heartbeats per second.

The model's parameters reduced to 24351, and memory usage per heartbeat is 4160 bytes. Our model delivers swift performance, low parameter count, and minimal computational cost.

V. DISCUSSION

A comprehensive analysis was conducted to evaluate the overall accuracy of the proposed model in this research paper, and compared it with the results from previous studies. This comparison is summarized in Table VI. To ensure a fair comparison, we specifically focused on studies that also addressed the classification of 15 ECG heartbeats from the MIT–BIH arrhythmia database. These studies are referenced as [44-45-48, 51]. The obtained results clearly demonstrate that our proposed model, based on WSN combined with a CNN, outperforms the previous studies cited previously. Our model not only achieves higher accuracy in classifying heartbeats, but it also demonstrates robust performance across various types of heartbeats. It provides more accurate results and shows great potential for accurately detecting different types of heartbeats.

Our proposed method outperforms previous approaches by achieving high classification accuracy across the majority of ECG heartbeat types. Unlike existing models, our optimized technique enhances the model’s performance to a level that enables clinical application. By addressing the limitations of prior methods such as suboptimal feature extraction and insufficient classification accuracy our approach ensures more reliable heartbeat classification. As a result, the proposed

model surpasses state-of-the-art methods, achieving an overall accuracy of 98.50%.

In this study, the proposed approach for ECG heartbeat classification is illustrated in Fig. 12. It begins with QRS detection, using Pan-Tompkins algorithm, which identifies the dynamic features of heartbeats. These heartbeats are then fed into the WSN, where morphological characteristics are extracted. Following that, the most relevant morphological features are selected using the MRMR method combined with the proposed algorithm. After fusing the dynamic and morphological features, a feature matrix of size 20x30 is obtained. This matrix is then input into a specially designed CNN to classify the heartbeats into 15 distinct categories.

TABLE VI. COMPARISON WITH OTHER PREVIOUS WORKS

Studies	No. of Classes	Methodology	Overall Accuracy (%)
Rodriguez et al. [44]	**15**	PDA + Decision trees	96.12
Chen et al. [45]	**15**	Projection + WRR + SVM	98.46
Ihasanto et al. [46]	10	MLP	97
Melgani et al. [47]	10	PSO + SVM	89.72
Alqudah et al. [48]	13 **15** 17	Iris spectrum + CNN	99.13 98.23 97.49
Alqudah et al. [49]	6	STFT + CNN	93.8
Rajkumar et al. [50]	7	End to end CNN	93.6
Shaker et al. [51]	**15**	GAN + end to end CNN	98.0
Ramkumar et al. [52]	3	DWT + ICA + MLP	96.50
Arslan et al. [53]	6	Autoencoder + Classifier	99.99
Vavekanand et al. [54]	2	Deep CNN	94.6
Zhou et al. [55]	5	CNN model with FCA	99.4
Proposed Method	**15**	WSN + MRMR + CNN	**98.50**

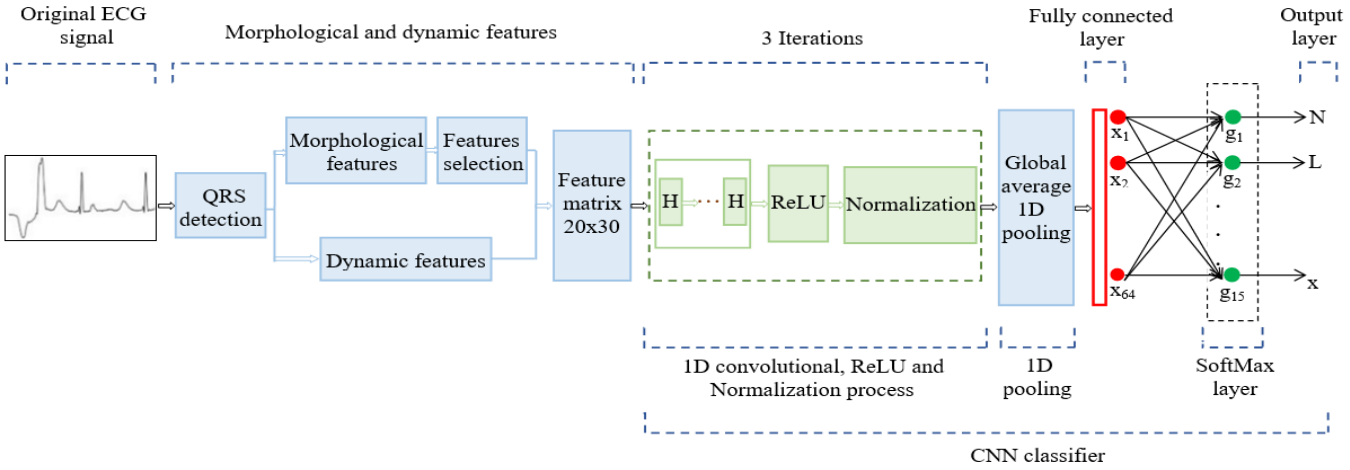


Fig. 12. Proposed framework for ECG heartbeats classification.

VI. CONCLUSION AND PROSPECTS

This study introduces an efficient approach for classifying ECG heartbeats. Our proposed model begins by detecting heartbeats in ECG signals, followed by utilizing the WSN for feature extraction. WSN is a robust technique that captures both temporal and spectral characteristics. Feature fusion is performed, and CNN employed for the classification of ECG heartbeats. Our model outperforms existing systems in terms of prediction accuracy and precision, achieving 98.50% and 97.60%, respectively, and also demonstrating low computational requirements due to the application of an innovative selection path techniques.

Although the WSN has proven effective in extracting important time and frequency domain features, it still suffers from high computational cost due to the convolution with multiple wavelets. Our approach addresses this issue by employing the MRMR technique in order to optimize the model and reduce computational costs, by selecting the most suitable wavelet for the convolution process.

The results of our proposed approach further demonstrate its effectiveness in detecting 15 types of ECG heartbeats from the MIT–BIH arrhythmia database, making it highly applicable for clinical use. This method has the potential to aid in early diagnosis and save numerous lives worldwide. Moving forward, our future work aims to enhance accuracy on testing data and further reduce computational costs to enable implementation on devices such as smartphones.

AVAILABILITY OF DATA AND MATERIALS

ECG readings were taken from [56].

CONFLICT OF INTEREST

We confirm that all authors declare no conflicts of interest.

ACKNOWLEDGMENT

This work was supported by the National Center for Scientific and Technical Research (CNRST), Rabat, Morocco.

REFERENCES

[1] A. Kennedy,D. D. Finlay, D. Guldenring, R. Bond, K. Moran, J. McLaughlin, "The cardiac conduction system: generation and conduction of the cardiac impulse," Critical Care Nursing Clinics, 28(3), pp. 269-79, 2016.

[2] A. D. John, L. A. Fleisher, "Electrocardiography: the ECG. Anesthesiology Clinics of North America," 24(4), pp. 697-715, 2006.

[3] M. S. Manikandan, S. Dandapat, "Wavelet energy based diagnostic distortion measure for ECG," Biomedical Signal Processing and Control, 2(2), pp. 80-96, 2007.

[4] E. J. Benjamin, P. Muntner, A. Alonso, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, A. R. Chang, S. Cheng, S. R. Das, F. N. Delling, "heart disease and stroke statistics—2019 update: a report from the American Heart Association," Circulation, 139(10), pp. 56-28, 2019.

[5] T. A. Gaziano, A. Bitton, S. Anand, S. Abrahams-Gessel, A. Murphy, "Growing epidemic of coronary heart disease in low-and middle-income countries," Current problems in cardiology, 35(2), pp. 72-115, 2010.

[6] U. R. Acharya, H. Fujita, S. L. Oh, Y. Hagiwara, J. H. Tan, M. Adam, "Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals," Information Sciences, 415, pp. 190-8, 2017.

[7] U. R. Acharya, H. Fujita, O. S. Lih, Y. Hagiwara, J. H. Tan, M. Adam, "Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network," Information sciences, 405, pp. 81-90, 2017.

[8] M. Kumar, R. B. Pachori, U. R. Acharya, "Automated diagnosis of myocardial infarction ECG signals using sample entropy in flexible analytic wavelet transform framework," Entropy, 19(9), pp. 488, 2017.

[9] J. H. Tan, Y. Hagiwara, W. Pang, I. Lim, S. L. Oh, M. Adam, R. San Tan, M. Chen, U. R. Acharya, "Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals," Computers in biology and medicine, 94, pp. 19-26, 2018.

[10] S. W. Chen, S. L. Wang, X. Z. Qi, S. M. Samuri, C. Yang, "Review of ECG detection and classification based on deep learning: Coherent taxonomy, motivation, open challenges and recommendations," Biomedical Signal Processing and Control, 74, pp. 103493, 2022.

[11] F. H. Fenton, E. M. Cherry, L. Glass, "Cardiac arrhythmia," Scholarpedia, 3(7), pp. 1665, 2008.

[12] S. Rosara, M. Borgarelli, M. Perego, J. Häggström, G. La Rosa, A. Tarducci, R. A. Santilli, "Holter monitoring in 36 dogs with myxomatous mitral valve disease," Australian veterinary journal, 88(10), pp. 386-92, 2010.

[13] R. Tung, N. G. Boyle, K. Shivkumar, "Catheter ablation of ventricular tachycardia," Circulation, 122(3), pp. 389-91, 2010.

[14] S. H. Jambukia, V. K. Dabhi, H. B. Prajapati, "Classification of ECG signals using machine learning techniques: A survey," In: Proc. International Conference on Advances in Computer Engineering and Applications, 2015.

[15] S. Celin, K. Vasanth, "ECG signal classification using various machine learning techniques," Journal of medical systems, 42(12), pp. 241, 2018.

[16] M. G. Shankar, C. G. Babu, "An exploration of ECG signal feature selection and classification using machine learning techniques," Int. J. Innovative Technol. Exploring Eng. Regul, 9(3), pp. 797-804, 2020.

[17] M. Alfaras, M. C. Soriano, S. Ortín, "A fast machine learning model for ECG-based heartbeat classification and arrhythmia detection," Frontiers in Physics, pp. 103, 2019.

[18] L. Zhang, H. Peng, C. Yu, "An approach for ECG classification based on wavelet feature extraction and decision tree," In: Proc. International conference on wireless communications & signal processing (WCSP), 2010.

[19] Y. C. Yeh, W. J. Wang, C. W. Chiou, "Cardiac arrhythmia diagnosis method using linear discriminant analysis on ECG signals," Measurement, 42(5), pp. 778-89, 2009.

[20] M. A. Escalona-Morán, M. C. Soriano, I. Fischer, C. R. Mirasso, "Electrocardiogram classification using reservoir computing with logistic regression," IEEE Journal of Biomedical and health Informatics, 19(3), pp. 892-8, 2014.

[21] S. Padmavathi, E. Ramanujam, "Naïve Bayes classifier for ECG abnormalities using multivariate maximal time series motif," Procedia Computer Science, 47, pp. 222-8, 2015.

[22] R. Saini, N. Bindal, P. Bansal, "Classification of heart diseases from ECG signals using wavelet transform and kNN classifier," In: Proc. International Conference on Computing, Communication & Automation, 2015.

[23] Q. Zhao, L. Zhang, "ECG feature extraction and classification using wavelet transform and support vector machines," In: Proc. International Conference on Neural Networks and Brain, 2015.

[24] Y. Ozbay, B. Karlik, "A recognition of ECG arrhytihemias using artificial neural networks," In: Proc. Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2001.

[25] M. E. A. Bourkha, A. Hatim, D. Nasir, S. E. Beid, "Enhanced Atrial Fibrillation Detection-based Wavelet Scattering Transform with Time Window Selection and Neural Network Integration," International Journal of Advanced Computer Science and Applications (IJACSA), 14(12), 2023.

[26] R. G. Afkhami, G. Azarnia, M. A. Tinati, "Cardiac arrhythmia classification using statistical and mixture modeling features of ECG signals," Pattern Recognition Letters, 70, pp. 45-51, 2016.

[27] C. H. Lin, "Frequency-domain features for ECG beat discrimination using grey relational analysis-based classifier," Computers & Mathematics with Applications, 55(4), pp. 680-90, 2008.

[28] M. Kropf, D. Hayn, G. Schreier, "ECG classification based on time and frequency domain features using random forests," In: Proc. Computing in Cardiology (CinC), 2017.

[29] S. Banerjee, M. Mitra, "Application of cross wavelet transform for ECG pattern analysis and classification," IEEE transactions on instrumentation and measurement, 63(2), pp. 326-33, 2013.

[30] T. Li, M. Zhou, "ECG classification using wavelet packet entropy and random forests," Entropy, 18(8), pp. 285, 2016.

[31] J. Huang, B. Chen, B. Yao, W. He, "ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network," IEEE access, 7, pp. 92871-80, 2019.

[32] X. Zhai, C. Tin, "Automated ECG classification using dual heartbeat coupling based on convolutional neural network," IEEE Access, 6, pp. 27465-72, 2018.

[33] S. Singh, S. K. Pandey, U. Pawar, R. R. Janghel, "Classification of ECG arrhythmia using recurrent neural networks," Procedia computer science, 132, pp. 1290-7, 2018.

[34] S. Saadatnejad, M. Oveisi, M. Hashemi, "LSTM-based ECG classification for continuous monitoring on personal wearable devices,"

IEEE journal of biomedical and health informatics, 24(2), pp. 515-23, 2019.

[35] I. Silva, G. B. Moody, "An open-source toolbox for analysing and processing physionet databases in matlab and octave," Journal of open research software, 2(1), 2014.

[36] C. T. Arsene, R. Hankins, H. Yin, "Deep learning models for denoising ECG signals," In: Proc. 27th European Signal Processing Conference (EUSIPCO), 2019.

[37] M. A. Ozdemir, O. Guren, O. K. Cura, A. Akan, A. Onan, "Abnormal ecg beat detection based on convolutional neural networks," In: Proc. Medical technologies congress (TIPTEKNO), 2020.

[38] L. Y. Di Marco, W. Duan, M. Bojarnejad, D. Zheng, S. King, M. Murray, P. Langley, "Evaluation of an algorithm based on single-condition decision rules for binary classification of 12-lead ambulatory ECG recording quality," Physiological measurement, 33(9), pp. 1435, 2012.

[39] A. Çınar, S. A. Tuncer, "Classification of normal sinus rhythm, abnormal arrhythmia and congestive heart failure ECG signals using LSTM and hybrid CNN-SVM deep neural networks," Computer methods in biomechanics and biomedical engineering, 24(2), pp. 203-14, 2021.

[40] S. Nahak, A. Pathak, G. Saha, "Evaluation of handcrafted features and learned representations for the classification of arrhythmia and congestive heart failure in ECG," Biomedical Signal Processing and Control, 79, pp. 104230, 2023.

[41] M. E. A. Bourkha, A. Hatim, D. Nasir, S. E. Beid., & A. S. Tahiri, "A Novel Inter Patient ECG Arrhythmia Classification Approach with Deep Feature Extraction and 1D Convolutional Neural Network," International Journal of Advanced Computer Science and Applications (IJACSA), 15(2), 2024.

[42] J. Wang, X. Qiao, C. Liu, X. Wang, Y. Liu, L. Yao, H. Zhang, "Automated ECG classification using a non-local convolutional block attention module," Computer Methods and Programs in Biomedicine, 203, pp. 106006, 2021.

[43] S. Osowski, L. T. Hoai, T. Markiewicz, "Support vector machine-based expert system for reliable heartbeat recognition," IEEE transactions on biomedical engineering, 51(4), pp. 582-9, 2004.

[44] J. Rodriguez, A. Goni, A. Illarramendi, "Real-time classification of ECGs on a PDA," IEEE Transactions on information Technology in biomedicine, 9(1), pp. 23-34, 2005.

[45] S. Chen, W. Hua, Z. Li, J. Li, X. Gao, "Heartbeat classification using projected and dynamic features of ECG signal," Biomedical Signal Processing and Control, 31, pp. 165-73, 2017.

[46] E. Ihsanto, K. Ramli, D. Sudiana, "Real-time classification for cardiac arrhythmia ECG beat," In: Proc. 16th International Conference on Quality in Research (QIR): International Symposium on Electrical and Computer Engineering, 2019.

[47] F. Melgani, Y. Bazi, "Classification of electrocardiogram signals with support vector machines and particle swarm optimization," IEEE transactions on information technology in biomedicine, 12(5), pp. 667-77, 2008.

[48] A. M. Alqudah, A. Alqudah, "Deep learning for single-lead ECG beat arrhythmia-type detection using novel iris spectrogram representation," Soft Computing, 26(3), pp. 1123-39, 2022.

[49] A. M. Alqudah, S. Qazan, L. Al-Ebbini, H. Alquran, I. A. Qasmieh, "ECG heartbeat arrhythmias classification: A comparison study between different types of spectrum representation and convolutional neural networks architectures," Journal of Ambient Intelligence and Humanized Computing, pp. 1-31, 2021.

[50] A. Rajkumar, M. Ganesan, R. Lavanya, "Arrhythmia classification on ECG using Deep Learning," In: Proc. 5th international conference on advanced computing & communication systems (ICACCS), 2019.

[51] A. M. Shaker, M. Tantawi, H. A. Shedeed, M. F. Tolba MF, "Generalization of convolutional neural networks for ECG classification using generative adversarial networks," IEEE access, 8, pp. 35592-605, 2020.

[52] M. Ramkumar, C. G. Babu, K. V. Kumar, D. Hepsiba, A. Manjunathan, R. S. Kumar, "ECG cardiac arrhythmias classification using DWT, ICA and MLP neural networks," In: Proc. Journal of Physics: Conference Series, 2012.

[53] N. N. Arslan, D. Ozdemir, H. Temurtas, "ECG heartbeats classification with dilated convolutional autoencoder," Signal, Image and Video Processing, 18(1), pp. 417-426, 2024.

[54] R. Vavekanand, K. Sam, S. Kumar, T. Kumar, "CardiacNet: A Neural Networks Based Heartbeat Classifications using ECG Signals," Studies in Medical and Health Sciences, 1(2), pp. 1-17, 2024.

[55] F. Zhou, D. Fang, "Multimodal ECG heartbeat classification method based on a convolutional neural network embedded with FCA," Scientific Reports, 14(1), pp. 8804, 2024.

[56] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," Circulation, 101(23), pp. 215-20, 2000.

[57] G. B. Moody, R. G. Mark, "The impact of the MIT-BIH arrhythmia database," IEEE engineering in medicine and biology magazine, 20(3), pp. 45-50, 2001.

[58] J. Pan, W. J. Tompkins, "A real-time QRS detection algorithm," IEEE transactions on biomedical engineering, pp. 230-6, 1985.

[59] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research, 16, pp. 321-57, 2002.

[60] J. Bruna, S. Mallat, "Invariant scattering convolution networks," IEEE transactions on pattern analysis and machine intelligence, 35(8), pp. 1872-86, 2013.

[61] R. Leonarduzzi, H. Liu, Y. Wang, "Scattering transform and sparse linear classifiers for art authentication," Signal Processing, 150, pp. 11-9, 2018.

[62] B. S. Manjunath, "Gabor wavelet transform and application to problems in early vision," In: Proc. Asilomar conference on signals systems and computers, 1992.

[63] X. Jin, E. W. Ma, L. L. Cheng, M. Pecht, "Health monitoring of cooling fans based on Mahalanobis distance with mRMR feature selection," IEEE Transactions on Instrumentation and Measurement, 61(8), pp. 2222-9, 2012.

[64] C. Ding, H. Peng, "Minimum redundancy feature selection from microarray gene expression data," Journal of bioinformatics and computational biology, 3(02), pp. 185-205, 2005.

Personalized Motion Scheme Generation System Design for Motion Software Based on Cloud Computing

Jinkai Duan
Department of Public Courses, Inner Mongolia Technical College of Construction, Hohhot, 010070, China

***Abstract*—The increase of national attention has also promoted the growth of the scale of sports health industry. However, for ordinary people who lack professional knowledge, intuitive data cannot make correct sports planning. Therefore, aiming at the problem that it is difficult for ordinary people to make correct exercise plan according to intuitive data, a personalized exercise plan generation system based on cloud computing is proposed. By analyzing the user's movement and physical data, the system uses cloud computing resources and machine learning algorithms to provide customized exercise recommendations for users. The key innovation of the research is the combination of improved random forest algorithm and reinforcement learning, while improving the performance of the algorithm on unbalanced sample sets. The results indicated that the accuracy of the improved random forest was 0.985 higher than that of the precision weighted random forest. The research algorithm was 9.04% higher on average than the original random forest algorithm and 2.71% higher than the accuracy weighted random forest algorithm. In terms of the accuracy of personalized motion scheme generation of motion software, the improved algorithm reached 95.05% at most, and its recall rate reached 83.46% at most. Compared with the existing sports software solutions, the research system can generate personalized sports programs more accurately, promote the development of the sports health industry and improve the national physical health level. The system can provide users with personalized sports suggestions, and utilize the powerful computing power of cloud computing to realize real-time processing and analysis of large-scale user data, providing users with timely sports feedback and suggestions.**

Keywords*—Cloud computing; sports; random forest algorithm; personalization; system

I. Introduction

Nowadays, China's economic level is constantly improving, but the increasing cost of living has also had a negative impact on the physical health of its citizens [1]. Physical exercise can maintain the daily health of human body. Studies have shown that taking appropriate exercise methods and maintaining appropriate amount of exercise can improve health level [2]. The increase of people's attention has also promoted the growth of the scale of the sports health industry. At present, the domestic outdoor sports industry market has great room for growth. At the same time, with the deep popularization of the Internet, China's online sports health market is also growing. The huge user base brings a massive amount of information data. Therefore, how to mine the user characteristic information hidden in these data, analyze the user's body development trend, and provide more reasonable suggestions is the key research direction of sports health application in the intelligent era. Zheng W et al. proposed a programming practice recommendation algorithm based on knowledge structure tree (KSTER). This algorithm was predicated on the quantification of students' cognitive level and knowledge needs, the construction of a KSTER, and the combination of a learning goal prediction method to realize personalized programming practice recommendation. Experiments demonstrated the superiority of this algorithm in terms of precision and recall ratio, as well as its effective improvement of students' learning efficiency [3]. Netz Y et al. evaluated a new tool that uses smartphone sensors to remotely assess and deliver personalized exercise plans. Fifty-two healthy older volunteers participated in the test, which showed significant improvements in strength, flexibility, and balance. Research confirmed the potential of this tool to deliver personalized exercise programs in a home setting [4]. Due to the difference in human body constitution, everyone has a large difference in the exercise methods and amount of exercise. Similar to medical consultation, customized exercise recommendations based on the user's physical characteristics have formed the concept of exercise prescription. Cloud computing integrates computing and storage into a scalable resource pool to improve resource utilization [5]. This study aims to overcome the shortcomings of the random forest (RF) algorithm and improve the classification performance and accuracy of unbalanced sample sets by introducing weighting factors and optimizing the voting mechanism based on the traditional RF algorithm. By collecting users' movement and physical sign data, cloud computing and machine learning algorithms are used to generate a sports health cloud platform for sports programs, and promote the concept of sports health. The innovation of research methods is mainly reflected in five aspects. First, cloud computing integration. The cloud computing platform used integrates computing and storage resources, improving the processing capability and real-time feedback effect of large-scale data. Second, the RF is optimized. The traditional RF algorithm is improved by adding weight factors to improve the classification performance on unbalanced data sets. Third, reinforcement learning application. Reinforcement learning strategies are used to optimize exercise plan generation, dynamically adjust plans based on user feedback, and improve user experience. Fourth, it provides a personalized exercise plan based on the user's physique and goals, and realizes an innovative breakthrough in exercise software. Fifth, compared with traditional algorithms, the advantages of the improved algorithm in accuracy rate, recall

rate and F1 score are verified.

The research mainly includes six sections. Section I introduces the background and significance of the research on cloud platform and intelligent motion system. Section II summarizes the intelligent sports software, mainly for the detailed analysis of the achievements of the current domestic and foreign experts and scholars in the direction of sports health assistance system design. Section III is the research method, which is mainly divided into two sections. In subsection one, the research proposes the software scheme system design. In subsection two, the research proposes a personalized motion scheme generation system for motion software based on cloud computing and improved RF algorithm. Section IV is the performance analysis of the personalized motion scheme generation system of sports software. Discussion is given in Section V. Section VI summarizes the research methods and results analysis. At the same time, the shortcomings of research methods and future research directions are proposed.

II. Related Work

Sports can maintain the daily health of human body. A large number of studies have shown that taking appropriate exercise can improve people's health level. Yang C and others proposed to monitor sports energy consumption through cloud services and the Internet of Things and establish an effective data model for difficult operation and implementation of the method of detecting sports energy consumption. Research showed that the detection error of this method for sports energy consumption was less than 2% [6]. In order to promote the clinical application of metabolomics, Castelli F. et al. addressed problems such as insufficient standardization of data generation, low metabolite recognition rate, and complex data processing. These challenges affected the interoperability and reusability of data. In addition, metabolomics outputs were complex and expensive, and features needed to be simplified for on-site applications [7]. Xie D et al. designed a wearable energy-saving fitness tracking system for health monitoring of athletes based on the integrated of deep learning. The algorithm used smart phone applications to track the steps of using Internet of Things technology. Research showed that energy efficiency system performance was improved [8]. Virtual reality technology integrated digital image processing, computer, artificial intelligence, multimedia, sensors, and other information technologies to provide support for the creation and experience of virtual worlds. Its combination with cultural relic protection created a new field. Ling Z et al. explored the personalized health care pavilion display system based on VR and deep learning. A new computational model was designed to improve the system performance. Its effect was verified through application scenario testing [9].

To understand the cognition of high school physical education teachers to the rules of basketball competition, Saputro A et al. used data collection technology and descriptive research methods to investigate the high school physical education teachers in their local area. The results showed that most physical education teachers had a higher cognition of basketball game rules [10]. Men Y et al. proposed a motion recognition algorithm to improve prediction efficiency. A database model was constructed to analyze sports training and competitions. The results showed that this research improved the prediction performance of vicious energy motion [11]. Wang L et al. established the SVM model. Based on texture feature extraction method and in-depth analysis of non-downsampling contour wave transformation, an intelligent motion feature recognition system was constructed. The research showed that the system effectively and accurately extracted and recognized motion features [12]. Sun C M D et al. introduced time control factor to design classification strategy when using support vector machine for classification. Experiments showed that the research method achieved fast target discrimination when the ambient brightness changed, and improved recognition accuracy of complex human posture [13]. Wu B et al. proposed a motion assistant system to identify, query and analyze motion features through deep learning, aiming at the low detection performance of target detection technology in complex background. The experiment showed that the system had good performance in recognizing and classifying human motion behavior characteristics [14].

To sum up, the current research and market application of sports health applications lack effective user data. The application research of data algorithm in sports health utilizes the data uploaded by the sports health monitoring system. However, with the popularity of intelligent sports devices such as smart bracelets, the scale of users' motion data is growing. Cloud computing has powerful computing power, storage capacity and other advantages. Therefore, this study combines cloud computing with machine learning to build a cloud platform. The system uses the improved RF algorithm to solve the unbalance of sample data. Second, the powerful processing power of cloud computing is used to analyze the user's motion data in real time and provide personalized suggestions through machine learning algorithms to ensure the accuracy of the motion scheme.

III. Personalized Motion Scheme Generation System Design and Research for Motion Software Based on Cloud Computing

In this study, an improved RF algorithm is used to generate a personalized motion scheme based on the user's motion and physical data. It comprehensively considers individual differences such as the user's age, health status and medical background. The machine learning classification algorithm is used to accurately identify individual characteristics through a large number of physical test data. The improved algorithm uses a weighted voting mechanism to optimize the classification effect of the decision tree, thus improving the performance of the model when dealing with unbalanced sample sets. The system has adaptive capabilities and can process user data in real time. The high-performance computing capabilities of cloud computing platforms can also be used to dynamically adjust exercise suggestions to ensure personalized and adaptive solutions. Through intelligent analysis and real-time feedback, the system is able to provide users with highly personalized exercise plans that take into account individual differences.

A. Software System Scheme Design

In traditional enterprise software development, computer hardware infrastructure is an important part. Some small and

medium-sized enterprises need to purchase physical machine servers with high cost from server manufacturers [15]. However, large software enterprises or server manufacturers have the ability to manufacture large-scale physical machine servers. Therefore, computing and storage devices will be redundant or idle [16]. Cloud computing provides computing resources over the Internet, giving users remote access to data storage, computing power, and applications. Cloud service types typically include infrastructure as a service (IaaS), platform as a Service (PaaS), and software as a service (SaaS). IaaS provides virtualized hardware resources. PaaS supports application development and management. SaaS provides standardized application services. Major cloud providers such as AWS and Microsoft Azure offer a wide range of services for applications such as artificial intelligence, the Internet of Things. To ensure data security and privacy, these services use measures such as data backup, encrypted transmission, and access control, and comply with regulations such as GDPR. Cloud computing not only improves data analysis capabilities, but also provides personalized solutions while ensuring compliance. In terms of usage mode, the cloud computing service users use is to obtain the service provider's storage hardware, underlying network, storage space, and other resources through remote technology. Fig. 1 shows the conceptual model of cloud computing [17].

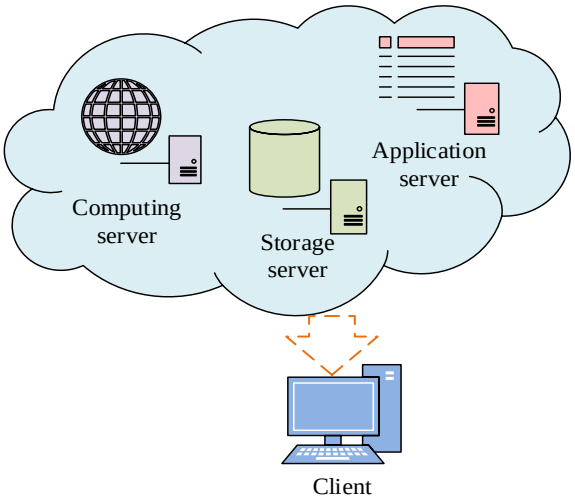


Fig. 1. Conceptual model of cloud computing.

In Fig. 1, cloud computing includes two parts, cloud and terminal. Cloud refers to the collection of network, storage, service and security devices. Cloud boundary is determined by the size of various hardware devices. Terminals refer to terminal devices that can connect to the cloud and help users access and use cloud computing services, as well as various intelligent devices with networking capabilities. The architecture of cloud computing system has three layers, namely core service, service management and user access [18]. The role of the core service is to virtualize the software and hardware so that user services have better performance in terms of reliability, availability, and other aspects to meet the diverse high-quality service needs of users. The service management layer guarantees the stability, reliability and high quality of the cloud platform during operation. The user access function is to provide the user access interface. The cloud platform mainly includes cloud server, database and other components. The main function of the server is to communicate with the client and perform business functions. The research platform is to accurately generate motion schemes for users, and there are many commonly used classification algorithms. RF is is mainly used for regression and classification. The algorithm can freely choose the type of decision tree, and investigates the use of classification and regression tree (CART) as the generation model of a single RF classifier. CART uses GINI as feature selection index, as shown in Formula (1) [19].

$$GINI(P)=\sum_{i=1}^{k} p_i(1-p_i) \tag{1}$$

In Formula (1), k represents the number of samples classified. p_i is the probability that the sample belongs to this category. For a given sample set D, the expression is shown in Formula (2).

$$GINI(D)=1-\sum_{k=1}^{k}\left(\frac{|d_k|}{D}\right)^2 \tag{2}$$

In Formula (2), d_k represents the classification of samples in D. CART is a binary tree. If the training set is divided into two categories because of A attributes, the expression of GINI is shown in Formula (3).

$$GINI(D,A)=\frac{D_1}{D}GINI(D_1)+\frac{D_2}{D}GINI(D_2) \tag{3}$$

The classification result of a RF is determined by the classification pattern of all decision trees, as shown in Formula (4).

$$I(x)=\arg\max_{Y}\sum_{i=1}^{k} I(h_i(x)=y) \tag{4}$$

In Formula (4), y represents the classification category. $h_i(x)$ is the tree of decision tree. Y represents the classification result of a single decision tree. The cloud computing platform is a SaaS service. Its main function is to obtain, store and process data such as web pages and intelligent terminals through servers in Alibaba Cloud [20]. Cloud platform system architecture includes user, data transmission, control, business logic, and data. The user layer is a channel for human-computer interaction and data upload. App and web pages can also be used as human-computer interaction terminals. The app can accept the data uploaded by the body feature sensor, while the web can allow users to upload manually entered data. Data transmission transmits data through the Internet. Control uses Nginx as the front-end server for handling high load access. The business logic is the core platform structure layer, which is realized by running the code of each business function module through the Tomcat server cluster. Data function response operation requirements, mainly

through the Redis database and MySQL database, storing physique data, motion data, and scheme library. The hybrid replication is used to achieve data synchronization in the master and slave databases, as shown in Fig. 2.

In Fig. 2, first of all, the main node writes the corresponding SQL statement to the binary text before each transaction completes data update. After the transaction is completed, the main node submits the transaction through the database storage engine. Then, the slave node is connected to the master node by the I/O process, and the slave node determines the initial location of the required binary code. After the master node accepts the request, it parses the request from the slave node. Then, the log message after reading the binary file at that location according to the bit given in the request is published in the I/O process of the node. The main contents of the log information are the file name and the end position of the file. This position is defaulted to the start position of the next request. After the master node returns the data to the slave node, the slave node adds the content to the end of the relay file in order. Meanwhile, the returned file description information is added to the master info file. The slave node detects the newly added files through the SQL thread, and processes the SQL statements of the corresponding nodes, thus realizing the master slave data replication.

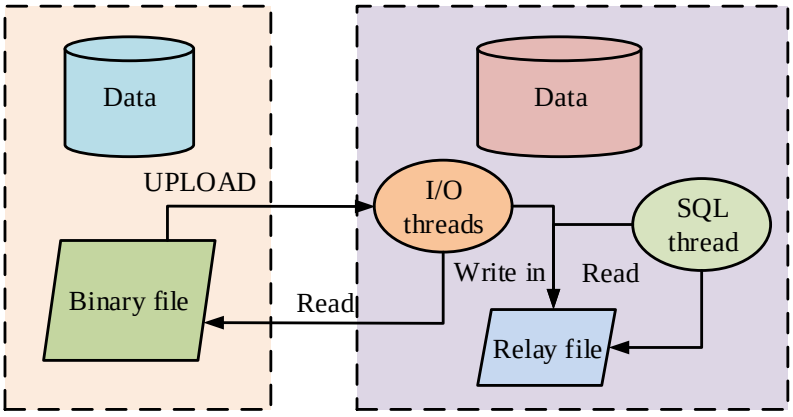


Fig. 2. Master / Slave synchronization execution process.

B. *Design of Personalized Motion Scheme Generation System for Motion Software Based on Cloud Computing and Improved Random Forest Algorithm*

Exercise is an effective way to keep healthy. However, exercise method and amount suitable for everyone are different. Therefore, the cloud computing platform provides users with reasonable online training programs and plays an auxiliary role in sports and fitness. Although the RF classification effect and generalization ability are good, the algorithm also shows an over-fitting phenomenon with too much noise. In addition, the performance of the algorithm under unbalanced sample sets needs to be improved. Therefore, the research improves RF shortcomings, and uses the improved algorithm as the core function of the cloud platform motion scheme generation algorithm. The research mainly improves RF from two aspects: improving the voting mechanism and improving the performance of the RF algorithm in unbalanced sample set. For RF, the input value is set to X and the output value is set to Y. Then, there is a definable interval function $mg(X,Y)$, as shown in Formula (5) [21].

$$mg\left(X,Y\right)=P_{\theta}\left(h\left(X,\theta\right)=Y\right)-\max_{j\neq Y}P_{\theta}\left(h\left(X,\theta\right)=j\right) \tag{5}$$

In formula (5), θ represents the random input value formed in the random selection of training sets from a single subtree in the forest. $h\left(X,\theta\right)$ represents the output value. RF classification intensity is the expected value of the interval function. The calculation expression is shown in Formula (6).

$$S=E_{X,Y}\left(mg\left(X,Y\right)\right) \tag{6}$$

The generalization error of RF is shown in Formula (7).

$$PE^{*}\leq\overline{\rho}\left(1-S^{2}\right)/S^{2} \tag{7}$$

In formula (7), S indicates the classification strength. $\overline{\rho}$ represents the average similarity between decision trees. If the performance of RF is to be improved, its error should be reduced. The classification result function of RF is improved by introducing weight factor, as shown in Formula (8).

$$I\left(x\right)=\arg\max_{Y}\sum_{i=1}^{k}W_{i}*I\left(h_{i}\left(x\right)=y\right) \tag{8}$$

The probability of RF classification for any result α is shown in Formula (9).

$$P_{\theta}\left(h\left(X,\theta\right)=\alpha\right)=\frac{W_{\alpha}}{W_{t}} \tag{9}$$

In Formula (9), W_{α} is the total weight of the α decision tree. W_{t} represents the total weight of all decision trees in the forest. The interval function of RF can be obtained, as shown in Formula (10).

$$mg(X,Y)=\frac{W_Y}{W_t}-\frac{\arg\max_{j\neq Y} W_j}{W_t}=\frac{W_Y}{W_t}-\frac{W_\beta}{W_t} \tag{10}$$

It is assumed that the tree weights of the same classification result are the same, then Formula (11) can be obtained.

$$w_Y * N_Y > w_\beta * N_\beta \tag{11}$$

In Formula (11), N_Y represents the number of decision trees with classification results. Y represents the decision tree classified as the weight of other results and the maximum result. If the formula does not hold, formula (12) is obtained.

$$\frac{N_Y}{N_\beta}>\frac{w_\beta}{w_Y} \tag{12}$$

When the voting mode of RF is equal voting, there is $w_Y = w_\beta$, as shown in Formula (13).

$$\frac{N_Y}{N_\beta}>1 \tag{13}$$

When voting with equal rights, $N_Y > N_\beta$ can get the correct classification results. When different weights are used, if the weight of the incorrectly classified tree is less than that of the correctly classified tree, Formula (12) exists.

$$\frac{N_Y}{N_\beta}>\lambda, \lambda<1 \tag{14}$$

In Formula (12), in classification, the number of trees with correct classification results can be less than the number of trees with wrong classification results. Meanwhile, compared with equal voting, when correct and wrong trees in the classification result remains the same, the weight of the correct tree in RF classification is higher, indicating that the weight of the incorrect tree is smaller. Good classification performance indicates smaller generalization error. Fig. 3 shows the steps of the improved RF algorithm after adding weight factors.

In Fig. 3, the improved RF algorithm uses test set pairs to test after the forest is constructed according to the original algorithm flow. The AUC value of each decision tree in RF is obtained and used as the corresponding weight value of classification result of each subtree. Finally, the classification results are obtained by weighted voting. Formula (15) displays the classification results of the improved RF.

$$I(x)=\arg\max_Y \sum_{i=1}^{k} weight_{AUC}(i) * I(h_i(x)=y) \tag{15}$$

To satisfy high traffic, stability and other access, a sports health cloud platform is built on Alibaba Cloud using frameworks and component technologies such as Springboot, React, Nginx, MySQL, and Redis. The database module of the platform is very critical, mainly composed of Redis database and MySQL database. Its main function is to provide data storage, query, modification and other operations. Fig. 4 displays the workflow of the database module.

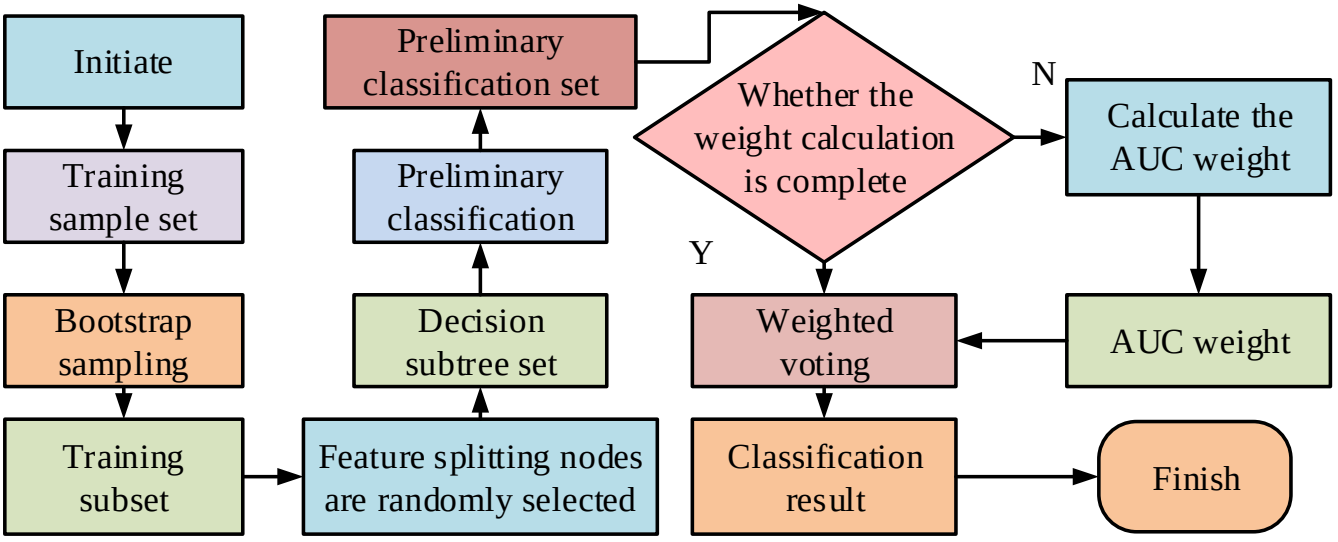


Fig. 3. Steps of the improved RF algorithm.

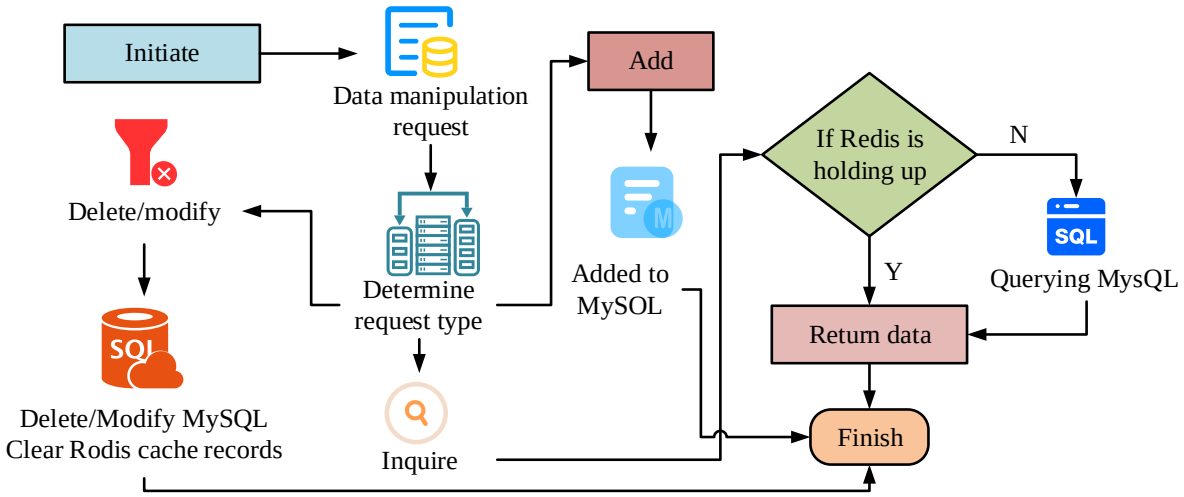


Fig. 4. Workflow of the database module.

In Fig. 4, in the case of high concurrent connections, since MySQL is a traditional relational database, all operations are completed on the hard disk. Therefore, the efficiency of IO is very low, and the database must be used as the buffer. The workflow of the database module is as follows. First, when the user requests a database query, ask whether there is corresponding data in the Redis database. If it exists, it will return directly. If it does not exist, it will query the MySQL database and cache the Redis database after returning data. When the user requests to add an operation, it directly executes the add task in the MySQL database. When the client performs delete and update operations, the MySQL database records are updated directly. If there is a corresponding cache in the Redis database, it is deleted directly. The user's current prescription display module is used to display the user's current exercise prescription. Fig. 5 displays the workflow of the current scheme.

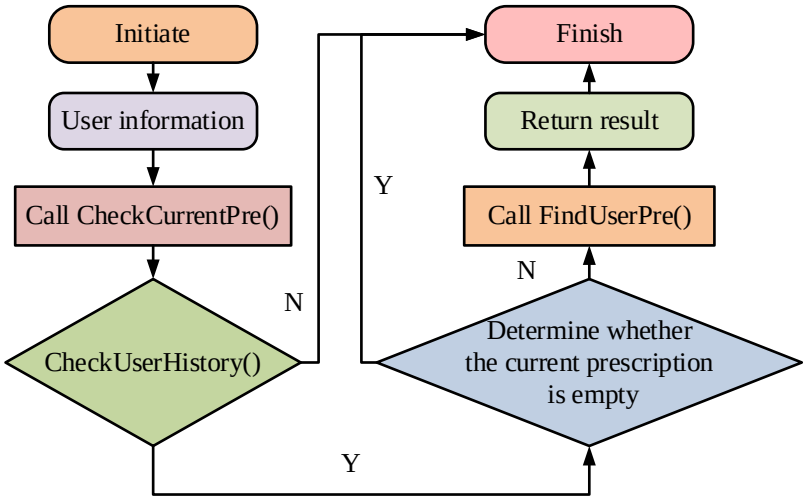


Fig. 5. Workflow in the current scenario.

In Fig. 5, the user only needs to press the current scheme key to submit the request and user information to the platform. After receiving the user's request and information, the platform calls the overall packaging function of *CheckCurrentPre*( ). First, the function *CheckUserHistory*() is called to check whether the user's history file exists. If it exists, whether the ID number of the ID item exe_ Pre_ is empty is queried. If it does not exist or exe_ pre_ id item is empty, the query failure information is returned. If the user record exists and exe_ pre_ id item is not empty, the FindUserPre function is called. According to the motion scheme ID number, the corresponding content is retrieved from the exercise plan library and sent back to the homepage for display.

IV. Performance Experiment Analysis of Personalized Motion Scheme Generation System of Sports Software

The personalized motion scheme generation system uses an enhanced RF algorithm, combines cloud computing resources and machine learning technology. The system analyzes the user's motion and physical data to provide personalized motion recommendations through improved RF and reinforcement learning. The key innovation is the enhancement of the RF algorithm to improve the performance of unbalanced sample sets. The system is trained and verified using a dataset of physical fitness tests in colleges and universities, including multiple test data from male and female students. The cloud

platform is implemented by Springboot, React and other technologies. The database module is composed of Redis and MySQL, which supports high concurrency and data caching to ensure low latency and high scalability of the system. The constructed platform is to accurately generate personalized sports programs for users. In order to complete the exercise program database, the experiment collected the physical test data of a university from 2017 to 2022, with a total of 18162 items. There are 9064 for boys and 9098 for girls. Because the data size of this study cannot meet the requirements of deep neural network, the algorithms used to achieve the core functions in the experiment are mainly compared and selected in the traditional machine learning classification algorithms. The performance comparison between RF and other classification algorithms is shown in Fig. 6.

In Fig. 6, in terms of accuracy, RF algorithm value is higher than other four algorithms, which is 87.2%. In terms of recall rate, the ROC value of RF algorithm is still higher than traditional classification methods, which is 0.957. In terms of F1 score, the RF algorithm is 0.861. This algorithm still has a higher F1 score than traditional classification methods such as KNN, SVM, ANN, and decision trees. In conclusion, the RF algorithm has better classification effect than traditional classification algorithms. The experimental test environment is Python 3.6+Sklearm, Intel (R) i5-6300HQCPU@2 Platform. The experimental data are divided into balanced sample set and unbalanced sample set. The specific sample set is shown below.

The total number of samples is 20, with 3 imbalanced negative samples and 10 balanced samples in the sample set. As shown in Fig. 7, the AUC value represents the classification performance under different balance levels of sample sets.

From Fig. 7 (a), when the number of negative categories mistakenly classified as positive categories increases, the AUC value decreases. Therefore, the AUC value has a great impact on the classification effect under the unbalanced sample size. In Fig. 7 (b), AUC decreases as the number of negative classes wrongly classified into positive classes increases. To sum up, AUC value can be used to evaluate the classification effect of algorithms under both balanced and unbalanced sample sets. In addition, it can also be used as a weight factor in the voting process of the improved RF. Figure 8 displays the influence result of RF subtrees.

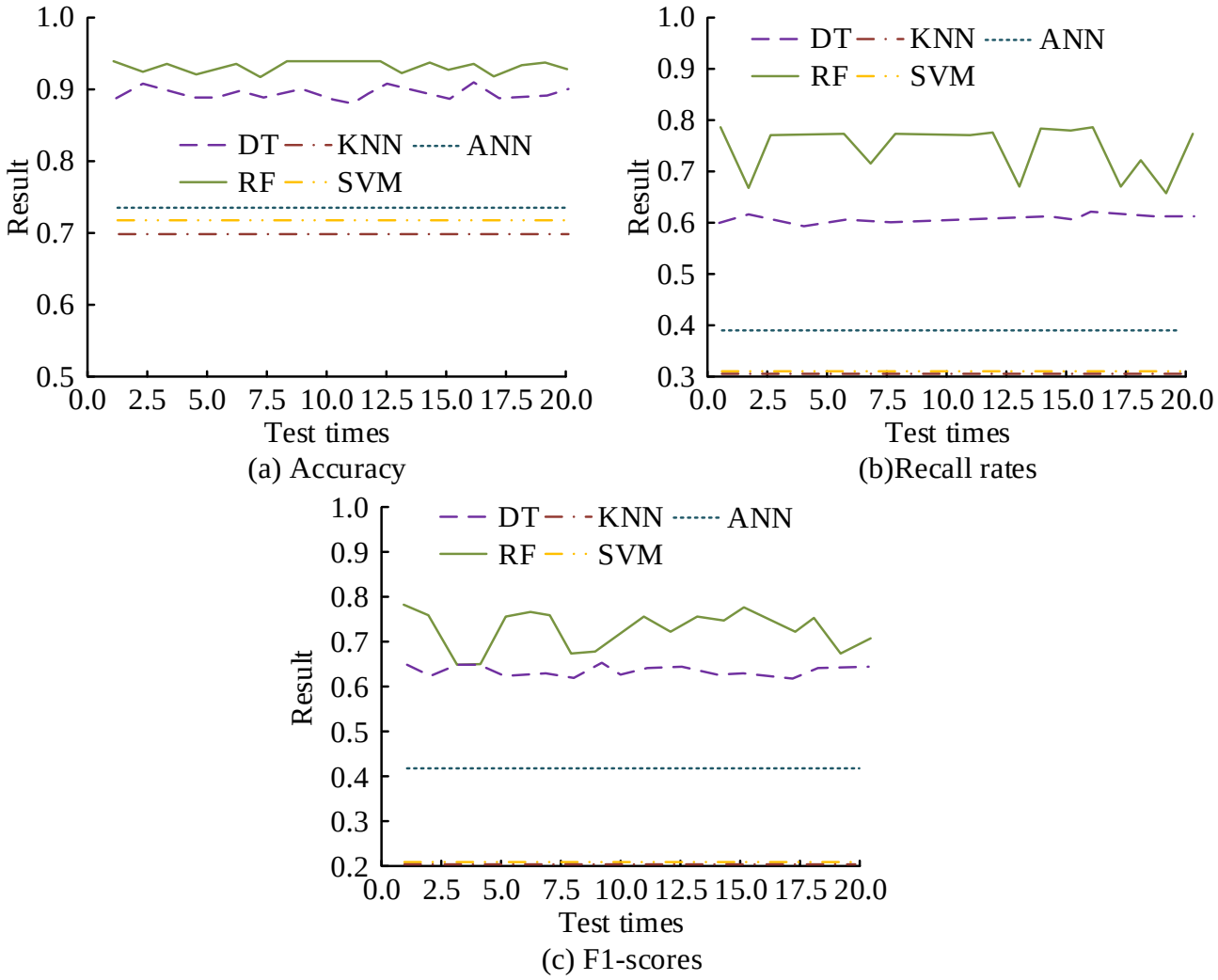
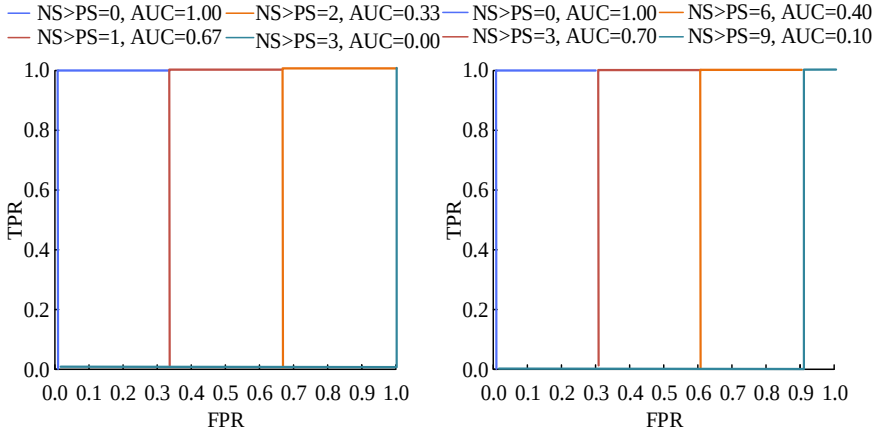


Fig. 6. Performance comparison results of five classification algorithms.

TABLE I. EXPERIMENTAL SAMPLE SET

Sample set		Sample size	Characteristic dimension	Positive sample proposal
Balanced sample set	Banana	5300	2	55.2
	data_banknote_authentication	1371	5	55.5
	Spambase	4600	57	60.6
	Vehicle	846	18	50
Unbal-anced sample set	Ecoli3	336	35	89.6
	Yeast3	1484	8	71.4
	Abalone	580	8	97.7
	Poker_8_9	1459	10	98.3



(a) AUC value of classification performance change under unbalanced sample set

(b) AUC value of classification performance change under balanced sample set

Fig. 7. AUC values representing classification performance in sample sets with different levels of balance.

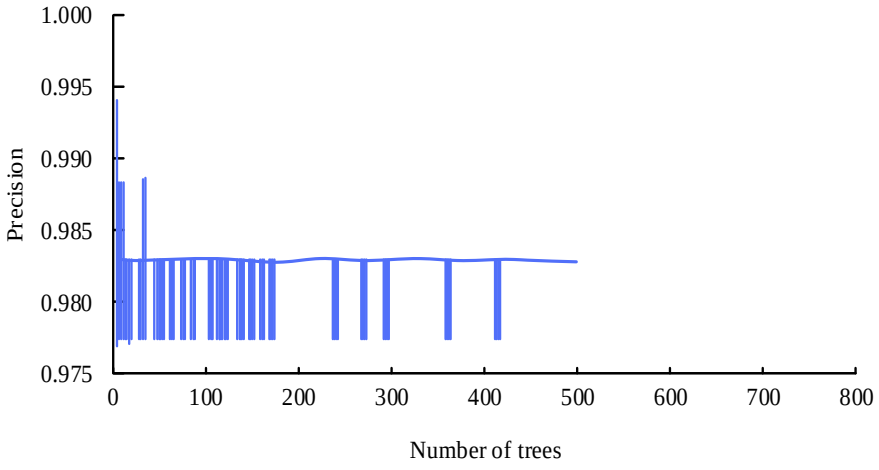


Fig. 8. Influence results of random forest subtrees.

In Fig. 8, after processing with the original RF, Precision gradually converges with the increase of trees. When the subtree is greater than 200, they entered a stable state. Therefore, the number of subtrees in each forest is considered to be 200. Due to the randomness in constructing RF, each algorithm is tested 20 times and then stable values are taken. The experiment uses Recall, Specificity, and G-mean evaluation algorithms to process the results of unbalanced sample sets. Fig. 9 shows experimental result of balanced sample set.

As shown in Fig. 9, in data_ Banknote_authentication dataset, the accuracy of improved RF is 0.985 higher. The precision, recall and F1 score of this algorithm are higher than those of the other two algorithms, with values of 0.977, 0.988, and 0.985. On Spambase, the improved RF algorithm had higher values than the other two algorithms, which are 0.947, 0.977, 0.988, and 0.985, respectively. On Bnana, the improved RF algorithm outperforms the other two algorithms, which are 0.866, 0.898, 0.802, and 0.832, respectively. On vehicle, the improved RF algorithm outperforms the other two algorithms, which is 0.771, 0.763, 0.784, and 0.769, respectively. In conclusion, when the sample set is balanced, the performance of improved RF has certain advantages. Fig. 10 shows the experimental results of an imbalanced sample set.

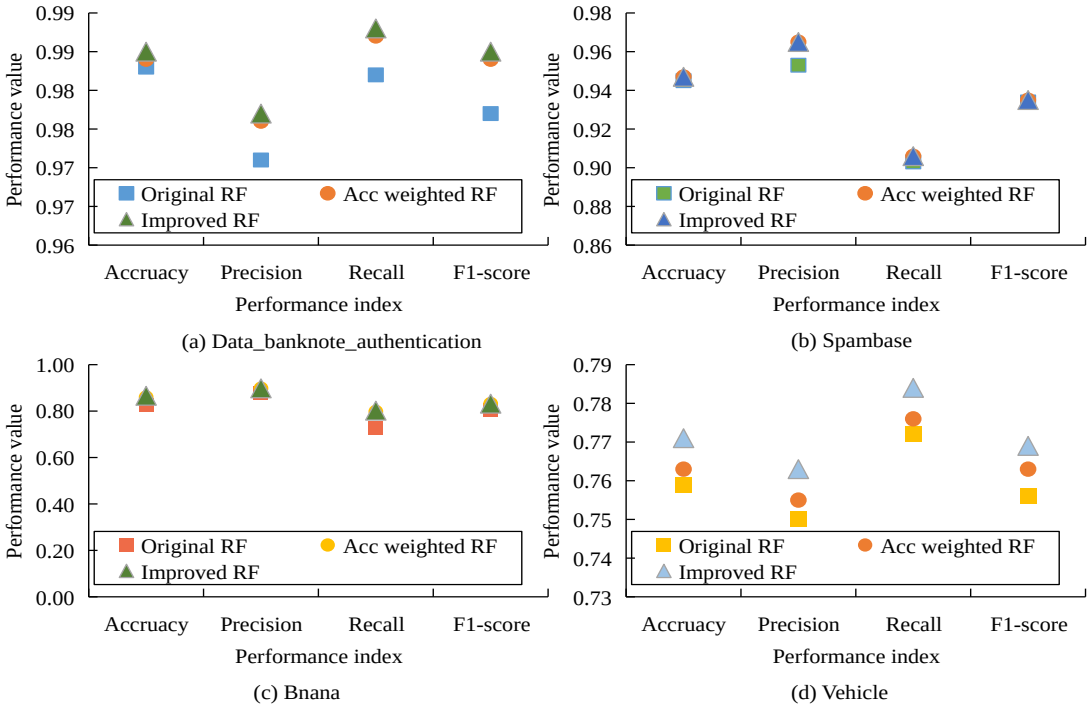


Fig. 9. Experimental results of balanced sample set and unbalanced sample set.

From Fig. 10, the recall, specificity, and G-mean of improved RF are higher than those of the other two algorithms in the Yeast 3 dataset, with values of 0.473, 0.945, and 0.668 respectively. In Ecoli 3, the recall, specificity and G-mean of improved RF are higher than those of the other two algorithms, with values of 0.485, 0.969, and 0.685. In Abalone, the recall, specificity, and G-mean of improved RF are higher than those of the other two algorithms, and their values are 0.994, 0.666, and 0.814, respectively. In Poker_8_9 data set, the recall, specificity, and G-mean of improved RF are higher than those of the other two algorithms, with values of 0.952, 1.0, and 0.975, respectively. To sum up, when the sample set is unbalanced, its specificity representing the accuracy of minority classification is increased by 33.4%, and the comprehensive performance index G-mean is increased by 24.01%. The experimental process is to import the data from the training set into the improved RF algorithm classifier, and train and construct the core function algorithm model. Fig. 11 shows performance test results of sports personalization generation scheme obtained from 20 training sessions.

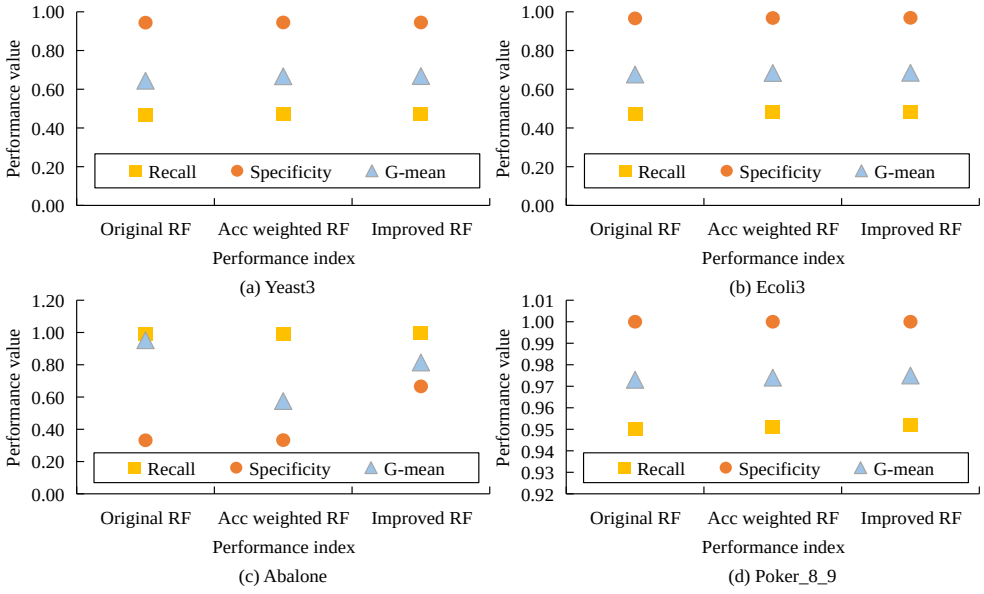


Fig. 10. Experimental results of non-balanced sample set.

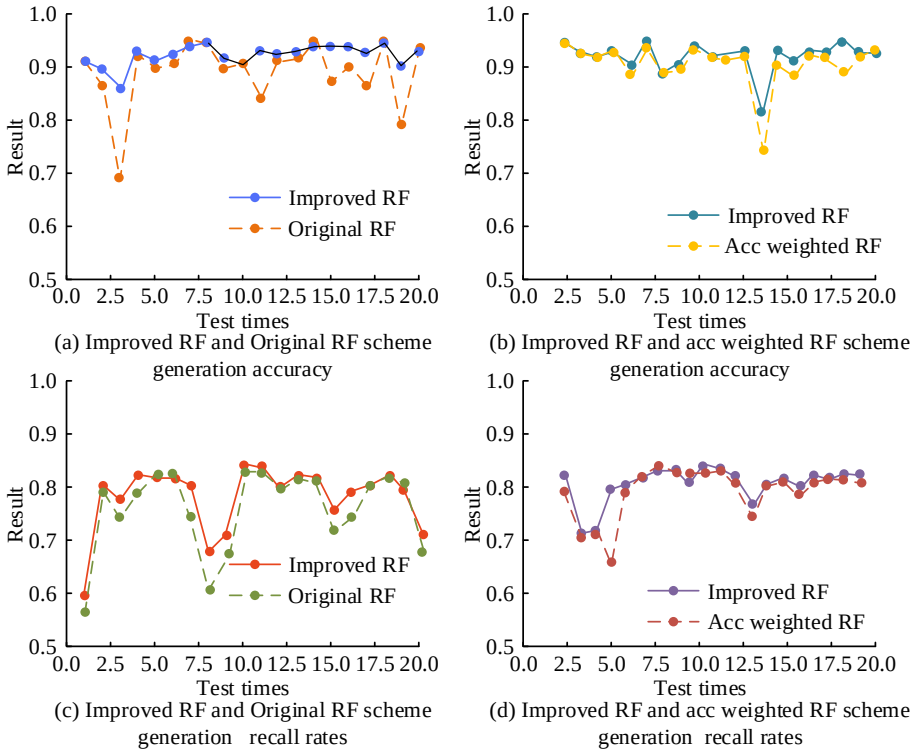


Fig. 11. Performance test results of the motion personalization generation scheme.

In Fig. 11 (a), the accuracy curve generated by the research algorithm is 13 times higher. The highest generation accuracy of the improved algorithm increases by 18.27%. In Fig. 11 (b), the accuracy of the research algorithm is higher than that of the weighted RF for 12 times, with the maximum increase of 4%. The research algorithm is 9.04% higher than the original RF on average, and 2.71% higher than the weighted RF algorithm on average. In Fig. 11 (c), the recall rate of improved algorithm is 13.43% higher than that of the original RF. In Fig. 11 (d), in terms of recall rate, the research algorithm is 7.32% higher than that of the original RF, and 2.68% higher than that of the weighted RF. To sum up, in terms of the accuracy of personalized motion scheme generation of motion software, the improved algorithm reaches 95.05% at most, and its recall rate reaches 83.46% at most.

The ACTIVITYNET Dataset is a large-scale video dataset used to study motor behavior recognition and personalized training plan generation. Video data contains a variety of sports activities, which can be used to analyze and generate personalized sports programs. To more comprehensively verify the performance of the improved RF algorithm proposed in this study in a personalized motion scheme generation system, the research method is compared with several other similar methods on the ACTIVITYNET dataset. The specific results are presented in Table II.

TABLE II. PERFORMANCE OF EACH METHOD'S PERSONALIZED MOTION SCHEME GENERATION ON THE ACTIVITYNET DATASET

Indicators/methods	Traditional RF	Weighted RF	Literature [6]	Deep learning-based method	Literature [11]	Literature [13]	Research method
Accuracy (%)	0.854	0.879	80.5	0.892	85.2	88.1	90.5
Recall	0.841	0.863		0.876			
F1 score	0.847	0.870	0.78	0.884	0.83	0.86	0.91
Computational efficiency (frames per second)	28.500	27.800	30	30.000	25	20	35
Resource consumption (CPU%)	45	42	40	38	35	50	30
Resource consumption (Memory MB)	1300	1250	1200	1400	1000	1500	800

Table II compares the performance of the different methods on the ACTIVITYNET dataset. The research method performs best with an accuracy of 90.5%, an F1 score of 0.91, and a computational efficiency of 35 frames per second. In terms of resource consumption, CPU usage is 30% and memory consumption is 800MB, both of which are the lowest. Compared with traditional RF, weighted RF, and the methods in literature [6], [11], and [13], the research method shows significant advantages in all performance indicators, especially in terms of accuracy, F1 score, and resource consumption. This

indicates that the research method has excellent performance and high resource utilization efficiency in the task of personalized motion plan generation.

V. Discussion

This study designed a personalized training plan generation system based on cloud computing and adopted the improved random forest algorithm to achieve in-depth analysis of users' training and physiological data. The results showed that the accuracy and recall rate of the improved algorithm reached 95.05% and 83.46%, respectively. Compared with the exercise energy consumption monitoring methods based on Internet of Things and cloud computing in the literature [6], this study not only focused on energy consumption monitoring, but also focused on providing personalized exercise recommendations through user data. The SVM-based model proposed in literature [11] and [13] performed well in feature recognition, but this study further improved the performance of random forest algorithm on unbalanced sample sets by introducing weight factors and optimizing the voting mechanism, which may be due to the improved algorithm's better handling of sample imbalance, thus improving the accuracy of classification. In addition, literature [8] and [9] discussed the application of deep learning in health monitoring and virtual reality technology. Although the deep learning method was not employed directly in this study, reinforcement learning strategies were utilized to optimize the generation of motion plans and dynamically adjust the plans to adapt to user feedback. This approach simulated the adaptive characteristics of the deep learning model to a certain extent. The enhanced algorithm's improved performance may be attributed to its more efficient capture and utilization of data characteristics, as well as its ability to respond promptly to user feedback, which were pivotal in achieving personalized motion scheme generation.

VI. Conclusion

In view of the high demand of current users for personalized sports advice services in sports health applications, a sports health cloud platform system using cloud computing and improved RF algorithm was designed by combining cloud computing and machine learning algorithms. Experiments showed that in data_banknote_authentication dataset, the accuracy of the improved RF was 0.985 higher. The precision, recall and F1 score of this algorithm were higher than those of the other two algorithms. Similarly, in the Spambase dataset, the Vehicle dataset, and the Bnana dataset, the scores of the research algorithms were higher than the other two algorithms. In conclusion, when the sample set was balanced, improved RF performance had certain advantages. In the Yeast 3 dataset, Ecoli 3 dataset, Abalone dataset, Poker_ 8_ 9 dataset, the recall, specificity and G-mean of improved RF were higher than those of the other two algorithms. When the sample set was unbalanced, the specificity representing the accuracy of minority classification increased by 33.4%, and the comprehensive classification performance index G-mean increased by 24.01%. The research algorithm was 9.04% higher than the original RF on average, and 2.71% higher than the accuracy weighted RF algorithm on average. In terms of the accuracy of personalized motion scheme generation of motion software, the improved algorithm reached 95.05% at most, and its recall rate reached 83.46% at most. This study reduces the dependence on cloud computing resources by improving the RF algorithm, and improves the performance in resource-constrained environments. The system design focuses on modularity and scalability, enabling the rapid integration of new features through plug-ins or services, while optimizing the interface and interaction design to ensure ease of use and accessibility. The initial investment of cloud-based platform is large, but it can reduce maintenance costs and improve resource utilization in the long run. Small users can reduce upfront investment and hardware expenditure through the pay-as-you-go model, lowering the barrier to entry. However, data security and privacy protection need to be taken seriously and may bring additional costs. Despite the flexibility and economics of cloud platforms, as data volumes and users increase, subscription costs may rise as data volume and users increase, requiring small organizations to address the challenges of complex architecture and security management. Therefore, the cloud platform is feasible for small users in the short term, but the long-term cost and technical requirements need to be evaluated. The research has achieved good results. As the number of users and the amount of data increase, how to ensure the stability and accuracy of the algorithm is a challenge. Future research must solve the adaptability of the algorithm to large data sets and improve the generalization ability of the algorithm. In addition, the integrated application of advanced algorithms such as deep learning will further enhance the intelligent level of personalized motion pattern generation. At the same time, research should focus on user privacy and data security to ensure the confidentiality of user information. The exploration of cross-platform data sharing mechanisms to realize data interoperability between different devices and applications will also be an important direction of future research.

Acknowledgment

The research is supported by: Fund Project: 2022 Inner Mongolia Autonomous Region Higher Education Science and Technology Research Project “Research on Innovation and Entrepreneurship Education for Vocational College Students Based on Professional Education” (Project Number: NJSY22422).

References

[1] Vera-Baquero A, Phelan O. Open Source Software as the Main Driver for Evolving Software Systems Toward a Distributed and Performant E-Commerce Platform: A Zalando Fashion Store Case Study. IT Professional, 2021, 23(1):34-41.

[2] Dergaa I, Saad H B, El Omri A, Glenn J, Clark C, Washif, J, et al. Using artificial intelligence for exercise prescription in personalised health promotion: A critical evaluation of OpenAI’ s GPT-4 model. Biology of Sport, 2024, 41(2): 221-241.

[3] Zheng W, Du Q, Fan Y, Tan L, Xia C, Yang F A personalized programming exercise recommendation algorithm based on knowledge structure tree. Journal of Intelligent & Fuzzy Systems, 2022, 42(3): 2169-2180.

[4] Netz Y, Yekutieli Z, Arnon M, Argov E, Tchelet K, Benmoha E, Jacobs J M. Personalized exercise programs based upon remote assessment of motor fitness: a pilot study among healthy people aged 65 years and older. Gerontology, 2022, 68(4): 465-479.

[5] Prabhu N, Jain H, Tripathi A. MTL-FoUn: A Multi-Task Learning Approach to Form Understanding. 2021, 12917:377-388.

[6] Yang C, Ming H. Detection of sports energy consumption based on IoTs and cloud computing.Sustainable energy technologies and assessments, 2021, 46(Aug.):1-11.

[7] Castelli F A, Rosati G, Moguet C, et al. Metabolomics for personalized medicine: the input of analytical chemistry from biomarker discovery to point-of-care tests. Analytical and bioanalytical chemistry, 2022, 414(2): 759-789.

[8] Xie D, Zhang M, Kumar P M, Muthu B A.Wearable energy-efficient fitness tracking system for sports person health monitoring application.Journal of Intelligent and Fuzzy Systems, 2021(3):1-13.

[9] Ling Z. Personalized healthcare museum exhibition system design based on VR and deep learning driven multimedia and multimodal sensing. Personal and ubiquitous computing, 2023, 27(3):973-988.

[10] Saputro A A, Prasetyo G B. PHYSICAL EDUCATION TEACHER KNOWLEDGE LEVEL ABOUT RULES OF BASKETBALL GAME IN JOMBANG DISTRICT SENIOR HIGH SCHOOL. PHEDHERAL, 2021, 18(1): 1-9.

[11] Men Y. Intelligent sports prediction analysis system based on improved Gaussian fuzzy algorithm. Alexandria Engineering Journal, 2022, 61(7):5351-5359.

[12] Wang L, Sun J, Li T. Intelligent sports feature recognition system based on texture feature extraction and SVM parameter selection.Journal of intelligent & fuzzy systems: Applications in Engineering and Technology, 2020, 39(4 Pt.1):4847-4858.

[13] Sun C M D. SVM-based global vision system of sports competition and action recognition.Journal of intelligent & fuzzy systems: Applications in Engineering and Technology, 2021, 40(2):2265-2276.

[14] Wu B. Sports Intelligent Assistance System Based on Deep Learning. Hindawi Limited, 2021, 2021(Pt.11):1-9.

[15] Huang Z, Stakhiyevich P. A Time-Aware Hybrid Approach for Intelligent Recommendation Systems for Individual and Group Users.Complexity, 2021, 2021(2):1-19.

[16] Debnath S. Fuzzy quadripartitioned neutrosophic soft matrix theory and its decision-making approach. Journal of Computational and Cognitive Engineering, 2022, 1(2): 88-93.

[17] Meng X, Ren G, Huang W. A Quantitative Enhancement Mechanism of University Students' Employability and Entrepreneurship Based on Deep Learning in the Context of the Digital Era. Hindawi Limited, 2021, 10:3-12.

[18] Karthikeyan S, Kathirvalavakumar T. Statistical Inference Through Variable Adaptive Threshold Algorithm in Over-Sampling the Imbalanced Data Distribution Problem. 2022, 1404:267-278.

[19] Zhang S, Liu L, Chen Z, & Zhong H. Probabilistic matrix factorization with personalized differential privacy. Knowledge-Based Systems, 2019, 183:4-13.

[20] Ghale-Joogh H S, Hosseini-Nasab S. On mean derivative estimation of longitudinal and functional data: from sparse to dense. Statistical Papers, 2020, 62(4):2047-2066.

[21] Nguyen A D, Nguyen D T, Dao H N, & Tran N Q. Impact Analysis ofDifferent Effective Loss Functions byUsing Deep Convolutional Neural Network forFace Recognition. 2022, 13636:101-111.

Enhancing Emotion Prediction in Multimedia Content Through Multi-Task Learning

Wan Fan
Department of Culture and Communication, West Anhui University, Lu'an 237000, China

***Abstract*—This study presents a robust multimodal emotion analysis model aimed at improving emotion prediction in film and television communication. Addressing challenges in modal fusion and data association, the model integrates a Transformer-based framework with multi-task learning to capture emotional associations and temporal features across various modalities. It overcomes the limitations of single-modal labels by incorporating multi-task learning, and is tested on the Cmumosi dataset using both classification and regression tasks. The model achieves strong performance, with an average absolute error of 0.70, a Pearson correlation coefficient of 0.82, and an accuracy of 47.1% in a seven-class task. In a two-class task, it achieves an accuracy and F1 score of 88.4%. Predictions for specific video segments are highly consistent with actual labels, with predicted scores of 2.15 and 1.4. This research offers a new approach to multimodal emotion analysis, providing valuable insights for film and television content creation and setting the foundation for further advancements in this area.**

***Keywords*—Multi task learning; multimodal emotion analysis; timing; transformer; attention**

I. Introduction

With the rapid development of big data and artificial intelligence technologies, the application value of multimodal sentiment analysis in the field of film and television communication has become increasingly important. However, current studies are still lacking in multimodal information fusion, temporal feature modeling, and accurate capture of emotional expressions [1-2]. For example, many studies only rely on a single modality (e.g., textual or visual) for emotion recognition, failing to take full advantage of the synergy between different modalities, resulting in limited generalization ability of the models in complex scenes. Furthermore, extant methodologies demonstrate deficiencies in their capacity to effectively address the dynamic fluctuations in multimodal data, impeding the accurate capture of emotional nuances in film and television content. The paucity of labeled data further curtails the efficacy of model training [3-5]. Consequently, the efficient fusion of multimodal information, the enhancement of time-series modeling capability, and the mitigation of label scarcity problem are pivotal issues that must be addressed in the present context. To address the above challenges, this study proposes a temporal multimodal sentiment analysis model that combines Transformer structure and multi-task learning. The model uses Transformer to deep mine temporal features, and supplements the lack of unimodal labels with a self-supervised learning strategy to improve the accuracy of sentiment recognition. Meanwhile, the study adopts a multi-level feature fusion approach to enhance the complementarity of information between different modalities to improve the model's ability to capture dynamic changes in sentiment. In addition, a multi-task learning mechanism is introduced to enhance the adaptability of the model in different scenarios. The central objective of the research is to develop a model that can effectively integrate multimodal information and accurately analyze emotional changes. This model aims to address the issue of emotion recognition in film and television communication. The research has important theoretical and practical values. Theoretically, the research combines the transformer structure and multi-task learning to improve the existing multimodal sentiment analysis methods and provide a new technical framework for temporal feature modeling and multimodal fusion. Practically, the research results can be widely applied in the fields of film and television communication, social media sentiment analysis, online education, and mental health assessment. It can provide more accurate sentiment feedback for content creators, optimize user experience, and promote the further development of sentiment computing technology. The overall structure of the research includes seven sections: The initial parts summarizes the relevant research of multi-task learning and sentiment analysis. Section II, a temporal multimodal emotion analysis model based on multi task learning is proposed. Section III is to conduct experimental analysis on the proposed model. Section IV summarizes the experimental results, points out the shortcomings of the research, Discussion is given in Section V. Finally the paper is concluded in Section VI followed by future work in Section VII.

II. Related Works

With the enhancement of computing power and the surge in data volume, multi task learning gradually demonstrates its advantages in improving model efficiency and performance. Especially in the field of sentiment analysis, multi task learning was widely applied to understand and process various data modalities such as text, audio, and video [6-7]. Many scientists and research institutions have focused on developing advanced algorithms to more accurately identify and analyze human emotions. The following will introduce some related research by scientists and scholars. Yang et al. provided a new solution of a two-stage, multi-task multimodal emotional analyzing mechanism. It adopted a two-stage training strategy for full utilization. The experiment outcomes indicated that the proposed one outperformed the current model in most metrics on both datasets, proving its usefulness [8]. Barnes et al. provided a new solution of a multi task approach for sentimental analyzing solution. This model had hierarchical neural structures, and confirmed that using negation as a multi-task training was a useful way to upgrade the sentimental

analyzing performance, which had been proven on different datasets [9]. Yu et al. provided a new solution of a label generation solution to obtain unimodal supervision and emotional differentiation information, and conducted experiments on different datasets to jointly combine the modalities of multiple tasks. It demonstrated the reliability and stability of multimodal self-generated unimodal supervision [10]. Zhang et al. provided a new solution of a multi task learning multimodal solution. This model proposed cross modal attention and cross task attention. The former was designed to simulate multimodal feature fusion, while the other was designed to capture the interaction. The experiment outcomes indicated that this method had achieved good performance on the dataset assisting sentiment recognition [11].

Mamta et al. provided a new solution of a multimodal method for deep learning. They took the utilization of cross lingual word embeddings to map two languages to a shared semantic space on different languages. The experiment indicated that the proposed one had an accuracy of 0.70 on the movie review dataset, demonstrating good performance [12]. Gao et al. proposed a negative multi-task learning framework to alleviate gender bias. The experiment outcomes indicated that the proposed one could effectively alleviate bias while maintaining the utility of the model in sentiment analysis [13]. Yin et al. provided a new solution of a multi task joint sentiment analysis model. This model used joint training of multiple tasks to obtain local feature characteristics. The results indicated that the model proposed by the research was superior to state-of-the-art models [14]. Saha et al. provided a new solution of a multimodal speech act classification approach based on multi task learning. In addition, a network was combined to capture and effectively integrate shared semantic relationships across modalities. The experiment outcomes indicated that the proposed solution enhanced multimodal speech acts by benefiting from two secondary tasks compared to single modal and single task multimodal speech act classification [15].

In summary, although multi-task learning and sentiment analysis have been widely studied by numerous domestic and foreign experts and scholars, and have been successfully applied. However, shortcomings exists in the challenges of association mining and fusion, especially in the processing of multimodal data. Existing methods often struggle to effectively integrate interactive information between different modalities such as text, audio, and visual, thereby limiting the depth and accuracy of sentiment analysis. In response to this challenge, a temporal multimodal sentiment analysis model combining multi-task learning based on the Transformer framework is proposed. By combining multimodal fusion and temporal data analysis, this model provides a new method for complex sentiment analysis, which is of great positive significance for promoting sentiment analysis and improving the accuracy of sentiment recognition in film and television content.

III. Construction of the Temporal Multimodal Emotion Analysis Model based on Multitask Learning

The study first uses the Transformer framework to extract temporal features closely related to emotions. On this basis, a multi task learning temporal multimodal emotion analysis model is further proposed, aiming to solve the insufficient single modal label issue and improve the recognition and analysis ability of complex emotional states.

A. Time Series Multimodal Emotion Analysis Model Based on Transformer

In film and television communication, multimodal sentiment analysis refers to the use of different types of data, such as text, audio, and video analysis of emotions. Compared with single-modal methods, multimodal sentiment analysis can comprehensively utilize multiple sources of information and improve the accuracy of emotion recognition. To this end, a time-series multimodal sentiment analysis model based on Transformer is proposed, which uses linguistic guided Transformer (LGT) to explore emotional relationships between different modal data. Moreover, through the soft mapping (SM) module, emotion related features are mapped to higher dimensions, effectively integrating multimodal information. This method improves the accuracy of sentiment analysis and optimizes the processing of temporal data. Fig. 1 presents the architecture of this temporal multimodal emotion analysis model.

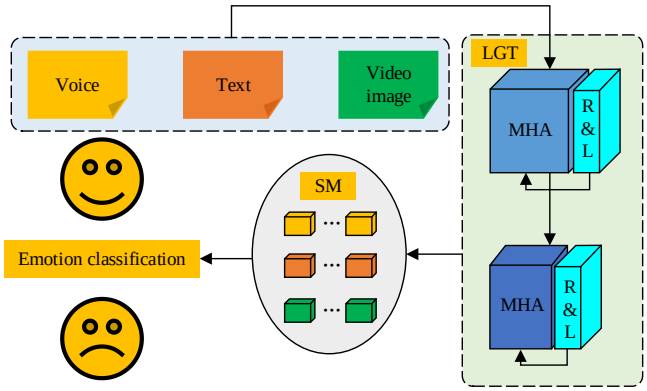


Fig. 1. Multimodal emotion analysis structure diagram.

In Fig. 1, the structure of this multimodal sentiment analysis model includes inputs in three modalities: speech, text, and video image. First, after feature extraction, each modal input is fed into the SM module, which maps the features of different modalities into a unified sentiment representation space. Then, the mapped features are fed into the multi-attention mechanism and the residual connection and normalization module to mine the correlations and affective features between different modalities. The language-guided framework further enhances affective feature extraction through the hierarchical multi-head attention (MHA) module. Finally, the model performs sentiment classification on the fused features and outputs the sentiment classification results. The traditional MHA mechanism is commonly used in machine translation, relying on parallel processing to accelerate the learning process. Its uniqueness lies in its ability to simultaneously process multiple data points without considering their temporal relationships [16-17]. The study applies this mechanism to multimodal emotion analysis to explore the mapping relationships between different modalities. Based on this logic, the LGT model is proposed, which utilizes MHA modules and forward neural networks (FNN) to explore emotional connections between different modalities. Although the modal data of this model

may not be synchronized in time, the model can process these data in parallel, effectively improving the efficiency of temporal data analysis. The architecture of LGT is shown in Fig. 2.

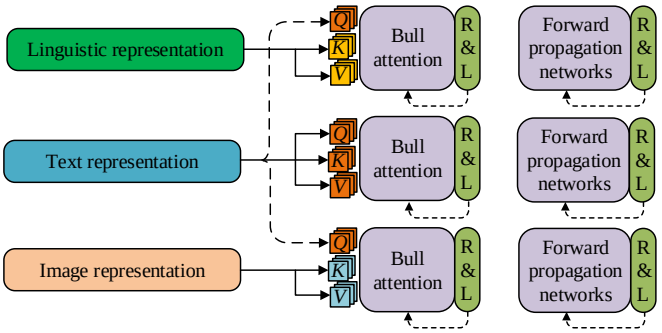


Fig. 2. Architecture diagram of LGT.

Fig. 2 shows a multimodal emotion analysis model dominated by text, supplemented by speech and image modalities. Each modal feature is independently inputted into the MHA mechanism. This setting allows emotional text data to guide speech and image data, thereby exploring the emotional connections between them. The text features are first decomposed into three vectors: query Q_l, key K_l, and value V_l, and these vectors are linearly transformed. Secondly, queries and key vectors are used to calculate attention scores. Finally, the attention score is combined with the value vector and the final result is calculated by weighted sum, as shown in Eq. (1).

$$Attention(Q_i, K_l, V_l) = soft\max((Q_l K_l^T)/\sqrt{d_k})V_l \quad (1)$$

In Eq. (1), d_k represents the dimension. This process is repeated multiple times, each representing an independent $head$. By combining the results of these heads, the final MHA output can be obtained, as shown in Eq. (2).

$$\begin{cases} head_i = Attention(Q_l W^Q, K_l W^K, V_l W^V) \\ F_{(l)} = MHA((Q_l, K_l, V_l)) = Concat(head_1, ..., head_h)W^O \end{cases} \quad (2)$$

After calculating the attention data, these results are immediately input into the FFN, aiming to explore the nonlinear connections between features and improve their expressive power. Eq. (3) provides the information of calculation process.

$$FFN = \mathrm{Re}lu(H'W^1 + b^1) + b^2 \quad (3)$$

The output of each layer of the LGT model is processed through residual connections and hierarchical normalization to ensure effective information transmission and stability, as shown in Eq. (4).

$$LayerNorm(x + Sublayer(x)) \quad (4)$$

In multimodal emotion analysis, when processing speech and image features, the MHA mechanism uses text modality as the source of the query vector. Meanwhile, keywords and truth vectors are derived from speech and image modalities. This method allows text features to introduce different representation spatial information [18], as shown in Eq. (5).

$$\begin{cases} F_{(a)} = MHA(Q_l, K_a, V_a) = Concat(head_1, ..., head_n)W^O \\ F_{(v)} = MHA(Q_l, K_v, V_v) = Concat(head_1, ..., head_n)W^O \end{cases} \quad (5)$$

In the next step of multimodal emotion analysis, the SM module is used to map the learned results of each modality to a new space, fuse them here, and then perform sentiment classification. Fig. 3 displays the schematic diagram of SM structure.

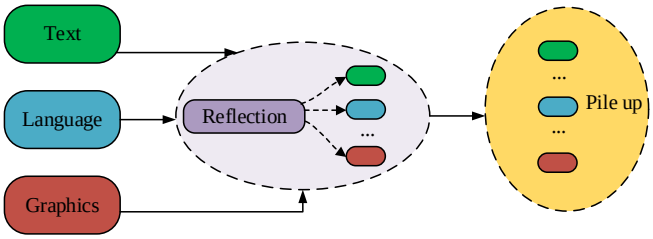


Fig. 3. SM structure diagram.

In Fig. 3, firstly, the input features of each modality, such as text, language, and graphics, enter the reflection module. It performs mapping processing on the input features and converts them into a unified representation format. It then stacks the mapped features by modality, laying the foundation for subsequent multimodal feature fusion and analysis. In the specific steps, the first step is to map the output from the FNN network to a higher dimensional space, and the calculation process is shown in Eq. (6).

$$NewMatrix = W_m M \quad (6)$$

In Eq. (6), W is the transformation matrix with a size of $2k \times k$, used to map the original matrix M to a higher dimension. Next, a vector set $\{v_i^p\}$ of $1 \times k$ size is used to perform soft attention calculation on each matrix in high-dimensional space, and these results are weighted and integrated into vector m_i to obtain the final calculation result of soft attention, as shown in Eq. (7).

$$\begin{aligned} p_i &= soft\max((v_i^p)^T (NewMatrix)) \\ SoftAttention_i(M) &= m_i = \sum_{j=0}^{N} (p_{ij} M_j) \end{aligned} \quad (7)$$

In the final stage of multimodal emotion analysis, the final output of SM is formed by stacking the results of the aforementioned computations. In addition, at the end of each step, a residual operation and hierarchical standardization processing are performed to ensure that the input of each round is integrated into the results of the previous round. Eq. (8) expresses the details.

$$\begin{cases} s = Stacking(\sum_{j=0}^{N} (m_j)) \\ M = LayerNorm(M + s) \end{cases} \quad (8)$$

In Eq. (8), M represents the output matrix, and s represents the independent output matrix obtained for each mode. In multimodal emotion analysis, it is necessary to accumulate the vectors corresponding to each modality in element order, and then perform classification prediction on this accumulated result according to Eq. (9).

$$y \sim p = W_p(LayerNorm(S_l + S_a + S_v)) \quad (9)$$

B. Time Series Multimodal Emotion Analysis Model Based on Multitask Learning

As a machine learning method, multi-task learning allows a model to perform several related tasks simultaneously in one training session, thereby improving the overall performance of the model. Transformer is a neural network structure for processing sequential data, such as text data, that can capture remote dependencies in the data, allowing the model to better understand contextual information. On the basis of a Transformer based temporal multimodal emotion analysis model, a self-supervised label generation module (SLGM) model with fusion multi-level correlation mining framework (MCMF) is proposed. The goal of MCMF is to explore the representation and correlation between multimodal data, improve existing models, and deeply explore emotional connections. Meanwhile, the study applies multi-task learning to multimodal emotion analysis to address the challenges of collaborative learning. By combining three single-modal tasks of text, speech, and image, collaborative learning is achieved through joint training using a multi-task learning framework. The goal of the SLGM model is to address the common problem of missing single modal labels in multimodal datasets. It generates single modal labels according to the mapping of modal representation and labels to meet the training requirements. Fig. 4 shows the structure of a temporal multimodal emotion analysis model based on multi task learning.

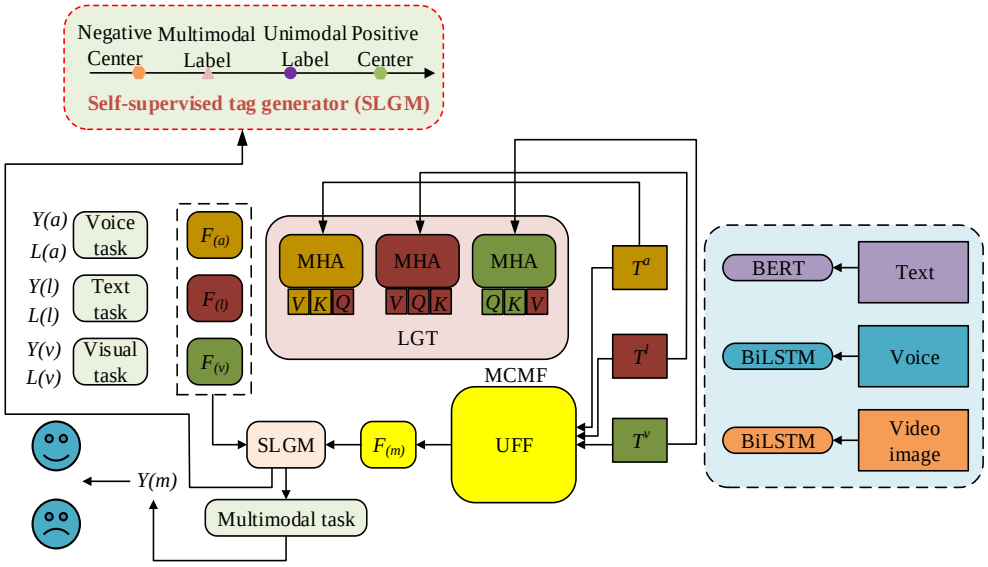


Fig. 4. Temporal multimodal emotion analysis supported by multi task learning.

In Fig. 4, the model mainly includes MCMF module and SLGM module. First, the model receives text, speech and visual data and extracts features through BERT, BiLSTM and other modules to solve the multimodal fusion problem in sentiment analysis. Second, SLGM generates unimodal and multimodal sentiment labels to compensate for the lack of label scarcity. Meanwhile, the unimodal features fusion (UFF) module in MCMF and LGT captures intermodal sentiment associations and fuses multimodal information through the MHA mechanism to improve the accuracy of sentiment prediction. Finally, the model realizes the comprehensive analysis of three-modal sentiment through multi-task learning. The research focuses on exploring the emotional connections between different multimodal representations from top to bottom. To discover the primary feature associations between these representations, a model agnostic fusion strategy is adopted in the study. Inspired by tensor fusion networks, UFF is used to overcome the limitations of traditional fusion methods, mapping single modal features to higher dimensions for fusion [19-20]. UFF fuses each single mode representation through triple Cartesian product, and captures the interaction between bimodal and trimodal through multi-level fusion, as shown in Fig. 5.

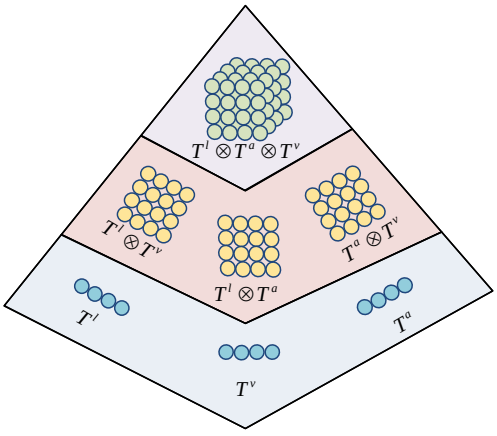


Fig. 5. Single mode feature mapping fusion UFF strategy.

In Fig. 5, the first step is to use T^l、T^a、T^v to capture the intrinsic characteristics of different modes. The second step focuses on exploring the interaction between the two modes, achieved through the calculation of $T^l \otimes T^a$, $T^l \otimes T^v$, and $T^a \otimes T^v$. Finally, the results of these stages are integrated to form a comprehensive fusion tensor, as shown in Eq. (10).

$$\begin{cases} \{(T^l, T^a, T^v) \mid T^l \in [{}_1^{T^l}], T^a \in [{}_1^{T^a}], T^v \in [{}_1^{T^v}]\} \\ F_{(m)} = [{}_1^{T^l}] \otimes [{}_1^{T^a}] \otimes [{}_1^{T^a}] \end{cases} \tag{10}$$

In Eq. (10), $\otimes$ represents the outer product operation, which is used to combine the features of multimodal m, text l, speech a, and visual v. By adding additional dimensions to the single modal tensor T^l、T^a、T^v and performing outer product operations, fused features are formed. Further application of hard parameter sharing mechanism in the framework to achieve sentiment analysis, where low-level networks share weights and high-level networks assign weights to specific tasks, as shown in Fig. 6.

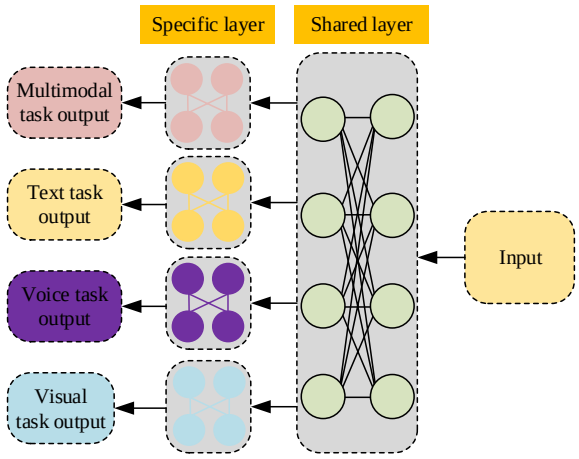


Fig. 6. Emotion analysis structure of multi task learning.

In Fig. 6, the study uses basic learning networks as shared layers, while independent prediction networks for each task are used as specific layers. Under this framework, a multimodal $F_{(m)}$ and three single modal tasks are set, with F^l, F^a, and F^v as inputs, respectively. Eq. (11) provides the definition of the dedicated layer.

$$\begin{cases} F_s^* = \mathrm{Re}\,LU(F_s W_s^{1T} + b_s^1) \\ y_s = F_s^* W_s^{2T} + b_s^2 \end{cases} \tag{11}$$

To adapt to multi-task learning frameworks, the SLGM model is proposed, which aims to generate single modal labels from multimodal data. SLGM is based on two core principles: the mapping relationship between modal representation and modal supervised values, and the proportional consistency of mapping relationships between different modalities. This method allows for the effective generation of labels required for single modal tasks, with specific calculations detailed in Eq. (12).

$$(F_m \# L_m) \infty (F_u \# L_u) \tag{12}$$

In Eq. (12), m represents multimodality and u represents covering $\{linguistic, acoustic, visual\}$ modes. F represents modal representation, L represents modal supervised value, $\#$ represents mapping relationship, and ∞ represents proportional relationship. These definitions indicate that modality and multimodal labels are highly correlated in terms of emotional polarity. However, in practical applications, there may be differences in the emotional polarity between single modal labels and multimodal labels. Among them, multimodal labels may be positive, but individual visual modalities may be marked as negative due to crying expressions. The strategy of SLGM to solve this problem is to divide modal representation as two categories abide by the emotional polarity, followed by calculating the center value of each category separately, and obtain the center representation of emotions from modalities, as shown in Eq. (13).

$$\begin{cases} C_p = \dfrac{\sum_{i=1}^{N} I(y(i) > 0) \cdot F_i}{\sum_{i=1}^{N} I(y(i) > 0)} \\ C_n = \dfrac{\sum_{i=1}^{N} I(y(i) < 0) \cdot F_i}{\sum_{i=1}^{N} I(y(i) < 0)} \end{cases} \tag{13}$$

In Eq. (13), $I(\cdot)$ is the symbol of the indicator function, N is the symbol of the number, and F represents the modal representation. Next, SLGM applies the Bartcharia distance coefficient for measuring the sample and its corresponding category center. The specific calculation is detailed in Eq.(14).

$$\begin{cases} S_p = \sum_{j=1}^{K} \sqrt{F(j) C_p(j)} \\ S_n = \sum_{j=1}^{K} \sqrt{F(j) C_n(j)} \end{cases} \tag{14}$$

In Eq. (14), K is the symbol of the total elements number in the modal representation. SLGM uses proportion and difference to reflect the mapping relationship, so the original Eq. (12) has been updated and expressed as Eq. (15).

$$\begin{cases} S_m / L_m = S_u / L_u, S_m - L_m = S_u - L_u \\ L_u = L_m + \dfrac{S_u - S_m}{2} * \dfrac{L_m + S_m}{S_m} \end{cases} \tag{15}$$

In Eq. (15), S_m and S_u represent the deviation degrees of multimodal and single-mode, respectively, while L_m and L_u refer to the markings of multimodal and single-mode. To address the issue of result fluctuations caused by iterations in the label generation process, SLGM has implemented a mechanism for dynamically updating single modal labels. The specific calculation method is shown in Eq. (16).

$$y_u^{(i)} = \frac{1}{2} * \frac{i-1}{i+1} * y_u^{(i-2)} + \frac{1}{2} * \frac{i-1}{i+1} * y_u^{(i-1)} + \frac{2}{i+1} * y_u^i, i \geq 3 \quad (16)$$

In summary, starting from 3rd iteration, each iteration i's outcome is associated with the first two rounds of $(i-1)$ and $(i-2)$, ensuring that as the number of iterations increases, the generated labels gradually stabilize.

IV. Test of Multimodal Emotion Analysis Model in Film and Television Communication

The research experiment first sets two datasets and their experimental parameters to conduct the experiment. Secondly, the results on these two datasets are compared and ablation experiments are performed to analyze the contributions of each component. Subsequently, performance comparisons are made with other models. Finally, the study conducts an analysis of the effectiveness of LGT and selects some film and television clips for in-depth analysis, demonstrating the emotional analysis ability of the proposed model.

A. Test of a Transformer Based Time-Series Multimodal Emotion Analysis Model

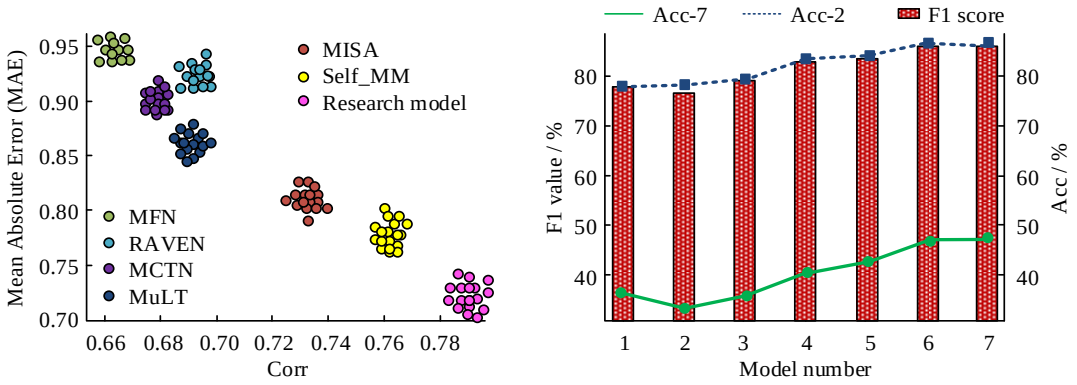
To validate the temporal multimodal emotion analysis model, two datasets, Cmumosi and Cmumosei, are selected for the study. The CMUMOSI and CMUMOSEI datasets contain 2,200 video clips and 24,000 YouTube comment clips, respectively. These cover a variety of data formats, including text, audio, and video modes. Bidirectional encoder representations from Transformers are used for feature extraction of text data, while Mel-Frequency Cepstral Coefficients are used for feature transformation of audio data. Video is sampled at 2 frames per second, scaled to 224x224 pixels, and visual features are extracted using a convolutional neural network. To ensure modal synchronization, the window alignment method is used to process multimodal data. In addition, to solve the problem of unbalanced emotion category, the study further applied a small amount of data enhancement strategy to improve the model's low-frequency emotion recognition ability.

Since the final model consists of MCMF and SLGM, MCMF includes UFF and LGT, and LGT consists of SM, MHA and FNN, ablation experiments are set up to test the benchmark performance of each module, and the results are shown in Table I. The LGT model based on Model 1, which consists of SM, MHA, and FNN, can extract basic emotional features. However, its performance is mediocre due to the lack of modal fusion and label generation. After UFF module is added to LGT in Model 2, the correlation between modes is enhanced and various indicators are improved. SLGM module is added to LGT in Model 3 to make the model capable of generating single mode labels, which effectively improves the accuracy of sentiment analysis. Model 4 is an MCMF module composed of UFF and LGT. Through the high-dimensional fusion of UFF and the emotional feature analysis of LGT, the performance of the model is close to the final model, and the synergistic effect of the two is verified. Model 5 is the final model. Combined with SLGM, the accuracy of multimodal emotion recognition is further improved. Moreover, the accuracy rate, recall rate and F1 score are all the best, which are 0.92, 0.93, and 0.92, respectively.

TABLE I. Ablation Test Results

Network structure	Accuracy value	Recall value	F1 score
SM+MHA+FNN (Model 1: LGT base model)	0.83	0.82	0.81
UFF+SM+MHA+FNN (Model 2: Add UFF module)	0.86	0.85	0.85
LGT+SLGM (Model 3: LGT combined with SLGM module)	0.88	0.87	0.87
UFF+LGT (Model 4: MCMF module)	0.89	0.89	0.88
UFF+LGT+SLGM (Model 5: MCMF+SLGM complete model)	0.92	0.93	0.92

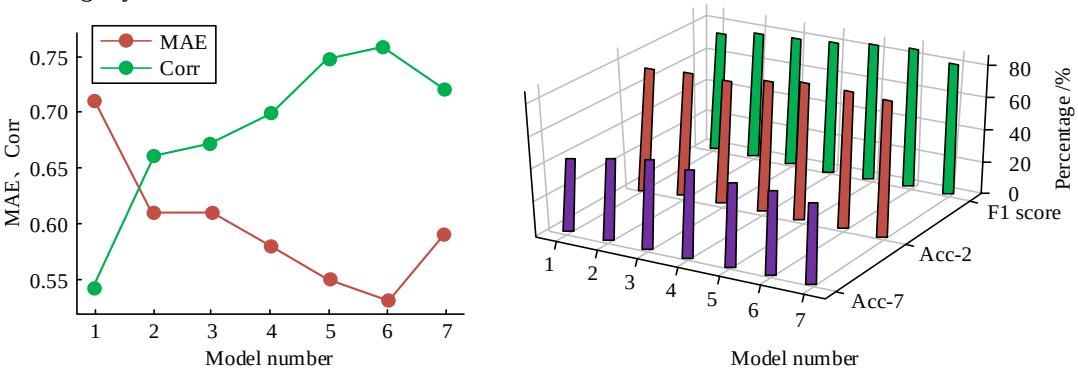
The testing is divided into regression and classification tasks, which hires Pearson correlation (Corr) together with mean absolute error (MAE) as indicators. The classification task is evaluated using seven class accuracy (Acc-7), two class accuracy (Acc-2), and F1 score (F1). The proposed one is planned to be compared with multiple existing models on the Cmumosi dataset, as shown in Fig. 7. The compared models include memory fusion network (MFN), recurrent attention variation embedding network (RAVEN), multimodal cyclic translation network (MCTN), multicurrent Transformer (MulT) Modality Invariant and Specific Representations for multimodal emotion analysis (MISA) and self-supervised multi task learning for multimodal emotion analysis (Self MM). These models are denoted as Model 1 to Model 6, respectively. In Fig. 7(a) and Fig. 7(b), the multi-task learning time-series multimodal sentiment analysis model proposed in the research outperforms the existing mainstream models in several key metrics. In the regression task, the MAE of the proposed model is 0.70, which is significantly lower than the 0.87 of MulT and the 0.78 of MISA. It indicates that the research model has a smaller error in sentiment prediction accuracy. Meanwhile, the Corr of the research model reaches 0.82, which is higher than the 0.70 of MulT and 0.76 of MISA. It indicates that the research model is more robust in modeling emotional features and can more accurately reflect the emotional associations between different modalities. In the categorization task, the Acc-7 of the research model is 47.1%, which is significantly better than the 42.3% of MISA and the 46.8% of Self-MM, indicating that the proposed model is able to discriminate different categories of emotions more accurately. In addition, the research model achieves 88.4% in both Acc-2 and F1 values, surpassing the 86.1% of Self-MM and the 83.1% of MulT. This result further validates the advantages of the multi-task learning framework in emotion recognition, which can effectively improve the ability to capture complex emotional features.



(a) MAE and Corr for Cmumosi (b) Acc-7, Acc-2, and F1 score for Cmumosi

Fig. 7. Comparison on the cmumosi dataset.

The next comparison on the Cmumosei dataset is shown in Fig. 8. In terms of the regression task, the MAE of the research model drops to 0.59, which is lower than both 0.61 for MCTN and 0.71 for MFN. It implies that its prediction of sentiment trends is more accurate. Meanwhile, the Corr of the research model reaches 0.72, which is significantly better than the 0.54 of MFN) and 0.67 of MCTN. It implies that the research model is able to extract multimodal features more stably and establish reasonable sentiment mapping relationships among different modalities. In the classification task, the research model's Acc-7 is 49.6%, which is slightly lower than MulT's 51.8% and Self-MM's 52.2%, but reaches 82.2% on Acc-2, which is higher than MCTN's 79.3% and MFN's 77.9%. Meanwhile, its F1 score is 82.2%, which showed a strong sentiment categorization ability. In summary, although the research model approaches the optimal model in some indicators, its overall performance exceeds that of most mainstream methods, particularly in the regression task. The research model demonstrates a more stable and robust sentiment prediction ability, ensuring accurate sentiment analysis of film and television communication content.



(a) MAE and Corr for Cmumosei (b) Acc-7, Acc-2, and F1 score for Cmumosei

Fig. 8. Different models are compared on the cmumosei dataset.

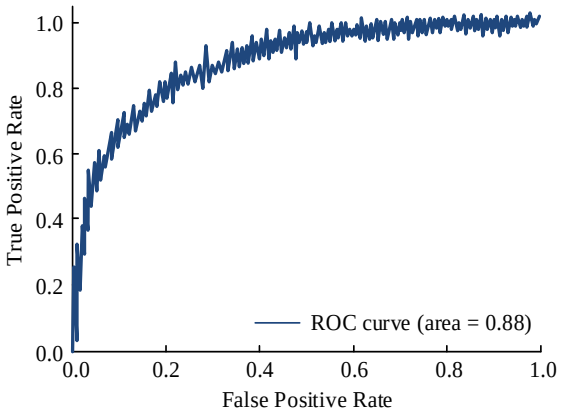


Fig. 9. Receiver operating characteristic curve.

Next, how the research model performing is validated by randomly selecting over 4000 samples from the test set for binary classification results testing. Through assessing the predicted labels with the real sample labels together, the true and false positives are calculated. Based on this data, the operating characteristic curves of the subjects are plotted, and the outcomes are given in Fig. 9. The model proposed in the study has a strong ability to distinguish between two types of emotional polarity, reflecting its advantage in accurately identifying and predicting emotional tendencies in film and television content.

The research model combines two mechanisms, LGT and SM, respectively, for modal interaction and mapping results to high-dimensional space for effective fusion. The study conducted ablation analysis on the Cmumosi dataset, as shown in Fig. 10. Both of these structures have a positive impact on

sentiment classification. The experiment is divided into 1-4 scenarios: no LGT and SM, only SM, only LGT, and both LGT+SM are used. In Fig. 10(a) and Fig. 10(b), the model performance decreases significantly with the removal of LGT and SM, indicating the importance of these two modules in sentiment modeling. When using only SM, although it improves some tasks, the overall performance is still limited by the lack of intermodal interaction capability. Using only LGT, although it improves modal interaction, it lacks the support of the soft-mapping mechanism, resulting in a decrease in the ability to fuse high-dimensional features. The complete model (LGT+SM) performs optimally and achieves the best results in all metrics, confirming the key role of LGT in enhancing intermodal affective interactions, while the SM module enhances the feature expression capability through high-dimensional mapping. The results of the ablation experiments illustrate that the combination of the two can maximize the emotion recognition ability of the model, making it more applicable to complex film and television communication scenarios.

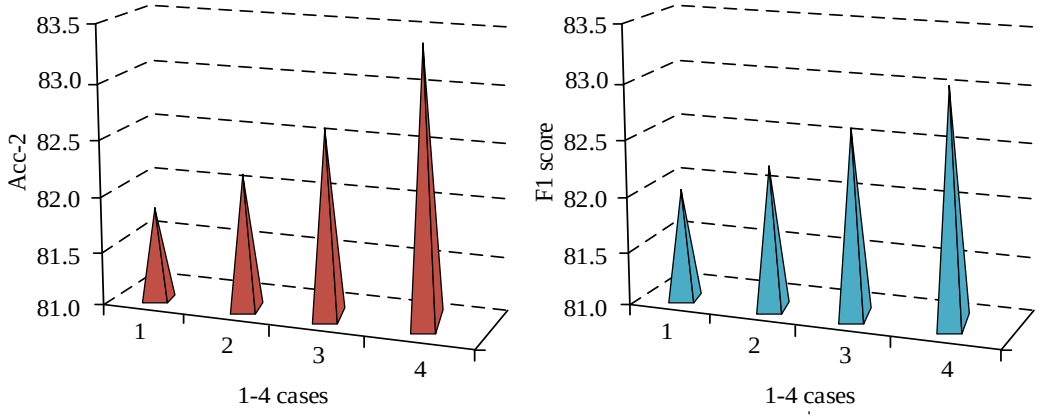


Fig. 10. Cmumosi dataset ablation analysis.

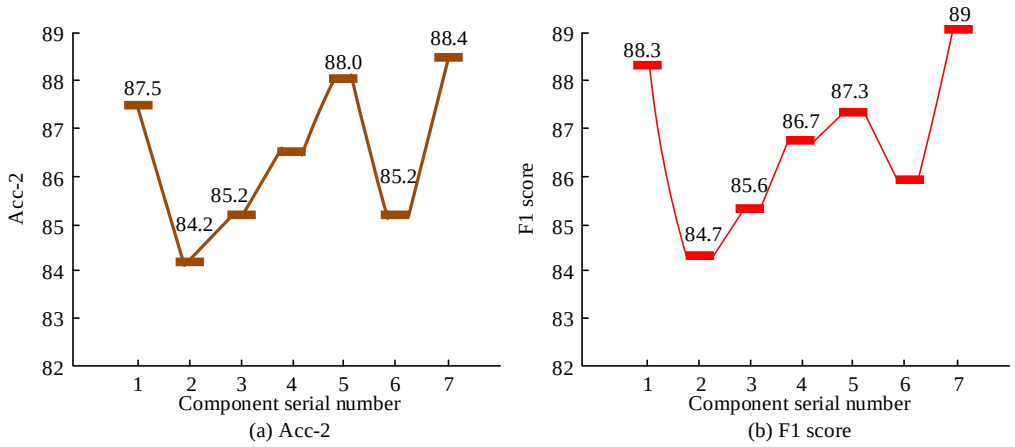
B. *Test of a Time-Series Multimodal Emotion Analysis Supported by Multitask Learning*

To compare the performance of different models, classification and regression evaluation indicators are used in the study, as shown in Table II. The research model outperforms the comparison model in all indicators: Acc-2 reaches 88.5%, which is 2.4% higher than Self_MM and 1.9% higher than Zhang Y et al.'s model. It indicates that the sentiment polarity determination is more accurate. Acc-7 reaches 47.2%, which outperforms MISA, Self_MM, and Yu W et al.'s model. It indicates that it has a stronger generalization ability in the task of complex sentiment classification. MAE is the lowest at 0.70, with better error control than Barnes J et al. (0.74) and Yang B et al. (0.76), which is more stable in prediction. Corr is the highest at 0.82, which is better able to accurately capture the trend of emotion change compared to Zhang Y et al.'s model at 0.79 and Self_MM's 0.80. Overall, the research model performs best on all indicators, validating the effectiveness of multi-task learning and temporal multimodal fusion methods to provide more accurate prediction results in complex film and television sentiment analysis tasks.

This study conducted ablation tests on the Cmumosi dataset to accomplish the effects assessment of various components of the model. The experiment set up seven components: UFF, LGT, SLGM, UFF+LGT, UFF+SLGM, LGT+SLGM, and UFF+LGT+SLGM, numbered 1 to 7, respectively. Fig. 11 gives the information that feature concatenation is used when UFF is missing, single modal representation is directly input when LGT is missing, and multimodal labels are used for training when SLGM is missing. The results show that UFF is the key to performance improvement, especially with the combination of UFF and SLGM achieving the highest Acc-2 and F1 scores of 88.0% and 89.0%, respectively. The combination of UFF+LGT and LGT+SLGM did not perform as expected, with 86.5% and 85.2%, respectively, due to the lack of single modal labels or basic correlations between modalities.

TABLE II. DIFFERENT MODELS ON THE CMUMOSI DATASET

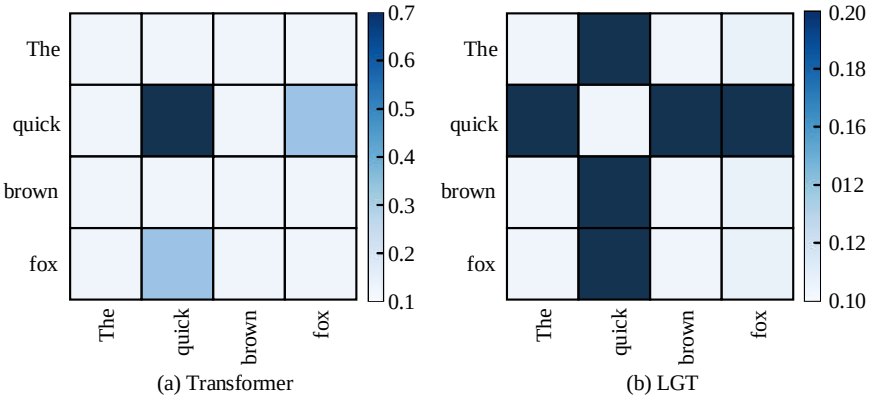
Type	Acc-2	Acc-7	MAE	Corr
MFN	77.9	36.3	0.95	0.66
RAVER	78.1	33.2	0.92	0.69
MCTN	79.3	35.6	0.91	0.68
MulT	83.1	40.2	0.87	0.7
MISA	83.4	42.3	0.78	0.76
Self_MM	86.1	46.8	0.71	0.8
The model of Yang et al.	85.6	45.59	0.74	0.8
The model of Barnes et al.	86.05	45.17	0.74	0.8
The model of Yu et al.	84.96	45.67	0.73	0.79
The model of Zhang et al.	84.42	46.01	0.72	0.79
Research model	88.5	47.2	0.7	0.82



(a) Acc-2 (b) F1 score

Fig. 11. Ablation analysis of different component structures.

For confirming the LGT model effectiveness, a test sentence "The quick brown fox" is used on the Cmumosi dataset, and the performance of LGT and Transformer in attention mechanism is compared. As shown in Fig. 12, the attention scores are represented by color depth, where a darker color is the symbol for higher weights and a lighter color is the symbol for lower weights. Comparing Fig. 12(a) and Fig. 12(b), LGT is more accurate in assigning word weights. For example, in the MHA mechanism of Transformer, "quick" pays significant attention to "fox", while the distribution of LGT is more uniform, allowing for a more balanced focus on the structure of the entire sentence. This demonstrates the advantage of LGT in finely adjusting attention weights, especially in capturing key vocabulary. It further demonstrates its effectiveness in processing language tasks, especially in understanding sentence structures.



(a) Transformer (b) LGT

Fig. 12. LGT attention matrix result graph.

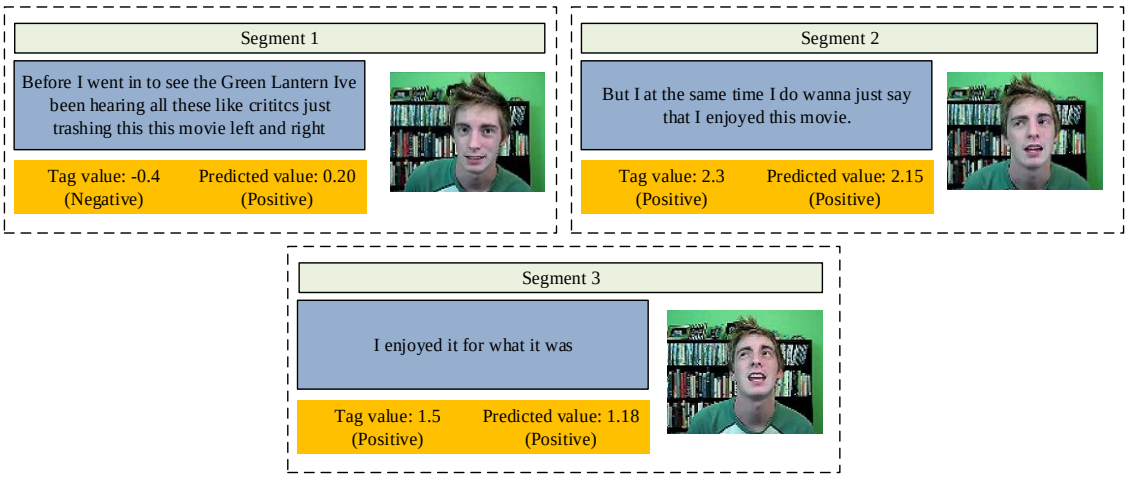


Fig. 13. Cmumosi dataset film and video emotion analysis results.

The study selects specific film and television video clips from the Cmumosi dataset to accomplish the evaluation of the the proposed one's analytical ability. Three video clips are selected for detailed testing in the experiment, with text content and keyframe images shown in Fig. 13, supplemented by experimental labels and prediction results. The observation can tell that the predictions of the second and third segments are highly consistent with the actual labels, demonstrating the accuracy of the model in sentiment analysis, with predicted scores of 2.15 and 1.18, respectively. However, there is a deviation in the results of the first segment, with a predicted value of only 0.20. This difference is attributed to the fact that the image at the beginning of the video can mislead the model's judgment. The combination of the smiling image of the man in the first clip with neutral text leads the model to lean towards positive predictions. In summary, this experiment reveals the complexity of the model in interpreting emotions and demonstrates its advantages in integrating visual and textual cues.

TABLE III. TEST RESULTS OF IEMOCAP AND MELD DATASETS

Type	Acc-2	Acc-7	F1-Score	MAE	Corr
MFN	0.78	0.65	0.72	0.9	0.65
RAVER	0.79	0.66	0.73	0.88	0.67
MCTN	0.8	0.68	0.75	0.85	0.69
MulT	0.83	0.7	0.78	0.82	0.72
MISA	0.85	0.72	0.8	0.78	0.74
Self_MM	0.86	0.73	0.82	0.76	0.76
The model of Yang et al.	0.84	0.71	0.81	0.78	0.76
The model of Barnes et al.	0.83	0.71	0.82	0.77	0.75
The model of Yu et al.	0.83	0.73	0.82	0.77	0.75
The model of Zhang et al.	0.84	0.71	0.83	0.77	0.73
Research model	0.88	0.75	0.85	0.73	0.78

To further verify the applicability of the model in different areas of emotion analysis, the study introduces additional datasets of IEMOCAP and MELD to test the performance of the model in conversational emotion recognition and multimodal emotion dialog, respectively. The IEMOCAP dataset contains multimodal dialog emotion labels, which can be used to test the emotion recognition effect of the model in the dialog scene. The MELD dataset provides multimodal emotion data of multi-round dialogues, which can be used to verify the emotion recognition ability of the model in complex scenes. In Table III, the proposed model outperforms the other models in all indicators on both IEMOCAP and MELD datasets. The models proposed by Yang B et al. Barnes J et al. Yu W et al. and Zhang Y et al. shows strong competitiveness in Acc-2, Acc-7, and F1-Score, but compared to the proposed model, they still fall slightly short of the proposed model, especially in Acc-2 and F1-Score. In particular, in terms of Acc-2 and F1-Score, the proposed model reaches 0.88 and 0.85, while the highest Acc-2 of the other models is only 0.84, and the highest F1-Score is 0.83. It suggests that the proposed model has more advantageous in terms of sentiment categorization accuracy. In addition, the MAE value of the proposed model is 0.73, which is more stable and exhibits a lower error rate compared to the 0.77-0.78 of the other models. It proves its reliability in fine-grained emotion recognition tasks. Overall, the proposed model of the study demonstrated better performance in the sentiment analysis task, further validating its application value in the field of film and television communication.

V. DISCUSSION

To improve the accuracy of multimodal sentiment analysis in film and television communication, this study presented a time-series model based on a Transformer structure with MCMF and SLGM modules. Compared with the two-stage multi-task learning model proposed by Yang B et al. [8], which improved feature classification, this approach addressed limitations when multimodal labels were sparse. By incorporating the SLGM module to generate single-modal labels, this model not only enriched the diversity of emotional labels, but also significantly improved the accuracy of complex emotion recognition. Results on IEMOCAP and MELD datasets showed F1 scores up to 0.85. In terms of efficiency, the model reduced the MAE on the CMUMOSEI dataset by 15%, demonstrating the effectiveness of the multi-task learning framework in sentiment analysis. Compared to Zhang Y et al.'s model [11], which used cross-modal attention and multi-task learning for accuracy, this model improved label stability and captures subtle emotions, achieving 47.1% accuracy in seven-category tasks and 88.4% accuracy in binary tasks. This improvement was due to the SLGM module, which ensured efficient emotion prediction even with sparse single-modal labels. In practical applications on social media platforms, the model could be utilized to analyze the emotional trends of users' video content, thereby providing real-time emotional feedback to content creators and platform operators for the purpose of optimizing content push strategies. In online education, the model could help identify students' emotional responses during video lessons so that teachers could adjust their teaching methods. In healthcare, the model helped assess mental health and monitor emotional states by analyzing video and voice recordings of patients.

For broader applications, the model could be extended to recognize multi-character interactions and nuanced emotions. Such scenes in film and television often involve multiple, complex emotions, where the recognition of subtle emotional states was crucial. Future studies could include these scenarios to verify the adaptability of the model. For real-time applications, the model showed high computational efficiency in binary emotion tasks, indicating potential for real-time deployment. Further optimization with a lightweight Transformer structure and a multi-tasking strategy could support real-time video analysis and meet the needs of efficient media analysis.

In summary, there are three main arguments of the study: First, SLGM improves the acquisition quality of unimodal sentiment labels and still maintains a stable sentiment recognition ability in multimodal datasets without complete annotation. Second, UFF allows different modal features to be fused more efficiently and improves the synergy between multimodal data, thus enhancing the robustness and generalization ability of the model. Third, MCMF enhances the

modeling capability of complex sentiment features by jointly optimizing LGT and SLGM, while improving the sentiment classification accuracy. These results show that the proposed model of the study not only outperforms existing methods on several benchmark datasets, but also makes a significant breakthrough in the accuracy and stability of multimodal data fusion and sentiment recognition. It lays a foundation for the application of sentiment analysis technology in more complex multimodal data applications, which has positive implications for improving the communication effects of movies and television and optimizing user experience.

VI. Conclusion

A temporal multimodal emotion analysis model supported by multi task learning was proposed to address the challenges of sentiment association mining and fusion in multimodal data. This model utilized a Transformer based framework to deeply explore temporal features related to emotions, and compensated for the shortcomings of single modal labels, enhancing the accuracy of emotion prediction. The experimental results showed that, specifically, in the binary sentiment analysis task on the CMUMOSEI dataset, the average processing time of the model was about 0.05 s per frame of data, enabling it to perform sentiment prediction in near real-time conditions. Furthermore, the ablation experiments demonstrated that the structural optimization of the UFF and LGT modules markedly reduced the computational overhead of the model and enhanced its computational speed by approximately 30% in comparison to traditional multimodal sentiment analysis models. It substantiated the superior performance of the proposed model over several existing models. These results demonstrated their clear advantages in accurately recognizing affective tendencies. To further validate the effectiveness of LGT, a comparative analysis was conducted on the differences in attention mechanisms between LGT and traditional Transformers. It was found that LGT was more precise in finely adjusting word weights, which was crucial for capturing the emotional meaning of sentences. In addition, by selecting specific film and television video segments for testing, the predictions of the second and third segments were highly consistent with the actual labels, with predicted scores of 2.15 and 1.4, respectively. It was proved that the research model can highly match the actual labels in most cases.

VII. Future Work

Despite the merits of the current model's superior emotion recognition performance as demonstrated in this study, there are notable deficiencies. First, the Transformer structure's application in temporal modeling incurs substantial computational complexity when dealing with video data of considerable duration. This may impede the efficiency of practical applications. Second, the multimodal fusion process primarily relies on global feature matching, neglecting the potential benefits of local information in sentiment analysis. This may result in a reduction in accuracy for certain nuanced sentiment recognition tasks. In addition, the research model still has room for improvement in the recognition of extreme emotion categories (e.g., intense anger or extreme happiness), and more fine-grained feature extraction methods can be explored in the future to improve the model's emotion classification ability. To address the aforementioned limitations, future research can be optimized in the following ways. First, a more efficient attention mechanism should be introduced to reduce computational complexity and improve the applicability of the model to long time series data. Second, local feature modeling should be combined to improve the model's sensitivity to subtle sentiment changes. Third, an adaptive multi-task learning strategy should be introduced to enhance the model's generalization ability in different application scenarios. Furthermore, at the level of the data, the diversity of the training data can be expanded in the future to encompass a more extensive range of movie and television styles and cultural backgrounds, thereby enhancing the robustness of the model.

In summary, this study improves the performance of multimodal sentiment analysis through an innovative model design, while pointing out the existing limitations and proposing appropriate improvement directions. The research results not only provide an effective solution for sentiment computation in the field of film and television communication, but also lay the foundation for the further development of multimodal sentiment analysis.

Acknowledgment

The research is supported by The Youth Project of Philosophy and Social Sciences Planning in Anhui Province in 2022: Research on the Influence of Social Short Video on the National Identity of Small Town Youth in Anhui Province (Project Number: AHSKQ2022D021).

References

[1] Singh A, Saha S, Hasanuzzaman M, Dey K. Multitask learning for complaint identification and sentiment analysis. Cognitive Computation, 2022, 1(1): 1-16.

[2] Zhang T, Gong X, Chen C L P. BMT-Net: Broad multitask transformer network for sentiment analysis. IEEE transactions on cybernetics, 2021, 52(7): 6232-6243.

[3] Mao R, Li X. Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification. Proceedings of the AAAI conference on artificial intelligence, 2021, 35(15): 13534-13542.

[4] Bie Y, Yang Y. A multitask multiview neural network for end-to-end aspect-based sentiment analysis. Big Data Mining and Analytics, 2021, 4(3): 195-207.

[5] Chen F, Yang Z, Huang Y. A multi-task learning framework for end-to-end aspect sentiment triplet extraction. Neurocomputing, 2022, 479(3): 12-21.

[6] Zhao M, Yang J, Qu L. A multi-task learning model with graph convolutional networks for aspect term extraction and polarity classification. Applied Intelligence, 2023, 53(6): 6585-6603.

[7] Purohit J, Dave R. Leveraging Deep Learning Techniques to Obtain Efficacious Segmentation Results. Archives of Advanced Engineering Science, 2023, 1(1): 11-26.

[8] Yang B, Wu L, Zhu J, Shao B, Lin X, Liu T Y. Multimodal sentiment analysis with two-phase multi-task learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30(1): 2015-2024.

[9] Barnes J, Velldal E, Øvrelid L. Improving sentiment analysis with multi-task learning of negation. Natural Language Engineering, 2021, 27(2): 249-269.

[10] Yu W, Xu H, Yuan Z, Wu J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. Proceedings of the AAAI conference on artificial intelligence, 2021, 35(12): 10790-10797.

[11] Zhang Y, Rong L, Li X, Chen R. Multimodal sentiment and emotion joint analysis with a deep attentive multi-task learning model. European Conference on Information Retrieval. Cham: Springer International Publishing, 2022, 1(1): 518-532.

[12] Mamta, Ekbal A, Bhattacharyya P. Exploring Multi-lingual, Multi-task, and Adversarial Learning for Low-resource Sentiment Analysis. Transactions on Asian and Low-Resource Language Information Processing, 2022, 21(5): 1-19.

[13] Gao L, Zhan H, Sheng V S. Mitigate gender bias using negative multi-task learning. Neural Processing Letters, 2023, 55(8): 11131-11146.

[14] Yin C, Chen Y, Zuo W. Multi-task deep neural networks for joint sarcasm detection and sentiment analysis. Pattern Recognition and Image Analysis, 2021, 31(8): 103-108.

[15] Saha T, Upadhyaya A, Saha S, Bhattacharyya P. A multitask multimodal ensemble model for sentiment-and emotion-aided tweet act classification. IEEE Transactions on Computational Social Systems, 2021, 9(2): 508-517.

[16] Tan Y Y, Chow C O, Kanesan J, Chuah J H, Lim Y. Sentiment Analysis and Sarcasm Detection using Deep Multi-Task Learning. Wireless personal communications, 2023, 129(3): 2213-2237.

[17] Zhang Y, Wang J, Liu Y, Rong L, Zheng Q, Song D, Qin J. A Multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations. Information Fusion, 2023, 93(1): 282-301.

[18] Maity K, Kumar A, Saha S. A Multitask Multimodal Framework for Sentiment and Emotion-Aided Cyberbullying Detection. IEEE Internet Computing, 2022, 26(4): 68-78.

[19] Abdullah T, Ahmet A. Deep learning in sentiment analysis: Recent architectures. ACM Computing Surveys, 2022, 55(8): 1-37.

[20] Hande A, Hegde S U, Chakravarthi B R. Multi-task learning in under-resourced Dravidian languages. Journal of Data, Information and Management, 2022, 4(2): 137-165.

Validation of an Adaptive Decision Support System Framework for Outcome-Based Blended Learning

Rahimah Abd Halim[1], Rosmayati Mohemad[2]*, Noraida Hj Ali[3], Anuar Abu Bakar[4], Hamimah Ujir[5]
Faculty of Computer Science and Mathematics, University Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia[1, 2, 3]
Faculty of Computer Science and Information Technology, University Malaysia Sarawak,
94300 Kota Samarahan, Sarawak, Malaysia[4, 5]

***Abstract*—The Adaptive Decision Support System Learning Framework (A-DSS-LF) was developed to address diverse learner needs in blended learning environments by integrating learning styles, cognitive levels, practical skills, and value practices. This study validates the framework using the Fuzzy Delphi Method (FDM), a consensus-building tool that synthesizes expert opinions and addresses uncertainties in subjective judgments. A panel of 15 experts evaluated the framework's constructs: Learning Process, Learning Assessment, Decision Support System, and Adaptive Learning Profile. All constructs met the FDM's consensus criterion, achieving threshold values between 0.087 and 0.118 (≤0.2), indicating high consistency and low variability. The defuzzification process confirmed values exceeding 0.5, with scores ranging from 0.873 to 0.922 and expert agreement surpassing 75 percent for all elements. These findings confirm the robustness and applicability of the A-DSS-LF, validating its role in enhancing personalized learning outcomes and supporting teachers in tailoring adaptive learning resources. The framework is scalable and can be implemented in secondary school computer science education and online learning platforms to create personalized learning paths, improve engagement, and bridge the gap between online and offline learning. This study reinforces the significance of expert validation in adaptive learning frameworks, ensuring their scalability and adaptability for future applications in diverse educational settings.**

***Keywords*—Learner needs; adaptive learning; blended learning; fuzzy delphi method; decision support system**

I. Introduction

In modern education, learners exhibit diverse needs, preferences, and abilities, requiring tailored educational approaches to enhance engagement and improve learning outcomes. According to study [1] and study [2], adaptive learning frameworks have emerged as a promising solution by leveraging computerized algorithms and data-driven methodologies to personalize learning experiences based on individual learner characteristics. Unlike traditional one-size-fits-all teaching methods, these frameworks dynamically adjust content and instructional strategies, allowing learners to receive materials tailored to their cognitive and behavioral profiles [3] [4]. This adaptability enhances student engagement and optimizes learning outcomes by ensuring that instructional content aligns with individual learning needs.

Particularly in blended learning environments, adaptive learning plays a crucial role in bridging online and offline learning components. As highlighted by study [5] and study [6], these frameworks create a more flexible and structured learning approach, allowing seamless integration between traditional and technology-enhanced learning experiences. This integration ensures that adaptive learning is not only personalized but also scalable and adaptable to different educational settings.

Several adaptive learning frameworks have been developed to facilitate personalized learning experiences, with many operating within Learning Management Systems (LMS). For instance, the study in [7] proposed a three-stage implementation model designed to scale adaptive learning in fully online education, focusing on faculty training and infrastructure development. Similarly, the study in [8] introduced a knowledge-based adaptive model that aligns instructional content with students' proficiency levels, while [9] developed an Adaptive Virtual Learning Environment (AVLE) structured around content, student, and adaptation models to provide personalized learning paths. The Adaptive Learning System-Knowledge Level (ALS-KL) by study [10] utilized pre-test and post-test assessments to classify learners and deliver content suited to their knowledge levels.

Despite their potential, existing adaptive learning frameworks face significant limitations that hinder their effectiveness in diverse educational settings. The studies in[5] and [9] highlight that many models rely heavily on Learning Management Systems (LMS), making them less adaptable for classroom-based learning environments that require seamless integration between online and offline instruction. Additionally, the studies in [11] and [12] emphasize that most frameworks primarily focus on cognitive aspects while neglecting other critical learner attributes, such as practical skills and value-based learning, which are essential for holistic education. The absence of these elements limits the ability of adaptive learning to fully support diverse learner needs.

Furthermore, the studies in [13] and [1] argue that many adaptive learning frameworks lack rigorous validation mechanisms, raising concerns about their effectiveness, scalability, and generalizability in different educational settings. Without systematic validation, these frameworks may fail to achieve consistent learning improvements across varied instructional contexts. Addressing these challenges requires a comprehensive and validated framework that integrates multiple learner characteristics while ensuring practical implementation in blended learning environments.

To bridge these gaps, this study introduces the Adaptive Decision Support System Learning Framework (A-DSS-LF), designed to provide a more holistic and data-driven adaptive

This research was supported by Universiti Malaysia Terengganu (TAPERG/2023/UMT/2564).

learning experience. A-DSS-LF differs from existing frameworks in several ways. Unlike traditional models that focus solely on cognitive skills, A-DSS-LF integrates learning styles, cognitive levels, practical skills, and value practices to offer a well-rounded personalized learning experience. Additionally, while many adaptive learning models are designed primarily for LMS-based environments, A-DSS-LF is structured to seamlessly integrate both online and offline learning environments, making it more suitable for blended learning. Another key feature of A-DSS-LF is its inclusion of a Decision Support System (DSS), which enables educators to make data-driven instructional decisions and adapt learning interventions based on students' needs, a capability often missing in many existing adaptive learning frameworks. To ensure its effectiveness, adaptability, and scalability, A-DSS-LF is rigorously validated using the Fuzzy Delphi Method (FDM), allowing expert consensus to confirm its applicability in real-world educational settings. These features position A-DSS-LF as a scalable and personalized approach to adaptive learning, supporting both student-centered learning and teacher-driven instructional strategies in blended learning environments.

The objective of this study is to present the validation process and findings of the A-DSS-LF using the Fuzzy Delphi Method (FDM). This study aims to establish expert consensus on the framework's constructs and components, ensuring its robustness, scalability, and applicability in blended learning environments. By detailing the validation process, this study examines how expert feedback informs the refinement and validation of A-DSS-LF to align with modern educational needs. Additionally, it evaluates the findings of the analysis, confirming the framework's effectiveness in supporting personalized learning pathways and educator-driven decision-making. Through this validation, the study contributes to the development of rigorously tested adaptive learning frameworks, ensuring their practical implementation in real-world educational settings to enhance adaptive learning experiences and improve instructional decision-making.

The remainder of this article is structured as follows. Section II reviews related work on the FDM, highlighting its significance in expert-based validation. Section III details the FDM validation methodology, including expert selection and analysis procedures. Section IV presents the key findings, followed by a discussion in Section V. Finally, Section VI concludes with implications and future research directions.

II. Related Work

To ensure the robustness and applicability of A-DSS-LF, a rigorous validation method is required. Traditional validation approaches may lack precision in expert-driven refinements, making them less suitable for validating adaptive learning frameworks. To address this challenge, this study employs the Fuzzy Delphi Method (FDM), a structured approach that systematically refines framework components through expert consensus. The following section explores FDM, its significance in validation research, and its applications in education.

A. Fuzzy Delphi Method (FDM)

Given the complexity of adaptive learning frameworks, robust validation measures are essential to ensure their effectiveness. The Fuzzy Delphi Method (FDM) was selected in this study for its ability to address uncertainty and systematically establish expert consensus. Originally introduced by [14] and later refined by study [15], FDM enhances the traditional Delphi Method by integrating fuzzy logic principles, allowing for quantitative evaluation of expert judgments. The study in [16] demonstrated FDM's extensive applications in education, technology, and policy-making, where iterative refinement of theoretical models is necessary.

Unlike conventional validation approaches, FDM refines framework components iteratively, ensuring expert consensus is achieved through multiple evaluation rounds. This process systematically reduces ambiguity and enhances precision in decision-making [17]. By incorporating expert-driven refinements, FDM ensures A-DSS-LF aligns with best practices in adaptive learning, strengthening its adaptability to blended learning environments.

Several validation methods exist for refining adaptive learning frameworks, yet each has notable limitations. Traditional expert review methods, as described by study [18], rely heavily on qualitative assessments and descriptive feedback, often introducing bias and inconsistencies. FDM, by contrast, provides a structured and quantifiable approach, ensuring expert evaluations are numerically analyzed rather than solely based on subjective agreement. Structural Equation Modeling (SEM), commonly used for model validation [19], requires large datasets and strong statistical assumptions, making it unsuitable for early-stage validation where expert input is prioritized. Similarly, Design-Based Research (DBR) emphasizes real-world implementation through iterative testing [20], but its time-intensive nature makes it less practical for preliminary validation stages.

In contrast, Analytic Hierarchy Process (AHP), as proposed by study [21], is widely used for ranking framework components based on weighted criteria. However, it lacks iterative expert feedback loops, making it less effective for dynamically evolving frameworks such as A-DSS-LF. Pilot testing with end-users, though essential for usability validation, is more beneficial in later stages once a framework has been theoretically refined [22]. Given these comparisons, FDM emerges as the most suitable validation method for A-DSS-LF, as it ensures a balance between theoretical validation and expert-driven iterative refinement.

The effectiveness of FDM has been widely demonstrated in prior research. The studies in [17] and [23] applied FDM to validate STEM teaching modules and immersive learning innovations, ensuring that expert recommendations were systematically incorporated into framework refinements. Similarly, the study in [24] demonstrated FDM's utility in validating hybrid learning strategies, synthesizing diverse expert opinions while maintaining theoretical and practical relevance. [25] and [16] confirm that FDM's consensus thresholds—such as a defuzzification coefficient ($d \leq 0.2$) and expert agreement exceeding 75%—enhance reliability and consistency in validation outcomes. These findings reinforce FDM's reliability as a consensus-driven method, confirming its adaptability to various educational settings.

While FDM serves as the primary validation method, complementary approaches such as DBR, AHP, and Pilot Testing contribute to specific aspects of framework validation. DBR enables iterative real-world refinement [20], AHP prioritizes framework components systematically [21], and pilot testing gathers usability insights for final-stage improvements [22]. Together, these methods contribute to a robust validation process, ensuring both theoretical soundness and practical applicability.

FDM's ability to quantify expert judgments, structure consensus, and support iterative refinements makes it an indispensable validation tool for educational research. Prior studies [17], [23], and [24] confirm its effectiveness in aligning theoretical models with real-world applications. Furthermore, FDM's alignment with contemporary challenges, including hybrid learning environments and emerging educational technologies, reinforces its continued relevance as a critical validation method for adaptive learning frameworks.

III. METHODOLOGY

As previously discussed, this study applies the FDM, introduced by study [15] to validate the proposed A-DSS-LF. The validation process follows a three-phase approach to ensure a systematic and structured evaluation of the framework. The following section provides a detailed explanation of each phase.

A. Expert Selection

To achieve consensus on the developed framework, purposive sampling was employed, a method particularly suited to the FDM, as noted by study [18]. The sample for this study comprised experts in the field of education. According to [26], an expert is an individual with extensive knowledge and skills in a specific domain, which in this context refers to subject matter experts. A panel of 23 experts was selected based on their roles, years of experience, and areas of expertise within the field of education. The number of experts selected aligns with study [18] recommendation that 10 to 50 participants are sufficient when the group is relatively uniform or homogenous. The selection criteria included educators with a minimum of 10 years of relevant experience, familiarity with teaching and learning practices, and holding various roles such as Head of Subject Panel, Chief Subject Assessor, School Improvement Specialist Coaches (SISC+), professors, Assistant Education Officers, and Senior Subject Teacher. This diverse panel ensured a comprehensive and well-rounded evaluation of the framework.

B. A-DSS-LF Expert Validation

The process was implemented in four key stages. The first stage involved the presentation of the A-DSS-LF to the expert panel through an online session. This presentation provided an overview of the framework, detailing its components, objectives, and application within blended learning environments. The session aimed to ensure that experts thoroughly understood the framework, enabling them to offer informed and constructive feedback. Following the presentation, a structured questionnaire was distributed to the experts. The questionnaire gathered insights on the framework's relevance, feasibility, and practicality. It incorporated closed-ended questions, evaluated using a fuzzy Likert scale, and used open-ended items to capture qualitative feedback. This combination ensured a comprehensive assessment of the framework.

The 7-point Likert scale was used in this research to identify the constructs and elements of the A-DSS-LF, similar to the approach taken by study [23] and study [27]. The study in [23] utilized the 7-point Likert scale to develop constructs and elements for a framework, emphasizing its accuracy and ability to reduce ambiguity compared to a 5-point scale. Similarly, the study in [27] employed a 7-point Likert scale in their study to evaluate expert agreement on mobile learning implementation in competency-based education, analysing responses using fuzzy logic techniques. These precedents highlight the scale's suitability for capturing nuanced expert feedback in framework development. The linguistic variables were aligned with fuzzy scales to facilitate expert responses, as shown in Table I. The table illustrates the mapping of linguistic variables (e.g., "Strongly disagree," "Moderately agree") to their corresponding Likert scale values and fuzzy scale representations. This approach ensures that the fuzzy logic analysis is grounded in systematically quantified expert input, enhancing the precision of consensus measurement.

TABLE I. LINGUISTIC VARIABLE SCALE

Linguistic Variables	Likert Scale	Fuzzy Scale
Strongly disagree	1	(0.0,0.0,0.1)
Moderately disagree	2	(0.0,0.1,0.1)
Slightly disagree	3	(0.1,0.3,0.5)
Neutral	4	(0.3,0.5,0.7)
Slightly agree	5	(0.5,0.7,0.9)
Moderately agree	6	(0.7,0.9,1.0)
Strongly agree	7	(0.9,1.0,1.0)

Source: [28]

In the third stage, feedback collection was conducted. Experts submitted responses that included both quantitative and qualitative data. Quantitative feedback measured levels of agreement on various aspects of the framework, while qualitative feedback provided additional insights and actionable suggestions for refinement. Finally, the application of fuzzy logic for analysis was carried out. The collected data was analysed using fuzzy logic principles to measure the degree of expert consensus systematically. This structured approach ensured that the validation of the A-DSS-LF was grounded in expert consensus, enhancing its robustness and practical applicability in addressing diverse learner needs in blended learning environments.

C. Validation of Constructs and Elements Through Fuzzy Delphi Analysis for the A-DSS-LF

The analysis of questionnaire data using the FDM involved three main steps: applying Triangular Fuzzy Numbers, calculating the Expert Consensus Percentage, and performing the Defuzzification Process. These steps systematically assessed the expert responses to validate the constructs and elements of the A-DSS-LF. The linguistic variable data obtained from experts in this study, as shown in Table I, must be converted into Triangular Fuzzy Numbers. The Triangular Fuzzy Number has three values, m_1, m_2, and m_3, indicating the minimum, reasonable, and maximum values, as shown in Fig. 1.

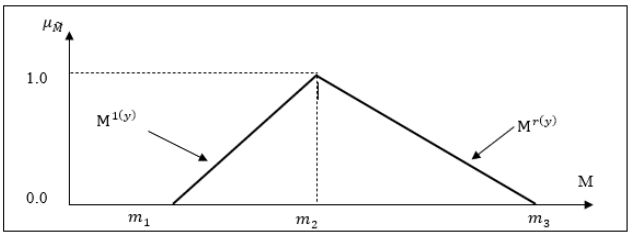


Fig. 1. The Triangular fuzzy number.

Next, the threshold value (*d***)** is calculated to measure the dispersion of expert opinions. An element is considered to have achieved expert consensus if $d \leq 0.2$, meaning that every element with a threshold value (*d*) equal to or less than 0.2 is accepted. As shown in Fig. 1, the threshold value (*d*) is calculated using Formula (1).

$$(\tilde{m}\ \tilde{n}) = \sqrt{\frac{1}{3}[(m_1 - n_1)^2 + (m_2 - n_2)^2 + (m_3 - n_3)^2]} \quad (1)$$

The second step involves calculating the percentage of expert agreement. According to the traditional Delphi technique, an item is accepted if the percentage of agreement among the expert group exceeds 75% [23], [29], [30]. Another requirement in the FDM is the defuzzification process. This step involves analysing the data by averaging fuzzy numbers to calculate the Fuzzy score (A). The Fuzzy score (A) must be greater than or equal to the median value (α-cut value) of 0.5 [19], [31], indicating that the element has achieved expert consensus. The Fuzzy score (A) determines the ranking and identifies acceptable elements based on expert agreement. An item is accepted if the Fuzzy score (A) is equal to or greater than 0.5; otherwise, it is rejected. The fuzzy score (A) is calculated using the formula shown in Formula (2).

$$A = (1/3) * (\mu_1 + \mu_2 + \mu_3) \quad (2)$$

The summary of conditions for the acceptable elements for the A-DSS-LF is shown in Table II.

TABLE II. KEY METRICS FOR FUZZY LOGIC ANALYSIS

Metric	Description	Conditions
Threshold Value (*d*)	Measures the dispersion of expert opinions. Consensus is achieved if $d \leq 0.2$.	$d \leq 0.2$
Percentage Agreement	The proportion of experts agreeing on an element.	> 75%
Defuzzification	Converts fuzzy values into crisp numbers for interpretation.	≥ α-cut value = 0.5

Following the criteria outlined in Table II, the analysis confirmed that elements meeting both the threshold value ($d \leq 0.2$) and achieving a Fuzzy score ($A \geq 0.5$) were accepted. Additionally, elements that achieved an expert consensus percentage of more than 75% were validated. These combined criteria include only elements with strong expert agreement and alignment. The ranked elements provide valuable insights into the framework's relevance and feasibility, reinforcing its robustness in addressing diverse learner needs.

IV. RESULT

This section presents the findings of the FDM analysis conducted to validate the constructs and elements of the A-DSS-LF. The findings include an overview of the expert demographic information and the outcomes of the FDM validation process.

A. Expert Demographic Information

Due to scheduling challenges and time constraints, only 15 experts were available for the final discussion session. This number still falls within the range of 10 to 15 experts suggested by study [32] for achieving a reliable consensus when the expert group is not homogeneous. The expert panel in this study reflects a diverse range of qualifications, expertise, and professional experience, ensuring robust and informed feedback during the validation process. Table III summarises their demographics, categorised into three key areas: Level of Education, Work Experience, and Field of Expertise.

TABLE III. SUMMARY OF EXPERT PANEL

Level of Education	Frequency
PhD	2
Master Degree	4
Bachelor Degree	9
Total	15
Work Experience (Years)	**Frequency**
11 to 15 years	1
16 to 20 years	5
More than 20 years	9
Total	15
Field of Experts	**Frequency**
Information Technology	1
Business Management	2
Education Technology	1
Coaching	1
Science and Mathematics	2
Accounting	1
Multimedia	1
History	1
Computer Science	1
Languages	2
Physical Education	1
Decision Support Systems	1
Total	15

This diversity of backgrounds, spanning fields such as educational technology, decision support systems, computer science, and language studies, adds significant value to the consensus-building process. It ensures a comprehensive evaluation of the A-DSS-LF and its applicability in blended learning environments.

B. Expert Consensus Findings Using the FDM

Data analysis from the closed-ended questionnaire was conducted systematically using a Microsoft Excel data sheet developed by study [33]. This framework has four constructs:

the Learning Process, the Learning Assessment, the Decision Support System, and the Adaptive Learning Profile. Each construct comprises three elements, except for the Learning Assessment, which includes seven

Refer to Table V, which shows th. The elements given to the experts are stated in Table IV.at the threshold value (*d*) for each construct is below the acceptable limit ($d \leq 0.2$), indicating that all constructs meet the FDM qualification criteria. Specifically, the overall threshold value (*d*) for the Learning Process construct is $d=0.114$, the Learning Assessment construct is $d=0.094$, the Decision Support System construct is d=0.092, and the Adaptive Learning Profile construct is $d=0.097$. These values confirm that all constructs are acceptable based on the Fuzzy Delphi process.

In addition, individual elements within the constructs also meet the Fuzzy qualification requirement, with $d \leq 0.2$ for each element. The details are below:

- Learning Process elements: E1 (d=0.111), E2 (d=0.112), and E3 (d=0.118) are all acceptable.
- Learning Assessment elements: E4 (d=0.094), E5 (d=0.087), E6 (d=0.094), E7 (d=0.112), E8
- (d=0.089), E9 (d=0.094), and E10 (d=0.094) meet the requirement.
- Decision Support System elements: E11 (d=0.087), E12 (d=0.087), and E13 (d=0.102) are within the threshold.
- Adaptive Learning Profile elements: E14 (d=0.108), E15 (d=0.092), and E16 (d=0.092) are also acceptable.

TABLE IV. ELEMENTS FOR THE A-DSS-LF ACCORDING TO THE CONSTRUCTS

Learning Process	
E1	Apply adaptive learning in the blended learning environment.
E2	Apply student-centred approach
E3	Apply self-regulated learning
Learning Assessment	
E4	Identify students' learning styles.
E5	Apply a learning style model to determine students' learning styles.
E6	Use systematic instruments to assess students' learning styles.
E7	Use systematic instruments to evaluate student's learning status.
E8	Use systematic instruments capable of measuring specific learning objectives.
E9	Use systematic instruments to assess various learning domains, including cognitive skills, practical skills, and value practices.
E10	Use systematic instruments with clear criteria to determine students' mastery levels for various learning domains.
Decision Support System	
E11	Utilise identified student characteristics (student model) to make adaptive decisions.
E12	Use a collection of information (learning object model) to support adaptive decision-making.
E13	Apply criteria to align student characteristics with available information for making adaptive decisions.
Adaptive Learning Profile	
E14	Recommend personalized learning paths based on students' characteristics.
E15	Provide tailored feedback aligned with students' characteristics.
E16	Suggest learning resources that match students' characteristics.

TABLE V. SUMMARY OF THRESHOLD VALUE FOR CONSTRUCTS AND ELEMENTS IN A-DSS-LF

Experts	**Learning Process**			**Learning Assessment**							**Decision Support System**			**Adaptive Learning Profile**		
	E1	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**E8**	**E9**	**E10**	**E11**	**E12**	**E13**	**E14**	**E15**	**E16**
1	0.102	0.092	0.138	0.132	0.107	0.086	0.092	0.096	0.066	0.086	0.107	0.107	0.122	0.112	0.096	0.096
2	0.055	0.065	0.031	0.027	0.047	0.067	0.065	0.057	0.066	0.067	0.047	0.047	0.036	0.045	0.057	0.057
3	0.102	0.092	0.138	0.132	0.107	0.086	0.092	0.096	0.066	0.086	0.107	0.107	0.122	0.112	0.096	0.096
4	0.102	0.092	0.138	0.132	0.107	0.086	0.092	0.096	0.066	0.086	0.047	0.047	0.036	0.045	0.096	0.096
5	0.055	0.092	0.031	0.027	0.047	0.067	0.301	0.057	0.088	0.067	0.047	0.047	0.273	0.282	0.057	0.057
6	0.102	0.301	0.256	0.027	0.047	0.067	0.092	0.096	0.066	0.086	0.047	0.047	0.036	0.112	0.096	0.096
7	0.102	0.092	0.138	0.132	0.107	0.086	0.092	0.096	0.066	0.086	0.107	0.107	0.122	0.112	0.096	0.096
8	0.102	0.092	0.138	0.263	0.047	0.086	0.092	0.096	0.066	0.086	0.107	0.107	0.122	0.112	0.096	0.096
9	0.102	0.092	0.031	0.027	0.047	0.086	0.092	0.057	0.066	0.086	0.047	0.107	0.036	0.112	0.057	0.057
10	0.292	0.065	0.256	0.027	0.289	0.308	0.065	0.057	0.327	0.067	0.047	0.047	0.273	0.045	0.057	0.057
11	0.055	0.065	0.256	0.027	0.047	0.067	0.301	0.299	0.088	0.308	0.047	0.289	0.036	0.282	0.299	0.299
12	0.292	0.301	0.031	0.263	0.047	0.067	0.092	0.057	0.088	0.067	0.047	0.047	0.036	0.045	0.057	0.057
13	0.055	0.092	0.031	0.027	0.107	0.086	0.065	0.057	0.088	0.067	0.107	0.047	0.122	0.045	0.057	0.057
14	0.055	0.065	0.031	0.027	0.047	0.067	0.065	0.057	0.066	0.067	0.289	0.047	0.036	0.045	0.057	0.057
15	0.102	0.092	0.138	0.132	0.107	0.086	0.092	0.096	0.066	0.086	0.107	0.107	0.122	0.112	0.096	0.096
Threshold Value (*d*) for each	**0.111**	**0.112**	**0.118**	**0.094**	**0.087**	**0.094**	**0.112**	**0.092**	**0.089**	**0.094**	**0.087**	**0.087**	**0.102**	**0.108**	**0.092**	**0.092**
Value (*d*) Construct	**0.114**			**0.094**							**0.092**			**0.097**		

Furthermore, the Fuzzy qualification requirement includes the percentage of expert consensus, which must exceed more than 75% for each element. The results demonstrate that all items meet this additional criterion, ensuring strong expert agreement for all elements and constructs. The threshold value (*d*), expert consensus percentage, defuzzification, and item position for the above elements are shown in Table VI.

Table V summarises the defuzzification process for the four constructs: Learning Process, Learning Assessment, Decision Support System, and Adaptive Learning Profile. The findings provide insight into the priority and significance of each element, as detailed below:

1) Learning process: The Learning Process construct comprises three elements, with Fuzzy scores (A) ranging from 0.873 to 0.904. All elements exceeded the FDM's α-cut value of A≥0.5 and met the expert consensus benchmark of more than 75%, confirming their acceptability and inclusion in the framework. The highest-ranked element, E2 (A=0.904, Rank1), demonstrates the most substantial expert agreement. E1 (A=0.898, Rank2) follows closely, reflecting its prioritization within the construct. Although ranked lowest, E3 (A=0.873, Rank3) still satisfies the Fuzzy Delphi criteria, validating its inclusion.

2) Learning assessment: The Learning Assessment construct comprises seven elements, with Fuzzy scores (A) ranging from 0.878 to 0.922. All elements exceeded the FDM's α-cut value of A≥0. and met the expert consensus benchmark of more than 75%, confirming their acceptability and inclusion in the framework. The highest-ranked element, E9 (A=0.922, Rank1), reflects the most substantial expert agreement, likely due to its alignment with the primary objectives of the construct. E6 (A=0.909, Rank2) and E10 (A=0.909, Rank2) share the second rank, highlighting their equal importance. E7 (A=0.904, Rank4) and E8 (A=0.902, Rank5) follow closely, demonstrating high levels of expert agreement. Although ranked lower, E5 (A=0.896, Rank6) and E4 (A=0.878, Rank7) remain valid, contributing to the construct's comprehensiveness.

3) Decision support system: The Decision Support System construct consists of three elements, with Fuzzy scores (A) ranging from 0.884 to 0.896. All elements exceeded the FDM's α-cut value of A≥0.5 and met the expert consensus benchmark of more than 75%, validating their inclusion in the framework. The highest-ranked element, E11 (A=0.896, Rank1), reflects strong expert prioritization. E12 (A=0.884, Rank2) and E13 (A=0.884, Rank2) share the second rank, indicating equal agreement and relevance within the construct.

4) Adaptive learning profile: The Adaptive Learning Profile construct comprises three elements, with Fuzzy scores (A) ranging from 0.891 to 0.902. All elements exceeded the FDM's α-cut value of A≥0.5 and met the expert consensus benchmark of more than 75%, confirming their acceptability. E15 (A=0.902, Rank1) and E16 (A=0.902, Rank1) share the top rank, reflecting the most substantial expert agreement and prioritization. Although ranked lower, E14 (A=0.891, Rank3) remains valid and aligned with the Fuzzy Delphi criteria, validating its inclusion in the construct.

TABLE VI. SUMMARY OF THE DEFUZZIFICATION PROCESS FOR CONSTRUCTS AND ELEMENTS IN A-DSS-LF

Elements	Triangular Fuzzy Numbers		Defuzzification Process				Expert Consensus	Acceptable Element	Ranking
	Threshold, *d*, value	% Expert Consensus	m_1	m_2	m_3	Fuzzy Score (A)			
Learning Process									
1	0.111	87%	0.780	0.927	0.987	0.898	Accepted	0.898	2
2	0.112	87%	0.793	0.933	0.987	0.904	Accepted	0.904	1
3	0.118	80%	0.740	0.900	0.980	0.873	Accepted	0.873	3
Learning Assessment									
4	0.094	87%	0.740	0.907	0.987	0.878	Accepted	0.878	7
5	0.087	93%	0.767	0.927	0.993	0.896	Accepted	0.896	6
6	0.094	93%	0.793	0.940	0.993	0.909	Accepted	0.909	2
7	0.112	87%	0.793	0.933	0.987	0.904	Accepted	0.904	4
8	0.092	93%	0.780	0.933	0.993	0.902	Accepted	0.902	5
9	0.089	93%	0.820	0.953	0.993	0.922	Accepted	0.922	1
10	0.094	93%	0.793	0.940	0.993	0.909	Accepted	0.909	2
Decision Support System									
11	0.087	93%	0.767	0.927	0.993	0.896	Accepted	0.896	2
12	0.087	93%	0.767	0.927	0.993	0.896	Accepted	0.896	1
13	0.102	87%	0.753	0.913	0.987	0.884	Accepted	0.884	3
Adaptive Learning Profile									
14	0.108	87%	0.767	0.920	0.987	0.891	Accepted	0.891	3
15	0.092	93%	0.780	0.933	0.993	0.902	Accepted	0.902	1
16	0.092	93%	0.780	0.933	0.993	0.902	Accepted	0.902	1

All elements' Fuzzy scores (A) range from 0.873 to 0.922, exceeding the FDM's α-cut value of A≥0.5 and the more than 75% expert consensus benchmark, confirming their validity. The findings highlight the prioritization of significant elements, such as E2 in the Learning Process, E9 in the Learning Assessment, E11 in the Decision Support System, and E15/E16 in the Adaptive Learning Profile. These results provide a robust foundation for the A-DSS-LF, ensuring its alignment with expert consensus and relevance in addressing diverse learner needs.

Table VII presents the summary ranking of all elements. According to Table VII, experts reached the highest agreement on E9 (A=0.922), emphasizing the importance of using systematic instruments to assess various learning domains. This element is the most essential component in the proposed A-DSS-LF framework.

TABLE VII. SUMMARY RANKING OF ALL A-DSS-LF ELEMENTS

Ranking	Acceptable Element	Elements
1	0.922	E9
2	0.909	E6
2	0.909	E10
4	0.904	E2
4	0.904	E7
6	0.902	E8
6	0.902	E15
6	0.902	E16
9	0.898	E1
10	0.896	E12
11	0.896	E5
11	0.896	E11
13	0.891	E14
14	0.884	E13
15	0.878	E4
16	0.873	E3

V. DISCUSSION

This section discusses the implications of the findings from the FDM analysis in validating the A-DSS-LF. It interprets the results, highlighting how the validated constructs and elements align with the study's objectives and contribute to a robust framework. Furthermore, the discussion explores the relevance of these findings in supporting personalized learning pathways within blended learning environments. Lastly, it identifies limitations and suggests directions for future research to enhance the framework's applicability and impact.

A. Expert Panel and Its Role in FDM Validation

In this study, the number of experts on the panel, 15, for the FDM process was deemed acceptable, although it did not fully meet the desired target. Prior research supports this approach, as studies such as [27], [23], [29], and [17] have employed expert panels ranging from 11 to 17 participants to achieve consensus on educational frameworks. Selecting an appropriate number of experts is critical to ensuring diverse perspectives, reliable consensus, and statistical robustness. This study's selection of 15 experts aligns with established FDM practices, thereby enhancing the credibility of the findings.

B. FDM Validation of A-DSS-LF

This section discusses the findings from the FDM validation, confirming the relevance of the A-DSS-LF. The results demonstrate that all elements across the four key constructs—Learning Process, Learning Assessment, Decision Support System, and Adaptive Learning Profile—achieved strong expert consensus. Each element successfully fulfilled the three essential criteria for FDM analysis:

- Meeting the required threshold value ($d \leq 0.2$).
- Achieving an expert agreement of >75%.
- Exceeding the α-cut defuzzification score ($A \geq 0.5$).

These findings validate the A-DSS-LF framework's ability to support adaptive learning in blended environments, reinforcing its role in enhancing engagement, personalization, and data-driven decision-making. The interactions between the four constructs—Learning Process, Learning Assessment, Decision Support System, and Adaptive Learning Profile—are illustrated in Fig. 2, showing how these components collectively enable adaptive learning experiences.

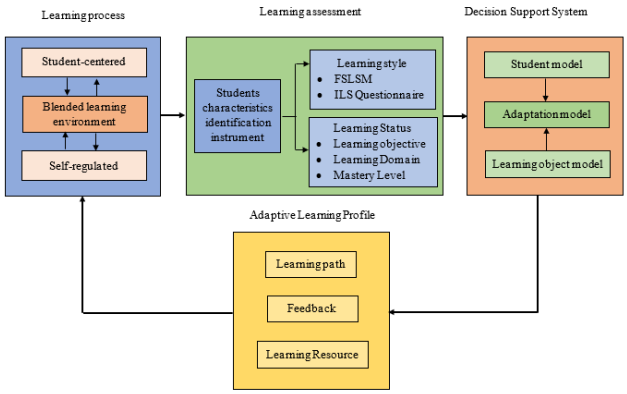


Fig. 2. Adaptive-decision support system-learning framework.

The following subsections provide a detailed discussion of the research findings, focusing on expert validation for each construct and the significance of the validated elements in supporting adaptive learning within the A-DSS-LF framework.

1) Learning process: The effectiveness of a structured learning process within a blended learning environment depends on its ability to adapt to individual learner needs, promote student-centered learning, and encourage self-regulation. Expert validation conducted in this study confirms the importance of these three key elements—blended learning environment, student-centered approaches, and self-regulated learning—as fundamental components in the learning process within the A-DSS-LF. All elements surpassed the α-cut threshold ($A \geq 0.5$) and the 75% consensus benchmark, confirming their necessity in the A-DSS-LF framework. Experts agreed that integrating these elements enhances learner

engagement and autonomy while aligning with best practices in adaptive learning and blended education models.

Among these, applying a student-centered approach (A = 0.904, Rank 1) received the highest expert consensus, emphasizing its role in fostering learner autonomy, engagement, and decision-making. The findings indicate that placing students at the center of the learning process improves motivation and enhances participation, reinforcing the importance of active learning strategies. This aligns with prior research, which highlights that student-centered learning fosters self-directed learning and critical thinking [34][35][36]. Applying adaptive learning in the blended learning environment (A = 0.898, Rank 2) was also strongly supported, as experts acknowledged that personalized learning paths, tailored feedback, and flexible progression help accommodate diverse learning needs. Previous studies similarly emphasize that adaptive learning improves instructional effectiveness by allowing students to advance at their own pace while receiving personalized support [7][37].

Although applying self-regulated learning (A = 0.873, Rank 3) ranked lowest, experts agreed on its significance in fostering learner independence, metacognitive skills, and academic performance. The findings suggest that students who engage in self-regulated learning demonstrate greater persistence and improved learning outcomes, particularly when they receive structured feedback and tracking tools. This is consistent with [38], who found that students with strong self-regulation skills achieve higher engagement and academic success in digital learning environments. Additionally, [39] emphasize that goal setting, adaptive scaffolding, and self-monitoring mechanisms are crucial in supporting self-regulation, reinforcing the importance of incorporating these strategies into the A-DSS-LF framework.

2) Learning assessment: A practical learning assessment is crucial for understanding student progress, identifying learning gaps, and personalizing instructional strategies. In Malaysian secondary education, the National Philosophy of Education emphasizes balancing cognitive, practical, and value-based domains. The FDM validation in this study confirms the importance of systematic learning assessments, with all elements surpassing the α-cut threshold ($A \geq 0.5$) and the 75% consensus benchmark, validating their inclusion in the A-DSS-LF.

Among these elements, systematic instruments with clear criteria to determine students' mastery levels (A = 0.922, Rank 1) received the most substantial expert consensus, highlighting the need for structured assessment frameworks. This aligns with [40], who demonstrated the effectiveness of DSS in behavioral modeling for personalized learning assessments. Similarly, [41] emphasizes that clear mastery criteria in formative assessments enhance student motivation and outcomes.

Assessing various learning domains (A = 0.909, Rank 2) also gained strong support, reinforcing the need for comprehensive, multidimensional assessments beyond cognitive evaluation to practical skills and value-based learning. The study in [1] validated this approach, showing that adaptive learning paths based on learner profiles improve engagement and achievement, a perspective supported by study [42], who highlight that digital assessment tools enhance data-driven instructional adjustments.

Identifying students' learning styles (A = 0.904, Rank 4) and applying a learning style model (A = 0.902, Rank 5) were also validated, reinforcing the importance of personalized education. The Felder-Silverman Learning Style Model (FSLSM) is a core component of the A-DSS-LF framework, aligning with studies that emphasize tailored instructional strategies [37] [43]. [44] further supports these findings, demonstrating that systematic learning style classification enhances adaptive learning effectiveness, ensuring that students receive personalized support, leading to better engagement and learning outcomes.

3) Decision support system: The DSS is a pivotal component of the A-DSS-LF, enabling adaptive learning decisions based on structured data analysis. The FDM validation confirms the DSS's importance, with all elements surpassing the α-cut threshold ($A \geq 0.5$) and the 75% consensus benchmark, validating its critical role in the framework.

Among these elements, utilizing identified student characteristics (A = 0.896, Rank 1) received the most substantial expert consensus, emphasizing the need for tailoring learning experiences based on student profiles. This aligns with [45], who highlighted the potential of ontology-based DSS for predictive learning and [46], who demonstrated that tailored educational technologies enhance engagement and satisfaction.

Experts also validated using a collection of information (A = 0.884, Rank 2) and applying criteria to align student characteristics with learning resources (A = 0.884, Rank 2), receiving equal expert consensus. These findings reinforce the importance of structured decision-making in adaptive learning, ensuring that DSS-driven systems effectively match learner needs with relevant instructional content. [47] emphasized that DSS enhances personalized learning by leveraging student models and learning resources and optimizing adaptive learning pathways. Similarly, the study in [9] highlighted that student models, content adaptation, and structured decision-making criteria are crucial in delivering tailored learning experiences, ensuring student profiles align with instructional content in adaptive systems.

Additionally, the study in [48] demonstrated that fuzzy weight-based rule systems dynamically adjust content complexity and volume based on cognitive-level analysis, reinforcing the importance of structured alignment between learner characteristics and learning resources. By integrating decision-making criteria with student profiles and learning objects, DSS ensures that personalized learning paths evolve dynamically, improving instructional effectiveness and learner engagement.

4) Adaptive learning profile: The Adaptive Learning Profile received unanimous expert agreement, reinforcing its critical role in enhancing personalized learning for both students and teachers. This construct focuses on three key elements: personalized learning paths, tailored feedback, and adaptive learning resources, ensuring data-driven, individualized learning experiences. The FDM validation confirms the importance of these elements, with all components

exceeding the α-cut threshold (A ≥ 0.5) and the 75% expert consensus benchmark, validating their essential role in the A-DSS-LF framework.

Among these elements, providing tailored feedback aligned with students' characteristics (A = 0.902, Rank 1) and suggesting learning resources that match students' characteristics (A = 0.902, Rank 1) received the highest expert consensus, reflecting their strong prioritization in adaptive learning environments. The findings emphasize that actionable feedback improves engagement and instructional strategies, ensuring that students receive personalized insights into their learning progress [9] [48]. Additionally, adaptive learning resources enhance accessibility and relevance by offering content aligned with students' cognitive levels and learning styles, reinforcing the importance of adaptive content curation [43].

Although recommending personalized learning paths based on students' characteristics (A = 0.891, Rank 3) was ranked slightly lower, it remains a fundamental component of adaptive learning. Personalized learning paths allow students to progress at their own pace, addressing knowledge gaps and supporting competency-based progression [2] [49]. The expert validation results confirm that while learning path recommendations are essential, they are most effective when paired with tailored feedback and adaptive learning resources, ensuring holistic and personalized educational outcomes.

VI. Conclusion

This study aimed to validate the A-DSS-LF using the FDM. Through the involvement of expert consensus, the findings confirmed the credibility and applicability of the proposed framework in supporting personalized adaptive learning within the blended learning environment. The FDM approach ensured robust validation of the constructs and elements, as evidenced in similar educational contexts where it effectively built consensus and established framework reliability.

The FDM results demonstrated a high level of agreement among experts, particularly regarding the essential components of the framework: the learner model, adaptation model, and learning object model. These findings align with studies emphasizing the significance of adaptive learning frameworks tailored to individual learner profiles. The validated A-DSS-LF provides a structured approach to integrating adaptive learning in blended settings, emphasizing personalized learning paths tailored to individual learner characteristics. Furthermore, this framework contributes to the ongoing dialogue on adaptive educational systems by incorporating unique elements such as value practices and practical skills, extending beyond traditional cognitive assessments.

The results of this study contribute significantly to the field by offering a validated framework tailored to the Malaysian education context, addressing theoretical and practical gaps. It sets a foundation for implementing the A-DSS-LF prototype and testing its impact in real-world education. Future work should refine the framework based on further qualitative feedback and usability testing. Additionally, implementing and evaluating the A-DSS-LF prototype in secondary school settings will provide valuable insights into its practical effectiveness and scalability. This study aims to enhance personalized education and optimize learning outcomes in blended learning environments by advancing adaptive learning frameworks.

Acknowledgment

We thank Universiti Malaysia Terengganu for providing funding support for this project (TAPERG/2023/UMT/2564).

References

[1] I. A. Alshalabi, S. E. Hamada, K. Elleithy, I. Badara, and S. Moslehpour, "Automated adaptive mobile learning system using shortest path algorithm and learning style," Int. J. Interact. Mob. Technol., vol. 12, no. 5, pp. 4–27, 2018, doi: 10.3991/ijim.v12i5.8186.

[2] H. Peng, S. Ma, and J. M. Spector, "Personalized adaptive learning: an emerging pedagogical approach enabled by a smart learning environment," Smart Learn. Environ., vol. 6, no. 1, pp. 1–14, 2019, doi: 10.1186/s40561-019-0089-y.

[3] N. Yotaman, K. Osathanunkul, P. Khoenkaw, and P. Pramokchon, "Teaching Support System by Clustering Students According to Learning Styles," in 2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON), 2020, pp. 137–140.

[4] I. Katsaris and N. Vidakis, "Adaptive e-learning systems through learning styles : A review of the literature Literature review Adaptive e-learning systems," Adv. Mob. Learn. Educ. Res., vol. 1, no. 2, pp. 124–145, 2021, doi: 10.25082/AMLER.2021.02.007.

[5] N. Morze, L. Varchenko-Trotsenko, T. Terletska, and E. Smyrnova-Trybulska, "Implementation of adaptive learning at higher education institutions by means of Moodle LMS," J. Phys. Conf. Ser., vol. 1840, no. 1, pp. 1–13, 2021, doi: 10.1088/1742-6596/1840/1/012062.

[6] B. Arsovic and N. Stefanovic, "E-learning based on the adaptive learning model: case study in Serbia," Sādhanā, vol. 45, no. 1, p. 266, 2020, doi: 10.1007/s12046-020-01499-8S.

[7] R. Abadia and S. Liu, "Low Adoption of Adaptive Learning Systems in Higher Education and How Can It Be Increased in Fully Online Courses," in 29th International Conference on Computers in Education Conference, ICCE 2021 - Proceedings, 2021, pp. 569–578.

[8] H. K. M. Al-Chalabi, A. M. A. Hussein, and U. C. Apoki, "An Adaptive Learning System Based on Learner ' s Knowledge Level," in 2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), IEE, 2021, pp. 1–4. doi: 10.1109/ECAI52376.2021.9515158.

[9] M. T. Alshammari, "Design and evaluation of an adaptive framework for virtual learning environments International Journal of Advanced and Applied Sciences," Int. J. Adv. Appl. Sci., vol. 7, no. 5, pp. 39–51, 2020, doi: 10.21833/ijaas.2020.05.006.

[10] S. Sfenrianto, Y. B. Hartarto, H. Akbar, M. Mukhtar, E. Efriadi, and M. Wahyudi, "An Adaptive Learning System based on Knowledge Level for English Learning," Int. J. Emerg. Technol. Learn., vol. 13, no. 12, pp. 191–200, 2018.

[11] F. E. Louhab, A. Bahnasse, F. Bensalah, A. Khiat, Y. Khiat, and M. Talea, "Novel approach for adaptive flipped classroom based on learning management system," Educ. Inf. Technol., vol. 25, pp. 755–773, 2019, doi: 10.1007/s10639-019-09994-0.

[12] M. S. Hasibuan, L. E. Nugroho, and P. I. Santosa, "Model E-learning MDP for Learning Style Detection using prior knowledge," Int. J. Eng. Technol., vol. 7, no. 4, pp. 118–122, 2018, doi: 10.14419/ijet.v7i4.40.24416.

[13] N. S. Raj and V. G. Renumol, "A systematic literature review on adaptive content recommenders in personalized learning environments," J. Comput. Educ., vol. 9, no. 1, pp. 113–148, 2021, doi: 10.1007/s40692-021-00199-4.

[14] T. J. Murray, L. L. Pipino, and J. P. Van Gigch, "A pilot study of fuzzy set modification of delphi," Hum. Syst. Manag., vol. 5, no. 1, pp. 76–80, 1985, doi: 10.3233/HSM-1985-5111.

[15] A. Kaufmann and M. M. Gupta, Fuzzy Mathematical Models in Engineering and Management Science. 1988.

[16] M. R. M. Jamil, A. T. M. Hashim, M. S. Othman, A. M. Ahmad, N. M. Noh, and M. F. M. Kamal, "Digital Pedagogy Policy in Technical and Vocational Education and Training (TVET) in Malaysia : Fuzzy Delphi Approach," J. Tech. Educ. Train., vol. 15, no. 2, pp. 1–10, 2023, doi: 10.30880/jtet.2023.15.02.001.

[17] R. Mustapha, M. Mahmud, N. M. Burhan, H. Awang, P. B. Sannagy, and M. F. Jafar, "An Exploration on Online Learning Challenges in Malaysian Higher Education : The Post COVID-19 Pandemic Outbreak," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 7, pp. 391–398, 2021.

[18] F. Hasson, S. Keeney, and H. McKenna, "Research guidelines for the Delphi survey technique," J. Adv. Nurs., vol. 32, no. 4, 2000, doi: 10.1046/j.1365-2648.2000.t01-1-01567.x.

[19] S. Bodjanova, "Median alpha-levels of a fuzzy number," Fuzzy sets Syst., vol. 157, no. 7, pp. 879–891, 2006.

[20] S. Barab and K. Squire, "Design-based research: Putting a stake in the ground. In Design-based Research," J. Learn. Sci., vol. 13, no. 1, pp. 1–14, 2016.

[21] T. L. Saaty, "The analytic hierarchy process (AHP)," J. Oper. Res. Soc., vol. 41, no. 11, pp. 1073–1076, 1980.

[22] Z. A. Hassan, P. Schattner, and D. Mazza, "Doing A Pilot Study: Why Is It Essential?," Malaysian Fam. physician Off. J. Acad. Fam. Physicians Malaysia, vol. 1, no. 2–3, p. 70, 2006.

[23] S. Jaya, R. Zaharudin, M. N. Yaakob, and M. A. Ithnin, "Application of Fuzzy Delphi Method (FDM) in Development of the Heutagogical and Technological Practices in Next Generation Learning Spaces (NGLS) Framework," J. Soc. Sci. Humanit., vol. 1, no. 2, pp. 39–51, 2022, doi: 10.53797/icccmjssh.v1i2.5.202.

[24] K. Naser, S. Alamassi, Z. Shana, J. Yousef, and S. H. Halili, "Designing of a Flipped STEM Classroom Engineering- Based Module : Fuzzy Delphi Approach," Int. J. Interact. Mob. Technol., vol. 17, no. 10, pp. 4–29, 2023, doi: 10.3991/ijim.v17i10.38217.

[25] L. H. Yeh et al., "Defining the Collaborative-Constructivism Based Learning and Teaching Approach in Malaysian Primary Schools in Supporting the Hybrid Learning of Visual Arts Education : A Fuzzy Delphi Method Study," J. Adv. Res. Appl. Sci. Eng. Technol., vol. 41, no. 2, pp. 62–81, 2024, doi: 10.37934/araset.41.2.6281 62.

[26] J. Baker, K. Lovell, and N. Harris, "How expert are the experts? An exploration of the concept of 'expert'within Delphi panel techniques," Nurse Res., vol. 14, no. 1, 2006.

[27] S. A. Abdullah et al., "Determining Elements in Mobile Learning Implemention Among Instructors in Vocational Colleges : A Fuzzy Delphi Method," IEEE Access, vol. 9, pp. 150839–150845, 2021, doi: 10.1109/ACCESS.2021.3121703.

[28] S. Saedah, M. R. T. L. Abdullah, and R. M. Rozkee, Pendekatan Penyelidikan Rekabentuk dan Pembangunan. 2021.

[29] A. F. Mohamed Yusoff, A. Hashim, N. Muhamad, and W. N. Wan Hamat, "Application of Fuzzy Delphi Technique to Identify the Elements for Designing and Developing the e-PBM PI-Poli Module," Asian J. Univ. Educ., vol. 17, no. 1, pp. 292–304, 2021, doi: 10.24191/ajue.v17i1.12625.

[30] J. W. Murry Jr and J. O. Hammons, "Delphi: A versatile methodology for conducting qualitative research," Rev. High. Educ., vol. 18, no. 4, pp. 423–436, 1995.

[31] G. A. M. Saido, S. Siraj, D. DeWitt, and O. S. Al-Amedy, "Development of an instructional model for higher order thinking in science among secondary school students: a fuzzy Delphi approach," Int. J. Sci. Educ., vol. 40, no. 8, pp. 847–866, 2018, doi: 10.1080/09500693.2018.1452307.

[32] M. Adler and E. Ziglio, Gazing into the oracle: The Delphi method and its application to social policy and public health. Jessica Kingsley Publishers, 1996.

[33] M. J. Mohd Ridhuan and M. N. Nurul Rabihah, Kepelbagaian metodologi dalam penyelidikan rekabentuk dan pembangunan. Qaisar Prestige Resources, 2020.

[34] D. J. P. Shet, "Adaptive and Blended Learning _ the Panacea for the Challenges of E-Learning," J. Emerg. Technol. Innov. Res., vol. 7, no. 5, pp. 707–720, 2020.

[35] B. R. Kerns, "A Case Study Of A Flipped Curriculum Using Collaborative And Active Learning With An Adaptive Learning System," Indiana State University, 2019. [Online]. Available: https://scholars.indianastate.edu/etds/1514

[36] I. Lestari, "The English Teacher ' s Perspective and Challenge on Implementing Merdeka Curriculum," RETORIKA J. Ilmu Bhs., vol. 9, no. 3, pp. 331–339, 2023.

[37] M. T. Alshammari and A. Qtaish, "Effective Adaptive E-Learning Systems According to Learning Style and Knowledge Level," J. Inf. Technol. Educ. Res., vol. 18, pp. 529–547, 2019, doi: 10.28945/4459.

[38] B. Yu, "Self-regulated learning : A key factor in the effectiveness of online learning for second language learners," Front. Psychol., vol. 13:1-51349, pp. 1–6, 2023, doi: 10.3389/fpsyg.2022.1051349.

[39] M. Sinkkonen and A. Tapani, "Review of the Concept ' Self -Regulated Learning ': Defined and Used in Different Educational Contexts," Int. J. Soc. Educ. Sci., vol. 6, no. 1, pp. 130–151, 2024, doi: 10.46328/ijonses.640.

[40] F. Mohd, W. F. F. W. Yahya, S. Ismail, M. A. Jalil, and N. M. M. Noor, "An Architecture of Decision Support System for Visual-Auditory-Kinesthetic (VAK) Learning Styles Detection Through Behavioral Modelling," Int. J. Innov. Enterp. Syst., vol. 3, no. 02, pp. 24–30, 2019.

[41] T. N. Hopfenbeck, "Classroom assessment , pedagogy and learning – twenty years after Black and Wiliam 1998," Assess. Educ. Princ. Policy Pract., vol. 25, no. 6, pp. 545–550, 2018, doi: 10.1080/0969594X.2018.1553695.

[42] S. Rahimi and V. J. Shute, "Stealth assessment : a theoretically grounded and psychometrically sound method to assess , support , and investigate learning in technology - rich environments," Educ. Technol. Res. Dev., vol. 116, no. 106647, 2023, doi: 10.1007/s11423-023-10232-1.

[43] A. Agarwal, D. S. Mishra, and S. V Kolekar, "Knowledge-based recommendation system using semantic web rules based on Learning styles for MOOCs Knowledge-based recommendation system using semantic web rules based on Learning styles for MOOCs," Cogent Eng., vol. 9, no. 1, pp. 1–24, 2022, doi: 10.1080/23311916.2021.2022568.

[44] H. Zhang et al., "A learning style classification approach based on deep belief network for large- scale online education," J. Cloud Comput., vol. 9, pp. 1–17, 2020.

[45] T. Hamim, F. Benabbou, and N. Sael, "An Ontology-based Decision Support System for Multi-objective Prediction Tasks," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 12, pp. 183–191, 2021, doi: 10.14569/IJACSA.2021.0121224.

[46] W. Yu and X. Du, "Implementation of a Blended Learning Model in Content-Based EFL Curriculum," Int. J. Emerg. Technol. Learn., vol. 14, no. 5, pp. 188–199, 2019, doi: 10.3991/ijet.v14i05.9612.

[47] L. T. Muharlisiani, W. G. Mulawarman, R. Rugaiyah, S. N. Azizah, and P. Karuru, "A decision support system for personalized learning in higher education," Al-Ishlah J. Pendidik., vol. 15, no. 4, pp. 5168–5175, 2023.

[48] C. Troussas, A. Krouska, and C. Sgouropoulou, "Collaboration and fuzzy-modeled personalization for mobile game-based learning in higher education," Comput. Educ., vol. 144, 2020, doi: 10.1016/j.compedu.2019.103698.

[49] T. Cavanagh, B. Chen, R. Ait, M. Lahcen, and R. James, "Constructing a Design Framework and Pedagogical Approach for Adaptive Learning in Higher Education : A Practitioner ' s Perspective Constructing a Design Framework and Pedagogical Approach for Adaptive Learning in Higher Education : A Practitioner ' s Per," Int. Rev. Res. Open Distrib. Learn., vol. 211, 2022.

Towards Two-Step Fine-Tuned Abstractive Summarization for Low-Resource Language Using Transformer T5

Salhazan Nasution[1], Ridi Ferdiana[2], Rudy Hartanto[3]
Department of Electrical and Information Engineering,
Faculty of Engineering, Universitas Gadjah Mada, Yogyakarta, Indonesia[1,2,3]
Department of Informatics Engineering-Faculty of Engineering, Universitas Riau, Pekanbaru, Indonesia[1]

***Abstract*—This study explores the potential of two-step fine-tuning for abstractive summarization in a low-resource language, focusing on Indonesian. Leveraging the Transformer-T5 model, the research investigates the impact of transfer learning across two tasks: machine translation and text summarization. Four configurations were evaluated, ranging from zero-shot to two-step fine-tuned models. The evaluation, conducted using the ROUGE metric, shows that the two-step fine-tuned model (T5-MT-SUM) achieved the best performance, with ROUGE-1: 0.7126, ROUGE-2: 0.6416, and ROUGE-L: 0.6816, outperforming all baselines. These findings demonstrate the effectiveness of task transferability in improving abstractive summarization performance for low-resource languages like Indonesian. This study provides a pathway for advancing natural language processing (NLP) in low-resource language through two-step transfer learning.**

***Keywords*—Abstractive summarization; low-resource language; Transformer T5; transfer learning**

I. Introduction

The rapid advancement of technology has accelerated the flow of information, with an increasing number of people opting to consume information digitally rather than through printed media [1]. However, as the volume of information available on the Internet continues fto grow, humans face the challenge of processing and understanding this information within a limited amount of time, while their reading speed is inherently constrained to an average of 300 words per minute [2]. Automatic text summarization systems are, therefore, essential to facilitate faster information retrieval and comprehension.

Text summarization can be categorized into two types: manual summarization and automatic summarization. Manual summarization involves the direct effort of humans, which is inherently labor-intensive, time-consuming, and costly. Consequently, automation is required to produce summaries more efficiently and at a lower cost [3].

Text summarization is a method for generating concise, accurate, and digestible summaries from lengthy textual documents. This technique is commonly encountered in everyday life, such as in news headlines, meeting minutes, movie synopses, or book reviews. The objective of text summarization is to produce a summary from a collection of articles or documents that retains the essential information while being significantly shorter than the original text [4].

Text summarization is a subfield of Natural Language Processing (NLP) and typically employs two main approaches: extractive summarization and abstractive summarization. Extractive summarization involves selecting words, phrases, or sentences directly from the source document, ranking their relevance, and assembling them into a summary [5]. In contrast, abstractive summarization generates summaries by creating new sentences or phrases that convey the same meaning as the source document [6].

Abstractive summarization aids readers in better understanding an article because it generates a new summary by paraphrasing the content. This approach is considered the most ideal, as it holds significant potential for producing human-like summaries [7]. By rephrasing and condensing the source material, abstractive summarization achieves a level of coherence and fluency that extractive methods often lack, making it a highly valuable tool in text summarization research and applications.

Research on text summarization in Indonesia has been conducted across various domains. Studies employing extractive methods include summarization of news articles [8][9][10][11], snippets for search engine results [12], book synopsis summarization [13] using ranking methods such as Maximal Marginal Relevance (MMR) [14] and Text Rank [15], as well as meeting minute summarization [16][17]. In contrast, abstractive summarization research for the Indonesian language remains limited due to resource constraints. Indonesian is categorized as a low-resource language due to the limited availability of datasets [18]. Although Indonesian is recognized as the fourth most widely used language on the Internet, research progress in NLP for this language has been slow due to a lack of resources and datasets [19].

To date, there are only two large-scale datasets available for Indonesian text summarization: IndoSum [20], consisting of 19,000 articles and summaries, and Liputan6 [21], containing 200,000 articles with summaries. Combining these datasets can enhance the model's knowledge by exposing it to more data, patterns, and variations. However, dataset integration carries potential risks, particularly if one dataset is less accurate. Therefore, thorough pre-processing and in-depth data analysis of both datasets are necessary before merging them, along with cross-validation to ensure optimal results.

Research in the field of summarization has seen rapid advancements. Some studies have implemented Bidirectional

Gated Recurrent Unit (BiGRU) to generate summaries of Indonesian-language journals [22]. Other approaches include using Genetic Semantic Graphs [23], Abstract Meaning Representation (AMR) graphs [7], and Semantic Role Labeling (SLR) [24] for summarizing news articles. Wijayanti et al. [25] investigated the performance of pre-trained BERT models and evaluated them using the ROUGE metric [26]. The experimental results revealed that English pre-trained models could generate summaries from Indonesian articles. The utilization of Indonesian pre-trained models in conjunction with the IndoSum dataset, a benchmark for Indonesian text summarizing, produced more effective summaries. Nevertheless, issues such as fragmented sentences, erroneous phrases, and redundant vocabulary were still noted.

The efficacy of transfer learning provides a substantial benefit compared to earlier methodologies. Large Language Models (LLMs) like Transformer-T5 [27], pre-trained on vast datasets, may be tailored for many purposes, including machine translation and summarization. T5, leveraging its transfer learning capabilities, may be fine-tuned and retrained with more specialized datasets for various languages. This adaptability enables T5 to overcome the constraints in Indonesian text summarizing research by integrating machine translation and summarization, so enhancing the model's knowledge and ultimately elevating the accuracy and quality of Indonesian text summaries.

The rapid proliferation of digital content has highlighted the necessity for automatic text summary to enhance information accessibility. Summarization can be classified into extractive and abstractive methods. Extractive summarization directly selects sentences from the source text, whereas abstractive summarization rephrases and condenses the content, resulting in human-like summaries. In spite of its ability to produce coherent and meaningful summaries, abstractive summarization remains a difficult endeavor, particularly for low-resource languages such as Indonesian.

Existing research on Indonesian summarization is largely extractive due to the lack of large-scale, high-quality datasets required for abstractive models. However, advancements in Transformer-based models, particularly the Text-to-Text Transfer Transformer (T5), have introduced a new paradigm for natural language processing (NLP). T5 unifies multiple NLP tasks into a text-to-text framework, enabling it to perform tasks such as summarization and machine translation under the same architecture. Moreover, its transfer learning capabilities allow T5 to adapt to new tasks and languages with minimal data.

This study addresses the research gap in Indonesian abstractive summarization by evaluating the effectiveness of two-step fine-tuning. Using the T5 model, the research investigates the impact of first fine-tuning on machine translation and then on summarization. The evaluation, conducted using the ROUGE metric, provides insights into how task transferability enhances summarization performance in low-resource languages. This study contributes to advancing NLP for Indonesian by presenting an effective method to overcome data scarcity and achieve high-quality abstractive summarization.

The rest of this paper is organized as follows: Section II presents the motivating scenario, discussing the challenges of abstractive summarization in low-resource languages and the rationale behind using the Transformer-T5 model. Section III reviews related work, highlighting previous studies on Indonesian text summarization and transfer learning approaches. Section IV details the methodology, including the two-step fine-tuning approach, dataset descriptions, experimental design, and evaluation metrics. Section V presents the results, comparing different fine-tuning strategies and their impact on summarization performance. Section VI discusses the findings, analyzing the advantages of the proposed approach, the challenges encountered, and potential directions for future research. Finally, Section VII concludes the study by summarizing key insights and contributions.

II. Motivating Scenario

Abstractive summarization, especially in a low-resource language like Indonesian, requires a nuanced understanding of the language's syntax, semantics, and cultural context. The Transformer-T5 model, with its unified text-to-text approach, is well-suited for such tasks, but it must first acquire foundational linguistic knowledge in the target language to perform effectively.

For the T5 model to generate coherent and human-like abstractive summaries in Indonesian, it must be fine-tuned on datasets that introduce it to the language. This step is vital as the model's pre-training on general datasets may not include sufficient exposure to Indonesian. By enabling the model to learn the intricacies of Indonesian through a targeted fine-tuning process, we establish a critical foundation for downstream tasks like summarization.

The most efficient approach to instructing a large language model in a new language is to fine-tune it using a high-quality parallel dataset for machine translation. In particular, a parallel corpus of English and Indonesian can be a valuable source of grammatical structures, idiomatic expressions, and linguistic patterns which are diverse. This fine-tuning phase guarantees that the model can accurately comprehend Indonesian texts and establishes the foundation for abstractive summarization.

In this study, we implement a two-step fine-tuning strategy to reconcile the disparity between task-specific training and foundational language comprehension. Initially, the T5 model is refined on English-Indonesian machine translation duties to enhance its linguistic proficiency. The model is subsequently fine-tuned on an Indonesian abstractive summarization dataset to enhance its summarization capabilities. This sequential fine-tuning approach is intended to optimize the model's performance by utilizing transfer learning to overcome the obstacles presented by the low-resource nature of the Indonesian language.

III. Related Work

Recent studies have explored the use of pre-trained models for Indonesian text summarization as shown in Table I. The previous work [25] fine-tuned BertSumAbs with an Indonesian pre-trained BERT model on the IndoSum dataset, achieving ROUGE-1: 0.67, ROUGE-2: 0.54, and ROUGE-L: 0.65, outperforming English pre-trained BERT models. However, challenges such as meaningless words and repetitive phrases persist. Similarly, [36] examined IndoBERT checkpoints and

TABLE I. COMPARISON OF ROUGE SCORES ACROSS MODELS

No.	Research	Model	Dataset	R-1	R-2	R-L
1	[28]	BART (Augmented)	Liputan6	0.4093	0.3409	0.4037
2	[29]	IndoBERT	Liputan6	0.4235	0.2415	0.3544
3	[30]	ALBERT	IndoSum	0.4528	0.4077	0.4439
4	[31]	EASum (IndoBART + Extractive)	IndoSum	0.3600	0.2300	0.3500
5	[32]	IndoBART (Augmented)	Liputan6	0.3822	0.2079	0.3124
6	[33]	mT5 (Augmented)	Liputan6	0.3856	0.2329	0.3263
7	[34]	BERT2GPT	IndoSum	0.6226	0.5613	0.6043
8	[35]	T5-Large	Wikipedia	0.4738	0.2927	0.4115
9	[36]	BERTSumAbs + IndoBERT-LEM + GPT Decoder	IndoSum	0.6920	0.6135	0.6836
10	[37]	CTRLSum (mBART50)	Liputan6	0.4341	0.2406	0.3963
11	[38]	Transformer (CLS + MWE)	IndoSum	0.4247	0.2226	0.3809
12	[25]	BertSumAbs (IndoBERT)	IndoSum	0.6700	0.5400	0.6500
13	**Our Research**	Transformer-T5 (two-step transfer learning)	IndoSum	**0.7126**	**0.6416**	**0.6816**

found that BERTSumAbs with IndoBERT-LEM and a GPT-like decoder performed best, achieving ROUGE-1: 0.6920, ROUGE-2: 0.6135, and ROUGE-L: 0.6836, highlighting the importance of decoder selection in transformer-based summarization. Beyond BERT models, [30] explored ALBERT, a lightweight variant of BERT, fine-tuning it on IndoSum. The best model achieved ROUGE-1: 0.4528, ROUGE-2: 0.4077, and ROUGE-L: 0.4439, demonstrating a viable alternative for resource-constrained environments while maintaining competitive accuracy.

Some studies have combined extractive and abstractive approaches to improve summarization performance. The author in [31] introduced EASum, a model that first generates extractive summaries using Doc2Vec and cosine similarity, then refines them using IndoBART. On IndoSum, it achieved ROUGE-1: 0.36, ROUGE-2: 0.23, and ROUGE-L: 0.35, while on Liputan6, it performed lower than mBART and IndoGPT. A different hybrid approach was explored in [34], which combined BERT as an encoder with GPT-2 as a decoder. The model, trained on IndoSum, achieved ROUGE-1: 0.6226, ROUGE-2: 0.5613, and ROUGE-L: 0.6043, outperforming BertSum-based models in bi-gram accuracy.

Data augmentation techniques have been widely applied to improve model generalization. The author in [32] utilized synonym replacement-based augmentation and fine-tuning on Liputan6, where the best-performing IndoBART model achieved ROUGE-1: 0.3822, ROUGE-2: 0.2079, and ROUGE-L: 0.3124. Similarly, [33] applied backtranslation-based augmentation on mT5, achieving ROUGE-1: 0.3856, ROUGE-2: 0.2329, and ROUGE-L: 0.3263, demonstrating improvements in coherence and informativeness. Another augmentation-based study, [28], fine-tuned BART on Liputan6, IndoSum, and ChatGPT-augmented summaries, generating over 36,000 new instances. The best model reached ROUGE-1: 0.4093, ROUGE-2: 0.3409, and ROUGE-L: 0.4037, highlighting the effectiveness of multi-dataset fine-tuning.

Cross-lingual and multilingual approaches have also been investigated to expand summarization capabilities. The author in [37] introduced CTRLSum, a keyword-controlled summarization method fine-tuned on mBART50, which achieved ROUGE-1: 0.4341, ROUGE-2: 0.2406, and ROUGE-L: 0.3963 on Liputan6. For cross-lingual summarization (CLS), [38] proposed an end-to-end CLS model integrating multilingual word embeddings (MWE). The model, trained on a translated IndoSum dataset, achieved ROUGE-1: 0.4247, ROUGE-2: 0.2226, and ROUGE-L: 0.3809, improving cross-lingual representation alignment.

Large language models (LLMs) have recently been explored as few-shot summarization tools. The author in [29] evaluated ChatGPT with prompt tuning on Liputan6, comparing it with IndoBART, IndoBERT, and mBART50. While IndoBERT performed best on Liputan6 (ROUGE-1: 0.4235, ROUGE-2: 0.2415, ROUGE-L: 0.3544), ChatGPT had lower automatic evaluation scores but excelled in human evaluation, suggesting superior fluency but lower adherence to reference summaries. Benchmarking efforts have assessed various transformer-based models. The author in [35] compared T5-Large, Pegasus-XSum, and ProphetNet-CNNDM on Wikipedia datasets in English and Indonesian, finding that T5-Large achieved the best ROUGE scores (ROUGE-1: 0.4738, ROUGE-2: 0.2927, ROUGE-L: 0.4115).

Despite the advancements in Indonesian abstractive summarization, most studies either focus on direct fine-tuning of transformer-based models or augmenting datasets to improve performance. However, few works explore multi-task learning approaches that leverage machine translation as a pre-training step for summarization. Existing studies, such as those using IndoBART, mT5, and BERT2GPT, demonstrate the effectiveness of pre-trained models but primarily fine-tune them only on summarization tasks.

This paper addresses this gap by proposing a two-step fine-tuning approach where a model is first fine-tuned on a machine translation task, followed by fine-tuning for abstractive summarization. The hypothesis is that exposing the model to translation tasks helps it learn better language representations and text reformation techniques, which could lead to more coherent, fluent, and information-rich summaries. Unlike previous works that rely solely on pre-trained language models or data augmentation, this research explores whether a structured pre-training strategy with translation improves the performance of Indonesian abstractive summarization models.

IV. METHODS

A. Text-to-Text Transfer Transformer (T5)

The Text-to-Text Transfer Transformer (T5) is a cutting-edge natural language processing (NLP) model developed by Google Research. It is structured to encapsulate all NLP tasks within a cohesive text-to-text format, wherein both the input and output are expressed as sequences of text. This approach enables T5 to do several tasks, including translation,

summarization, categorization, and question answering, with a consistent architecture and training methodology.

- Text-to-Text Format: T5 differentiates itself from conventional models, which are task-specific (such as categorization or generation), by reinterpreting every activity as a text creation challenge. For instance:
  - Translation: Input = "translate English to Indonesian: Where are you?", Output = "Di mana kamu?".
  - Summarization: Input = "summarize: The article discusses...", Output = "Key points of the article...".
  - Classification: Input = "classify sentiment: This movie was amazing!", Output = "positive".
- Unified Architecture: T5 utilizes the Transformer architecture, a deep learning model based on self-attention mechanisms. It employs an encoder-decoder design:
  - The encoder processes the input text to generate contextual representations.
  - The decoder generates the output text token by token.
- Pre-Training with Transfer Learning: T5 is pre-trained on a large corpus using a task called span corruption, where spans of text in the input are masked, and the model is trained to predict the masked spans. This approach helps the model learn a rich understanding of language, which can then be fine-tuned for specific downstream tasks.
- Scalability: T5 comes in various sizes, from small to large (e.g. T5-Small, T5-base, T5-Large, T5-XXL), allowing flexibility depending on computational resources and task complexity.
- Fine-Tuning: After pre-training, T5 can be fine-tuned on specific tasks with labeled data. This step ensures that the model learns task-specific nuances while leveraging the general language knowledge from pre-training.

The versatility of T5 makes it particularly suitable for tasks like summarization and machine translation, as it treats both tasks uniformly in the text-to-text format. This study leverages T5's capabilities in a multi-stage process: first fine-tuning it on machine translation tasks to leverage its language understanding capabilities, followed by additional fine-tuning on summarization tasks to optimize its performance for generating concise and coherent summaries. The model's capacity to generalize across tasks facilitates zero-shot testing, permitting assessment without task-specific fine-tuning.

B. Datasets

1) Summarization task: The dataset employed in this study is IndoSum [20]. IndoSum establishes a novel standard for text summarizing in Bahasa Indonesia. The IndoSum dataset comprises articles obtained from Indonesian online news outlets and contains almost 200 times the number of articles compared to prior research on Indonesian-language articles [39]. To be more precise, IndoSum includes 18,762 articles and their respective summaries, which are divided into six categories: entertainment, inspiration, sports, celebrity, headlines, and technology.

A number of well-known online news portals, including CNN Indonesia, Kumparan, Suara.com, Antaranews, and others, provided the data for IndoSum. The mean article length in this dataset is 292 words, with the biggest article being 1,228 words and the shortest item consisting of 36 words. The summaries average 58 words in length, with the greatest being 86 words and the smallest including 32 words.

2) Machine translation task: Datasets with parallel English-Indonesian sentences were utilized for pre-training, including OpenSubtitle [40], TED 2018 [41], TED 2020 [42], News Commentary [43], and CCMatrix [44]. These datasets were chosen especially for the purpose of training the T5 model for translating from English to Indonesian because of their complimentary qualities and applicability to the field of text translation.

The OpenSubtitle dataset, comprising a compilation of movie and television subtitles, offers a valuable repository of informal, conversational language. This dataset is very beneficial for training models to process daily language, colloquialisms, and conversational writing. The extensive range of themes and contexts guarantees that the model is exposed to multiple linguistic styles and structures.

The TED 2018 and TED 2020 datasets are composed of transcripts from TED Talks, which are renowned for their formal and informative communication style. Ideally, these datasets are utilized to train models that will translate academic, technical, and professional content. Additionally, they assist the model in dealing with a variety of subjects, as TED Talks encompass a broad spectrum of fields, including technology, science, and the arts and culture.

The News Commentary dataset emphasizes news stories and commentary, with a formal tone and terminology typically seen in journalistic literature. This dataset enables the model to manage domain-specific language, organized arguments, and subjective content characteristic of news sources.

Finally, the CCMatrix dataset constitutes a substantial compilation of parallel sentences derived from web-crawled data. The extensive size and varied content provide thorough coverage of linguistic structures and terminology, rendering it essential for enhancing the model's generalization capacity across distinct text kinds. For the CCMatrix dataset, we utilized just 600,000 data points due to constraints of computational resources and processing duration.

We sought to establish a strong and adaptable training basis for the T5 model by integrating these datasets. This selection guarantees the model's exposure to a wide array of text styles, tones, and domains, facilitating its effective performance in multiple translation scenarios, encompassing informal, formal, and technical contexts. The meticulous selection of datasets demonstrates our aim to develop a translation model that can tackle the intricacies and diversity present in actual English-to-Indonesian translation projects.

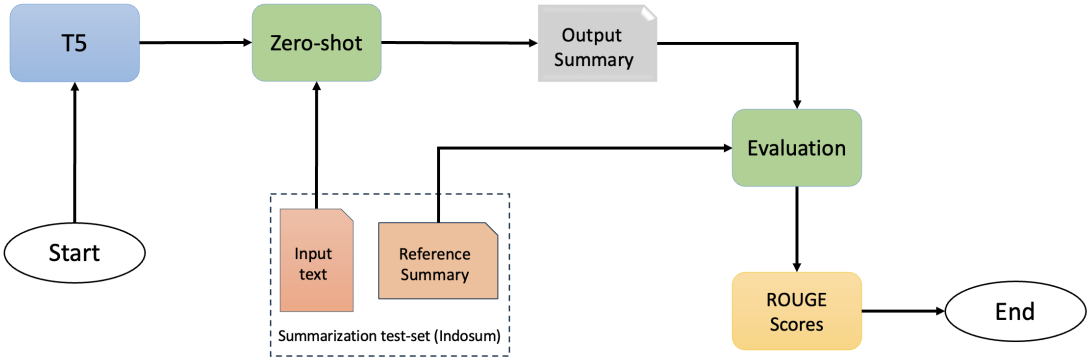


Fig. 1. T5 (Zero-shot).

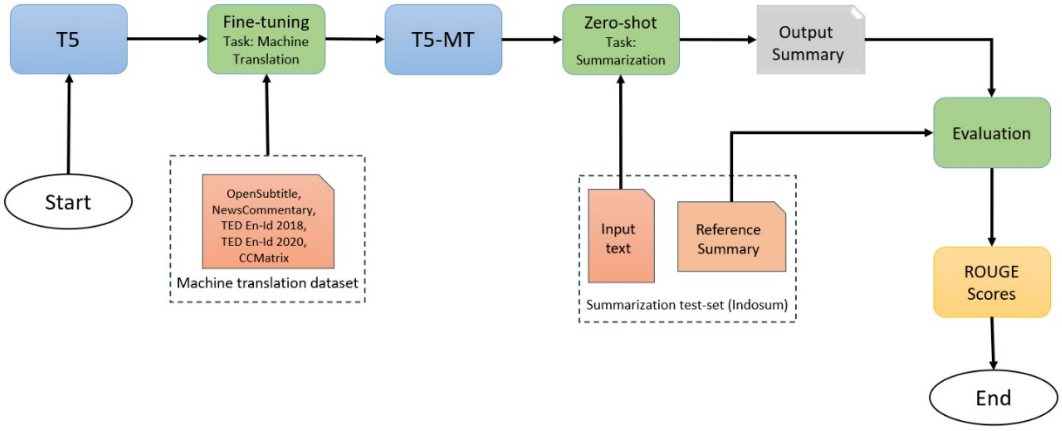


Fig. 2. T5-MT (Fine-tuning).

C. Experimental Design

A two-step fine-tuning approach is implemented in this study to assess the efficacy of the T5 model on Indonesian abstractive summarization. Four experimental configurations are included in the design, which has two tasks: machine translation and summarization.

1) Experiment 1: T5 Zero-shot: Fig. 1 illustrates the workflow for a zero-shot summarization task that employs the T5 (Text-to-Text Transfer Transformer) model. Starting with the pre-trained T5 model, which has not been refined on the specific summarization dataset, the procedure commences. The zero-shot approach involves explicitly testing the model on summarization tasks without any additional training. This demonstrates the T5 model's ability to generalize to unseen tasks by utilizing its pre-trained knowledge.

The summarization task employs the IndoSum dataset, comprising two primary elements: input text (the texts designated for summarization) and reference summaries (the authoritative summaries utilized for assessment). The T5 model analyzes the input text and produces a summary, which is subsequently compared to the reference summary to assess its quality.

The assessment procedure employs the ROUGE metric, which quantifies the overlap between the generated summaries and the reference summaries based on unigrams (ROUGE-1), bigrams (ROUGE-2), and the longest common subsequences (ROUGE-L). The resultant ROUGE scores quantify the T5 model's performance on the summarization job. The workflow concludes with the reporting of the ROUGE scores as the final result. This picture clearly delineates the process from job start to evaluation, highlighting the zero-shot capabilities of T5 in summarization tasks.

2) Experiment 2: T5-MT Fine-tuning: The T5 model is employed to perform two sequential tasks: fine-tuning for machine translation and zero-shot summarization. The workflow is illustrated in Fig. 2. It initiates with the T5-base model, which is refined on a machine translation task. This fine-tuning process uses diverse datasets such as OpenSubtitles, NewsCommentary, TED En-Id 2018, TED En-Id 2020, and CCMatrix, which consist of parallel text pairs for English-to-Indonesian translation. The outcome of this step is a specialized T5 model for machine translation, labeled as T5-MT.

The fine-tuned T5-MT model is then tested on a summarization task using the IndoSum dataset in a zero-shot manner, meaning the model is applied to summarization without additional task-specific fine-tuning. The summarization dataset includes input text (the articles to be summarized) and cor-

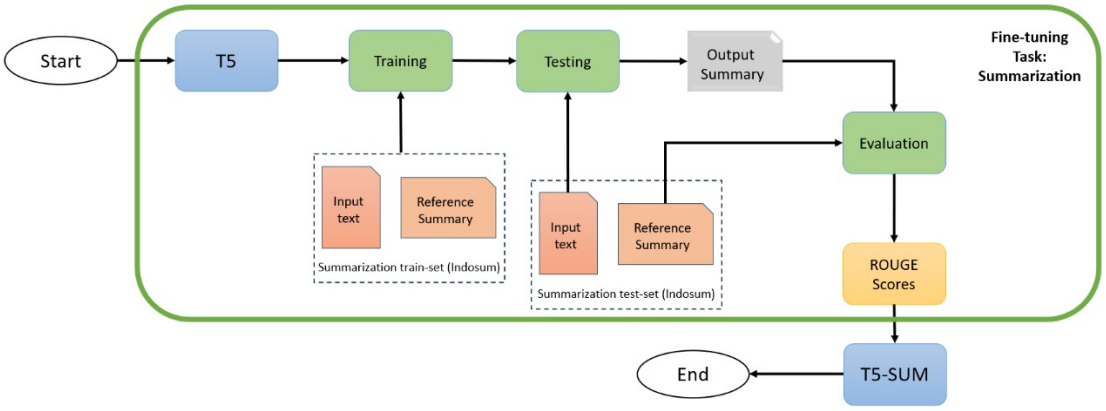


Fig. 3. T5-SUM (Fine-tuning).

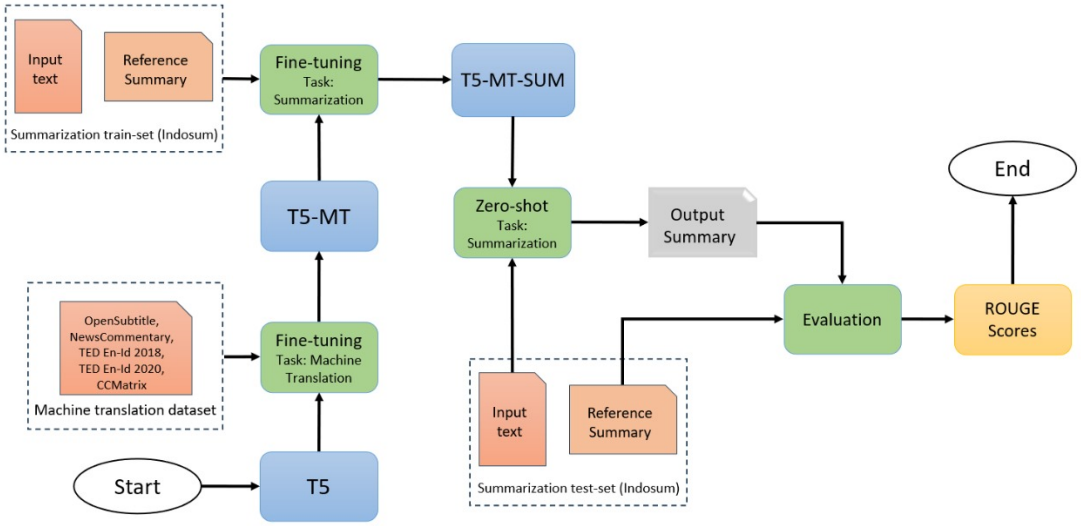


Fig. 4. T5-MT-SUM (Two-step Fine-tuning).

responding reference summaries (gold-standard summaries). The model processes the input text and generates an output summary.

The generated summaries are evaluated using the ROUGE metric, which computes the similarity between the output summaries and the reference summaries. The evaluation provides scores for ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), and ROUGE-L (longest common subsequence). These scores serve as a measure of the summarization performance. The workflow concludes with the ROUGE scores summarizing the model's effectiveness. This diagram showcases a pipeline approach, where a model fine-tuned for one task (machine translation) is tested for generalization in another task (summarization), illustrating the versatility of T5.

3) Experiment 3: T5-SUM Fine-tuning: Fig. 3 diagram represents the workflow for fine-tuning the T5 model specifically for a summarization task. It begins with the pre-trained T5 model, which undergoes fine-tuning using the IndoSum dataset. The dataset contains input text (articles or passages to be summarized) and their corresponding reference summaries (gold-standard summaries). The fine-tuning process adapts the T5 model to perform summarization tasks effectively, creating a specialized model referred to as T5-SUM.

The process is divided into two main phases: training and testing. During the training phase, the model learns patterns and relationships between the input text and the reference summaries from the training set of the IndoSum dataset. Once trained, the model is tested using the test set of the same dataset, where it generates summaries (labeled as Output Summary) for new, unseen input texts.

The ROUGE metric is employed to assess the summaries that are generated, comparing the overlap between the summaries and the reference summaries. The model is evaluated using ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), and ROUGE-L (longest common subsequences). The process culminates in the ROUGE scores, which serve as an indicator of the summarization task's performance of the fine-tuned T5-SUM model. This diagram effectively emphasizes

the significance of dataset preparation and evaluation metrics by illustrating a structured pipeline for training and evaluating a summarization model.

4) Experiment 4: T5-MT-SUM Two-Step Fine-tuning: Fig. 4 shows a two-step fine-tuning approach for modifying the T5 model to perform summarization tasks. It starts with the pre-trained T5 model, which is then fine-tuned on a machine translation task with datasets including OpenSubtitles, NewsCommentary, TED En-Id 2018, TED En-Id 2020, and CCMatrix. This step results in the T5-MT model, which is designed for machine translation.

During the second phase, the T5-MT model is subjected to an additional fine-tuning process for the summarizing job utilizing the IndoSum dataset, comprising input text and their respective reference summaries. Through further refinement, the T5-MT model is modified for summarizing tasks, resulting in the final model known as T5-MT-SUM. This two-step fine-tuning ensures that the model takes advantage of its machine translation knowledge before being improved for summarization, potentially enhancing performance.

The T5-MT-SUM model is subsequently evaluated on the IndoSum test set, producing output summaries for novel input texts in a zero-shot summarizing task. The produced summaries are assessed against the reference summaries utilizing the ROUGE metric, which measures summary quality through unigram, bigram, and longest common subsequence overlaps (ROUGE-1, ROUGE-2, and ROUGE-L, respectively). The procedure culminates with the ROUGE scores, which encapsulate the model's efficacy in the summarization task. This figure clearly illustrates the sequential fine-tuning method, demonstrating how knowledge transfer between tasks can improve performance.

D. Evaluation Metrics

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is a metric that is intended to assess the quality of summaries by corresponding them to one or more human-generated reference summaries. It measures the overlap between the candidate summary and reference summaries in terms of n-grams, sequences, or word pairs.

- ROUGE-1: Measures unigram overlap.
- ROUGE-2: Measures bigram overlap.
- ROUGE-L: Measures the longest common subsequence between generated and reference summaries.

ROUGE-N evaluates the n-gram overlap between a candidate summary and reference summaries. It is defined as:

$$\text{ROUGE-N} = \frac{\sum_{S \in \text{Reference Summaries}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \text{Reference Summaries}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)} \tag{1}$$

Where:

- $\text{Count}_{\text{match}}(\text{gram}_n)$: The number of n-grams in both the candidate summary and the reference summary.
- $\text{Count}(\text{gram}_n)$: The total number of n-grams in the reference summary.
- n: The length of the n-gram (e.g. 1 for unigrams, 2 for bigrams).

1) ROUGE-1 (Unigram overlap): ROUGE-1 measures the overlap of unigrams between the reference and candidate summaries. It is defined as:

$$\text{ROUGE-1} = \frac{\text{Number of Overlapping Unigrams}}{\text{Total Number of Unigrams in the Reference Summary}} \tag{2}$$

Example:

- Reference Summary: "I really like reading novels".
- Candidate Summary: "I like reading novels every day".
- Unigrams in Reference: I, really, like, reading, novels.
- Unigrams in Candidate: I, like, reading, novels, every, day.
- Overlapping Unigrams: I, like, reading, novels (4 overlaps).

The total number of unigrams in the reference summary is 5. Thus:

$$\text{ROUGE-1} = \frac{4}{5} = 0.8 \tag{3}$$

2) ROUGE-2 (Bigram Overlap): ROUGE-2 measures the overlap of bigrams (pairs of consecutive words) between the reference and candidate summaries. It is defined as:

$$\text{ROUGE-2} = \frac{\text{Number of Overlapping Bigrams}}{\text{Total Number of Bigrams in the Reference Summary}} \tag{4}$$

Example:

- Reference Bigrams: I really, really like, like reading, reading novels.
- Candidate Bigrams: I like, like reading, reading novels, novels every, every day.
- Overlapping Bigrams: like reading, reading novels (2 overlaps).

The total number of bigrams in the reference summary is 4. Thus:

$$\text{ROUGE-2} = \frac{2}{4} = 0.5 \tag{5}$$

3) ROUGE-L: ROUGE-L measures the longest common subsequence (LCS) between a candidate summary and a reference summary, capturing sentence-level structure. The precision, recall, and F-measure based on LCS are defined as:

$$P_{\text{LCS}} = \frac{\text{LCS}(X, Y)}{|Y|} \tag{6}$$

$$R_{\text{LCS}} = \frac{\text{LCS}(X, Y)}{|X|} \tag{7}$$

$$F_{\text{LCS}} = \frac{(1 + \beta^2) \cdot P_{\text{LCS}} \cdot R_{\text{LCS}}}{\beta^2 \cdot P_{\text{LCS}} + R_{\text{LCS}}} \tag{8}$$

Where:

- $\text{LCS}(X, Y)$: The length of the longest common subsequence between the reference summary X and the candidate summary Y.
- $|X|$: The length of the reference summary.
- $|Y|$: The length of the candidate summary.
- β: A weighting parameter, often set to 1 to equally weight precision and recall.

Example:

- Reference Summary: "I really like reading novels".
- Candidate Summary: "I like reading novels every day".
- LCS: "I like reading novels" (4 words).
- Length of Candidate: 6, Length of Reference: 5.

Calculate precision and recall:

$$P_{\text{LCS}} = \frac{4}{6} = 0.6667, \quad R_{\text{LCS}} = \frac{4}{5} = 0.8 \tag{9}$$

Calculate F-measure with $\beta = 1$:

$$F_{\text{LCS}} = \frac{(1 + 1^2) \cdot 0.6667 \cdot 0.8}{1^2 \cdot 0.6667 + 0.8} = 0.7273 \tag{10}$$

V. Results

The ablation study was conducted to evaluate the impact of fine-tuning at different stages on the Transformer-T5 model's performance in Indonesian abstractive summarization. Four configurations of the model were tested, as illustrated in Fig. 5, with the evaluation focusing on ROUGE-1, ROUGE-2, and ROUGE-L scores.

1) T5-base (Zero-shot): ROUGE-1: 0.4458, ROUGE-2: 0.3381, ROUGE-L: 0.3961. This configuration represents the baseline performance of the pre-trained T5 model without task-specific fine-tuning. The scores reflect the limited effectiveness of the zero-shot approach for a low-resource language like Indonesian, where the model has no prior exposure to Indonesian summarization or translation tasks.

2) T5-MT (Machine Translation fine-tuning): ROUGE-1: 0.6224, ROUGE-2: 0.5220, ROUGE-L: 0.5793. This model was fine-tuned on a machine translation task using the CC-Matrix dataset (300K parallel sentences). The improvement over the zero-shot baseline demonstrates the transferability of knowledge learned in machine translation to abstractive summarization.

3) T5-SUM (Summarization fine-tuning): ROUGE-1: 0.7106, ROUGE-2: 0.6393, ROUGE-L: 0.6790. Fine-tuning directly on the IndoSum dataset for summarization significantly enhanced performance compared to T5-MT. This result indicates that task-specific training is critical for improving the model's summarization capabilities.

4) T5-MT-SUM (Two-Step fine-tuning): In this configuration, the model was first fine-tuned on the machine translation task (OpenSubtitle, TED 2018, TED 2020, NewsCommentary, CCMatrix 600K) and subsequently fine-tuned on the summarization task (IndoSum). This two-step approach achieved the best performance with ROUGE-1: 0.7126, ROUGE-2: 0.6416, and ROUGE-L: 0.6816, underscoring the benefits of sequential task transferability.

VI. Discussion

The ablation study was designed to investigate how different fine-tuning strategies affect the T5 model's performance on Indonesian abstractive summarization. By systematically varying the stages of fine-tuning, the study aimed to uncover the relative contributions of machine translation and task-specific summarization fine-tuning to the final performance.

A. Ablation Study

1) Importance of related tasks: Fine-tuning on machine translation (T5-MT) significantly boosted performance compared to the zero-shot baseline. The ROUGE-1 improvement from 0.4458 to 0.6224 suggests that the knowledge gained in machine translation helps the model understand Indonesian syntax and semantics better.

2) Critical role of task-specific training: Fine-tuning directly on the summarization task (T5-SUM) led to further substantial gains, with ROUGE-1 reaching 0.7106. This highlights that task-specific training is essential for generating coherent and meaningful summaries.

3) Advantages of two-step fine-tuning: The incremental improvement of T5-MT-SUM over T5-SUM (ROUGE-1: 0.7126 vs. 0.7106) demonstrates the synergy of combining task-specific fine-tuning with knowledge from a related task like machine translation. While the improvement is small, it underscores the potential of two-step fine-tuning as a systematic approach to boosting model performance.

4) Implications for low-resource summarization: The findings from this ablation study provide valuable insights for Summarization research in low-resource languages:

- Leveraging Related Tasks: Transfer learning from machine translation to summarization is an effective strategy to overcome data scarcity in low-resource settings.
- Sequential Fine-Tuning: Two-step fine-tuning enhances the adaptability of pre-trained models by progressively refining their knowledge through related tasks.

B. Comparison with Existing Summarization Models

The current paper introduces a two-step transfer learning approach using Transformer-T5, where the model is first fine-tuned on a machine translation task before being fine-tuned for abstractive summarization on the IndoSum dataset. This structured pre-training strategy demonstrates significant improvements over previous works as shown in Table I. In

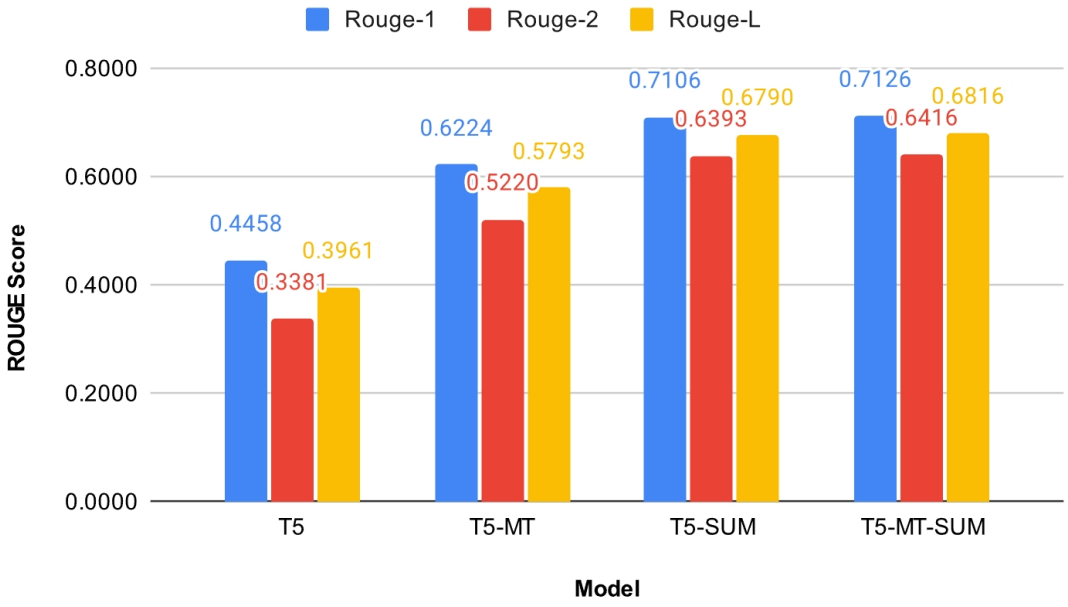


Fig. 5. Experiment result for two-step fine-tuned abstractive summarization.

terms of performance, the proposed model achieves ROUGE-1: 0.7126, ROUGE-2: 0.6416, and ROUGE-3: 0.6816, making it the highest-performing model across all three metrics. It surpasses the previous best result from BERTSum-Abs + IndoBERT-LEM + GPT Decoder (ROUGE-1: 0.6920, ROUGE-2: 0.6135, ROUGE-3: 0.6836), as well as BERT2GPT (ROUGE-1: 0.6226, ROUGE-2: 0.5613, ROUGE-3: 0.6043), demonstrating the effectiveness of combining translation-based pre-training with summarization fine-tuning.

Compared to data augmentation-based models, such as BART (Augmented) and mT5 (Augmented), which improved summarization performance but remained below ROUGE-1: 0.41, the proposed two-step approach proves to be more effective. Instead of solely relying on augmented training data, this study shows that a structured pre-training approach enhances the model's ability to generate coherent and fluent summaries. Additionally, the study outperforms multilingual and cross-lingual models, such as CTRLSum and Transformer (CLS + MWE), which achieved lower ROUGE scores despite their ability to generate summaries in multiple languages. The findings suggest that a monolingual, targeted approach with pre-training on a related NLP task (machine translation) yields better results than cross-lingual training.

Overall, this research sets a new benchmark for Indonesian abstractive summarization, showing that fine-tuning on machine translation before summarization enhances summary coherence and quality. The structured two-step learning strategy proves more effective than direct fine-tuning, data augmentation, and hybrid extractive-abstractive methods, establishing Transformer-T5 (two-step transfer learning) as the best-performing model for Indonesian text summarization compared to previous work.

C. Challenges and Limitations

Despite the promising results, the study faced several challenges:

- Marginal gains: The improvement from T5-SUM to T5-MT-SUM, while consistent, was relatively small. This suggests diminishing returns with additional fine-tuning stages and calls for further exploration of factors such as dataset size and quality.
- Evaluation metrics: ROUGE scores, while widely used, have limitations in capturing summary coherence and informativeness. Incorporating human evaluations could provide a more comprehensive assessment of summary quality.

D. Future Directions

This study paves the way for several promising research directions. First, future work could explore optimizing the two-step fine-tuning process by refining dataset selection or introducing pretraining techniques tailored to low-resource languages. Incorporating larger or more diverse datasets could further improve the model's ability to generalize across different summarization scenarios.

Second, qualitative evaluation of generated summaries through human assessments should be conducted to complement the ROUGE metric. This would provide deeper insights into summary coherence, readability, and informativeness, which are crucial for practical applications.

Third, expanding the approach to multilingual fine-tuning could assess its generalizability across other low-resource languages. The combination of cross-lingual transfer learning and domain-specific training could offer a robust framework for NLP applications in diverse linguistic contexts.

Finally, future research could explore adapting the two-step fine-tuning methodology for other NLP tasks, such as question answering, sentiment analysis, or content generation. These extensions would highlight the broader applicability of this approach and contribute to advancing natural language processing in low-resource settings.

VII. Conclusions

This research explored the effectiveness of a two-step fine-tuning approach using the Transformer-T5 model for abstractive text summarization in Bahasa Indonesia, a low-resource language. Four configurations were evaluated: zero-shot (T5-Base), single-task fine-tuning for machine translation (T5-MT), single-task fine-tuning for summarization (T5-SUM), and two-step fine-tuning combining machine translation and summarization tasks (T5-MT-SUM). The two-step fine-tuning approach achieved the best performance, with ROUGE-1: 0.7126, ROUGE-2: 0.6416, and ROUGE-L: 0.6816, demonstrating the benefits of task transferability in improving summarization quality.

The findings indicate that leveraging knowledge from a related task, such as machine translation, can enhance the performance of abstractive summarization models. Task-specific fine-tuning was also shown to play a critical role in generating coherent and accurate summaries. The results validate the potential of transfer learning to overcome the challenges posed by data scarcity in low-resource languages like Bahasa Indonesia.

Acknowledgment

This research is supported by Universitas Gadjah Mada.

References

[1] A. Mitchell, J. H. Holcomb, and R. Weisel, "State of the News Media 2016," *WD info*, p. 118, 2016. [Online]. Available: http://www.journalism.org/2016/06/15/state-of-the-news-media-2016/#

[2] S. Primativo, D. Spinelli, P. Zoccolotti, M. De Luca, and M. Martelli, "Perceptual and Cognitive Factors Imposing "Speed Limits" on reading rate: A study with the rapid serial visual presentation," *PLoS ONE*, vol. 11, no. 4, pp. 1–25, 2016.

[3] O. Klymenko, D. Braun, and F. Matthes, "Automatic text summarization: A state-of-the-art review," *ICEIS 2020 - Proceedings of the 22nd International Conference on Enterprise Information Systems*, vol. 1, no. Iceis, pp. 648–655, 2020.

[4] G. Sharma and D. Sharma, "Automatic Text Summarization Methods: A Comprehensive Review," *SN Computer Science*, vol. 4, no. 1, pp. 1–18, 2023. [Online]. Available: https://doi.org/10.1007/s42979-022-01446-w

[5] Y. J. Kumar, O. S. Goh, H. Basiron, N. H. Choon, and P. C. Suppiah, "A review on automatic text summarization approaches," pp. 178–190, apr 2016. [Online]. Available: https://thescipub.com/abstract/jcssp.2016.178.190

[6] A. Khan, N. Salim, and H. Farman, "Clustered genetic semantic graph approach for multi-document abstractive summarization," *2016 International Conference on Intelligent Systems Engineering, ICISE 2016*, vol. 77, no. 18, pp. 63–70, 2016.

[7] V. Severina and M. L. Khodra, "Multidocument Abstractive Summarization using Abstract Meaning Representation for Indonesian Language," in *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*. IEEE, sep 2019, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8904449/

[8] I. R. Musyaffanto, G. Budi Herwanto, and M. Riasetiawan, "Automatic extractive text summarization for indonesian news articles using maximal marginal relevance and non-negative matrix factorization," *Proceedings - 2019 5th International Conference on Science and Technology, ICST 2019*, pp. 1–6, 2019.

[9] D. Annisa and M. L. Khodra, "Query-based summarization for Indonesian news articles," *Proceedings - 2017 International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2017*, 2017.

[10] R. Reztaputra and M. L. Khodra, "Sentence structure-based summarization for Indonesian news articles," *Proceedings - 2017 International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2017*, pp. 0–5, 2017.

[11] R. B. Hutama, A. R. Barakbah, and A. Helen, "Indonesian news auto summarization in infrastructure development topic using 5W+1H consideration," *Proceedings - International Electronics Symposium on Knowledge Creation and Intelligent Computing, IES-KCIC 2017*, vol. 2017-Janua, pp. 258–264, 2017.

[12] N. F. Saraswati, Indriati, and R. S. Perdana, "Peringkasan Teks Otomatis Menggunakan Metode Maximum Marginal Relevance Pada Hasil Pencarian Sistem Temu Kembali Informasi Untuk Artikel Berbahasa Indonesia," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, vol. 2, no. 11, pp. 5494–5502, 2018.

[13] A. Indriani, "Maximum Marginal Relevance Untuk Peringkasan Teks Otomatis Sinopsis Buku Berbahasa Indonesia," *SEMNASTEKNOMEDIA ONLINE*, vol. 2, no. 1, pp. 3–5, 2014.

[14] J. Carbinell and J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries," *ACM SIGIR Forum*, vol. 51, no. 2, pp. 209–210, aug 2017. [Online]. Available: https://doi.org/10.1145/3130348.3130369https://dl.acm.org/doi/10.1145/3130348.3130369

[15] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411.

[16] M. T. Yulyanto and M. L. Khodra, "Automatic extractive summarization on Indonesian parliamentary meeting minutes," *Proceedings - 2017 International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2017*, 2017.

[17] G. H. Rachman and M. L. Khodra, "Automatic rhetorical sentence categorization on Indonesian meeting minutes," in *Proceedings of 2016 International Conference on Data and Software Engineering, ICoDSE 2016*. IEEE, oct 2017, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/7936103/

[18] C. D. F. M. Paul Lewis, Gary F. Simons, *Ethnologue: Languages of the World*, 18th ed. Dallas, TX: SIL International, 2015.

[19] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar, and A. Purwarianti, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," pp. 843–857, sep 2020. [Online]. Available: http://arxiv.org/abs/2009.05387

[20] K. Kurniawan and S. Louvan, "IndoSum: A New Benchmark Dataset for Indonesian Text Summarization," in *Proceedings of the 2018 International Conference on Asian Language Processing, IALP 2018*. IEEE, nov 2019, pp. 215–220. [Online]. Available: https://ieeexplore.ieee.org/document/8629109/

[21] F. Koto, J. H. Lau, and T. Baldwin, "Liputan6: A Large-scale Indonesian Dataset for Text Summarization," no. 1, pp. 598–608, nov 2020. [Online]. Available: http://arxiv.org/abs/2011.00679https://aclanthology.org/2020.aacl-main.60/

[22] R. Adelia, S. Suyanto, and U. N. Wisesty, "Indonesian Abstractive Text Summarization Using Bidirectional Gated Recurrent Unit," *Procedia Computer Science*, vol. 157, pp. 581–588, 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877050919311755

[23] R. S. Devianti and M. L. Khodra, "Abstractive Summarization using Genetic Semantic Graph for Indonesian News Articles," *Proceedings - 2019 International Conference on Advanced Informatics: Concepts, Theory, and Applications, ICAICTA 2019*, 2019.

[24] Y. H. B. Gunawan and M. L. Khodra, "Multi-document Summarization using Semantic Role Labeling and Semantic Graph for Indonesian News Article," *2020 7th International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2020*, 2020.

[25] R. Wijayanti, M. L. Khodra, and D. H. Widyantoro, "Single Document Summarization Using BertSum and Pointer Generator Network," *International Journal on Electrical Engineering and Informatics*, vol. 13, no. 4, pp. 916–930, dec 2021. [Online]. Available: http://ijeei.org/docs-204000029161f4c11fc91b1.pdf

[26] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," pp. 25–26, 2004. [Online]. Available: papers2://publication/uuid/5DDA0BB8-E59F-44C1-88E6-2AD316DAEF85

[27] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, oct 2020. [Online]. Available: http://arxiv.org/abs/1910.10683https://www.jmlr.org/beta/papers/v21/20-074.html

[28] M. Aurelia, S. Monica, and A. S. Girsang, "Transformer-based abstractive indonesian text summarization," *International Journal of Informatics and Communication Technology*, vol. 13, no. 3, pp. 388–399, 2024.

[29] R. A. Rahman and Suyanto, "Performance Analysis of ChatGPT for Indonesian Abstractive Text Summarization," *Proceedings - International Seminar on Intelligent Technology and its Applications, ISITIA*, no. 2024, pp. 477–482, 2024.

[30] A. L. Putra, Sanjaya, M. R. Fachruradzi, and A. Y. Zakiyyah, "Resource Efficient Abstractive Text Summarization in Indonesian with ALBERT," *2024 International Conference on Smart Computing, IoT and Machine Learning, SIML 2024*, pp. 81–85, 2024.

[31] Z. P. Ayudhia and S. Suyanto, "Deeper Investigation on Extractive-Abstractive Summarization for Indonesian Text," *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems, ICETSIS 2024*, no. 2022, pp. 1704–1708, 2024.

[32] M. D. Ferdiansyah and S. Suyanto, "Data Augmentation and Fine-Tuning to Improve IndoBART Performance," *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems, ICETSIS 2024*, pp. 1668–1672, 2024.

[33] A. S. Wijaya and A. S. Girsang, "Augmented-Based Indonesian Abstractive Text Summarization using Pre-Trained Model mT5," *International Journal of Engineering Trends and Technology*, vol. 71, no. 11, pp. 190–200, 2023.

[34] M. Nasari, A. Maulina, and A. S. Girsang, "Abstractive Indonesian News Summarization Using BERT2GPT," *Proceedings - 2023 IEEE 7th International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2023*, pp. 369–375, 2023.

[35] D. Puspitaningrum, "A Survey of Recent Abstract Summarization Techniques," 2022, pp. 783–801. [Online]. Available: https://link.springer.com/10.1007/978-981-16-2102-4$_$71

[36] H. Lucky and D. Suhartono, "Investigation Of Pre-Trained Bidirectional Encoder Representations From Transformers Checkpoints For Indonesian Abstractive Text Summarization," *Journal of Information and Communication Technology*, vol. 21, no. 1, pp. 71–94, nov 2022. [Online]. Available: https://e-journal.uum.edu.my/index.php/jict/article/view/13548

[37] R. E. Prasojo and A. A. Krisnadhi, "Controllable Abstractive Summarization Using Multilingual Pretrained Language Model," pp. 228–233, 2022.

[38] A. F. Abka, K. Azizah, and W. Jatmiko, "Transformer-based Cross-Lingual Summarization using Multilingual Word Embeddings for English - Bahasa Indonesia," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 12, pp. 636–645, 2022.

[39] A. Najibullah, "Indonesian Text Summarization based on Naïve Bayes Method," *International Seminar and Conference 2015 : The Golden Triangle (Indonesia-India-Tiongkok)*, pp. 67–78, 2015.

[40] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles," in *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 2016, pp. 923–929.

[41] Y. Qi, D. S. Sachan, M. Felix, S. J. Padmanabhan, and G. Neubig, "When and why are pre-trainedword embeddings useful for neural machine translation?" *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 2, pp. 529–535, 2018.

[42] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 4512–4525, 2020.

[43] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, pp. 2214–2218, 2012.

[44] H. Schwenk, G. Wenzek, S. Edunov, E. Grave, A. Joulin, and A. Fan, "CCMatrix: Mining billions of high-quality parallel sentences on the web," in *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2021, pp. 6490–6500.

AI-Driven Construction and Application of Gardens: Optimizing Design and Sustainability with Machine Learning

Jingyi Wang[1]*, Yan Song[2], Haozhong Yang[3], Han Li[4], Minglan Zhou[5]
School of Architecture, Xi'an University of Architecture & Technology, Xi'an 710055[1,3]
Survey Institute of Shaanxi Land Engineering Construction Group, Xi'an 710065, China[2]
CSCES AECOM CONSULTANTS CO., LTD, Lanzhou 730000[4]
School of Housing, Building and Planning, Universiti Sains Malaysia, Gelugor 11800, Penang, Malaysia[5]
PhD Candidates School of Art, Lanzhou University of Finance and Economics[5]

***Abstract*—Artificial intelligence (AI) integration into environmental analysis has revolutionized various fields. Including the construction and application of gardens, by enabling precise classification and decision-making for sustainable practices. This paper presents a strong AI-driven framework uses convolutional neural network (CNN) and pretrained models like VGG16 and InceptionV3 to classify eight distinct environmental classes. The CNN achieved superior performance Among the tested models and reaching an impressive 98% accuracy with optimized batch sizes. This demonstrate its effectiveness for precise environmental condition classification. This work highlights the crucial role of AI in advancing the construction and application of gardens. It offers insights into optimizing garden design through accurate environmental data analysis. The diverse dataset used ensures the framework's adaptability to real-world applications, making it a valuable resource for sustainable development and eco-friendly design strategies. This paper not only contributes to the field of AI-driven environmental analysis but also provides a foundation for future innovations in garden management and sustainability, paving the way for intelligent solutions in the evolving landscape of ecological design.**

***Keywords*—Artificial intelligence; machine learning; construction and application of garden design; convolutional neural network; VGG16; InceptionV3**

I. Introduction

The construction and application of gardens have long been intertwined with human civilization, serving as timeless symbols of beauty, solace, and ecological significance. From the glory of the Babylon Hanging Gardens, celebrated as one of the Seven Wonders of the Ancient World, to the serene and meticulously crafted Japanese Zen gardens [1]. These green spaces have continuously evolved to reflect cultural ideals, environmental adaptations, and technological advances [2]. More than just a testament to human creativity and harmony with nature, gardens also embody the intersection of sustainability and aesthetic appeal [3]. In today's rapidly urbanizing world, their role has expanded far beyond visual pleasure, as they now play a crucial part in addressing global challenges such as climate change, biodiversity loss, urban livability, and resource scarcity [4]. At the same time, they remain essential in promoting mental and physical well-being, offering spaces for relaxation, social interaction, and ecological balance. The integration of artificial intelligence has further revolutionized gardening, introducing data-driven solutions for optimizing garden layouts, automating maintenance, and enhancing biodiversity management. AI-powered systems now facilitate sustainable irrigation, real-time plant health monitoring, and predictive analysis for pest and disease control, making urban and rural green spaces more resilient and efficient. As technological advancements continue to shape the way gardens are designed and maintained, there is a growing need for ethical and sustainable approaches that balance innovation with ecological responsibility [5].

In this day and age of technological progress, the arrival of artificial intelligence (AI) and machine learning (ML) has opened evolutionary possibilities in the construction and application of gardens [6]. These cutting-edge technologies allow for the analysis of vast and complex datasets, enabling the optimization of garden design, the enhancement of sustainability, and the implementation of resource-efficient maintenance strategies. One good example of the potential of AI lies in land use scene classification, where ML models analyze satellite imagery as well as aerial photographs to classify various land types such as, *Forest*, *River*, agricultural areas etc. This capability is particularly instrumental in identifying suitable sites. This is useful for garden construction and to focus the design on specific ecological and climatic needs [7]. The application of these technologies marks the departure from traditional garden design practices, which often rely on manual analysis and experience methods. That can accidentally overlook crucial environmental and sustainability factors. The construction and application of gardens come with unique challenges. That AI is exceptionally well positioned to address. Traditional garden design methods, while rooted in artistic expression and intuition. Can be restricted by their ineffectiveness to integrate environmental and sustainability considerations widely [8].

By incorporating ML models into this process, designers can access data driven insights into land usability. And predict resource requirements and develop garden layouts that prioritize biodiversity and ecological harmony. For example, AI models can simulate various garden configurations. Taking into account variables such as sunlight exposure, water availability, soil quality, and plant compatibility ensures that

*Corresponding authors.

the resulting designs are functional and sustainable [9]. In addition, these AI-powered systems extend their utility beyond design by enabling post-construction monitoring. Providing real-time data to support versatile maintenance strategies that reduce resource consumption and environmental impact. These systems enable garden planners to address not only immediate design concerns, but also long-term sustainability objectives, ensuring gardens remain thriving ecosystems over time [10].

This research focuses on using land use scene classification, which revolutionizes the construction and application of gardens. It offers a comprehensive framework that integrates sustainability, innovative design principles, and cutting-edge technological advancements. Using the Land-Use Scene Classification dataset, this study examines specific land categories, such as *Agriculture*, *Forest*s, *River*s, and residential zones, which are directly relevant to the sustainable design of the garden. ML models trained in these categories aim to provide actionable insights into site selection, resource allocation, and design optimization. The proposed approach bridges the gap between traditional garden design methods and the powerful capabilities of modern AI technologies, demonstrating the potential of AI to transform green space planning into a data-driven, scalable, and ecologically sound endeavor. The insights generated by these models will not only enhance the visual and functional aspects of gardens but will also align them with broader global sustainability goals, such as carbon sequestration, biodiversity conservation, and efficient water use. The significance of this work lies in its ability to reimagine the construction and application of gardens in an era defined by rapid urbanization and environmental degradation. By introducing a scalable, adaptable and data-driven approach, this research demonstrates how AI can be used to create gardens that are aesthetically pleasing, ecologically sustainable, and socially impactful. Beyond their immediate applications, such gardens contribute to the broader vision of fostering greener, more resilient urban and rural landscapes. In addition, this paper underscores the vital role of AI in the resolution of global environmental challenges, highlighting the importance of innovation and technological solutions to foster a harmonious coexistence between human development and nature. The key contributions of this work are:

- Introduces an AI-powered framework for garden construction that integrates the classification of the land use scene, enabling sustainable and efficient garden design.
- Demonstrate how AI can improve resource efficiency, biodiversity, and environmental sustainability in garden construction.
- Conducts a detailed evaluation of traditional garden design methods versus AI-driven approaches to highlight efficiency, accuracy, and sustainability gains.

II. Literature Review

Artificial intelligence (AI) and machine learning (ML) in the construction and application of gardens has gained substantial attention in recent years [11]. Cities continue to enlarge the importance of sustainable, efficient and ecologically sound garden designs is becoming increasingly clear. AI and ML present an opportunity to optimize garden layouts, improve resource use, and improve sustainability [12]. These technologies enable designers to analyze and classify land use more precisely, creating gardens that are not only aesthetic but also environmentally beneficial. One of the key areas where AI has shown promise is in land use classification. The ability to accurately classify different types of land, such as agricultural zones and urban areas. *Forest*s, and water bodies can significantly influence the construction and application of gardens [13]. AI models can provide insight into the bionomics and environmental characteristics of the area, using satellite images and aerial imagery for the construction of gardens [14]. These techniques allow for a deeper understanding of soil quality and water availability. Also, understands climate for ensuring that gardens are designed with the local environment in mind.

In addition to land-use classification, AI-driven tools are being applied to urban planning and landscape architecture [15]. These tools enable the creation of generative designs in which multiple garden layouts are explored and tested for optimal performance. AI algorithms can evaluate different configurations, considering factors such as compatibility, sunlight exposure, and soil health [16]. This process ensures that gardens are not only functional but also sustainable. The design iterations produced by AI can adapt to environmental changes, making gardens more resilient to challenges such as climate change, loss of biodiversity, and water scarcity. Such innovations allow for the creation of green spaces that can thrive in a variety of conditions, from urban rooftops to expansive rural landscapes [17].

Another critical application of AI is in the field of precision *Agriculture*, which shares many principles with the construction of sustainable gardens [18]. By monitoring soil moisture levels, weather patterns, and water use, AI systems help optimize resource allocation in agriculture. These technologies have proven to be effective in reducing water waste and maximizing crop yield. Similarly, in the construction and application of gardens, AI can monitor and manage resources such as water, fertilizers, and energy, ensuring that gardens are not only beautiful, but also efficient and sustainable [19]. The application of such technologies allows for adaptive maintenance strategies that minimize environmental impact while keeping gardens thriving.

AI also plays a crucial role in enhancing biodiversity and supporting the ecological balance within gardens [20]. By simulating various environmental conditions and plant interactions, AI can suggest garden layouts that support a diverse range of species and foster healthier ecosystems. This is particularly important in urban areas, where biodiversity is often limited. AI models can identify plant species that are compatible with each other and the local environment, promoting plant diversity and reducing the need for chemical interventions [21].

Through this, gardens can become vital ecosystems that support local wildlife and contribute to the overall health of the urban environment. Although much of the current research has focused on individual aspects of garden design or broader environmental planning, the potential for AI to integrate land use classification with garden construction remains largely unexplored. Most of the existing efforts have focused on urban planning or agricultural optimization, leaving a gap in the

specific application of AI to the construction and application of gardens [22].

This presents an opportunity to bridge the gap and develop AI-driven tools that combine ecological sustainability with design optimization. By merging these two areas, AI can play a pivotal role in creating gardens that are not only functional and beautiful, but also environmentally resilient and resource-efficient [23].

The Motivation behind this research is to fill the gaps in existing research by specifically focusing on the intersection of AI-driven land use classification and the construction and application of gardens. although numerous researchers have explored human aspects of AI in urban planning, *Agriculture*, or ecological design, the application of these technologies in optimizing garden design and sustainability remains underexplored. This article aims to explain how AI can revolutionize the construction and application of gardens, ensuring that they are not only aesthetically pleasing but also ecologically viable and resource efficient. Using ML to integrate land-use classification with garden design, this research contributes to creating smarter, more sustainable green spaces that can adapt to the challenges posed by climate change, urbanization, and biodiversity loss. This paper seeks to push the boundaries of the construction and application of gardens, showing how AI can be harnessed to optimize both the process and the outcome of garden construction, thus fostering greener and more resilient landscapes.

III. Methodology

A. Dataset Collection

Data collection is the most important aspect of the research. Dataset utilized in this research is a curated from kaggle. Which is subset of a land-use scene classification dataset and is specifically designed to facilitate the construction and application of gardens.The dataset consists satellite images which represents eight diverse land-use scenarios such as, *Agriculture*, *Forest*, *River*, urban areas and more. Each class is selected for its critical role in sustainable garden design. This dataset captures a range of ecological and environmental scenarios which provides valuable insights into land characteristics. Which is essential for tailored garden planning which include agricultural suitability urban density as well as water management. The dataset underwent preprocessing to standardize image resolution and format to ensure consistency. To increase diversity and simulate real-world variations advanced augmentation techniques such as rotation, scaling, and flipping, were applied. This diverse dataset forms a strong foundation for integrating AI into the construction and application of gardens which enhances their sustainability, efficiency, and resilience.

This dataset was chosen based on its comprehensive coverage of diverse land-use types relevant to garden planning. It provides high-quality satellite imagery, ensuring reliable data for AI-driven analysis. Compared to other datasets, this one offers a well-balanced mix of natural and urban environments, making it particularly suitable for evaluating ecological and spatial factors in sustainable garden construction. This directly align with the objectives of this work, integrating AI into the construction and application of gardens. The dataset diverse classes provide a basis for training ML models which is capable of optimizing garden designs, accounting for ecological and spatial demands. Furthermore, the dataset detailed depiction of environmental scenarios makes it ideal for testing new approaches to sustainable and efficient garden construction. Enhancing the practicality and adaptability of the methodology. Detailed description of the dataset is shown in Table I, which highlights its significance in the advancement of AI-driven garden design solutions. Despite its strengths, the dataset has some limitations. The fixed satellite image resolution may impact fine-grained analysis of smaller garden structures. Additionally, while the dataset covers multiple land-use types, real-time environmental variations such as seasonal changes or soil conditions are not explicitly captured. Addressing these challenges in future work could involve integrating real-time remote sensing data or expanding the dataset to include dynamic environmental parameters.

B. Preprocessing

1) Data resizing: We observed significant variations in the resolution of images within the dataset during preprocessing. This variation could negatively impact the consistency and accuracy of the model. In order to address this issue all images were uniformly resized to 224 × 224 pixels. By doing this, we observed that this standardization ensures the consistency of the input. Facilitating efficient processing by the model as well as helping with reducing computational complexity. Furthermore, this resizing helps in maintaining the balance not only image quality but also performance. Which ensures optimal feature extraction during the training.

2) Normalization of data: Data normalization is a crucial step in data preprocessing to enhance the performance of ML models. Where we standardize or rescales the input to fall within a specific range between 0 and 1 or sometimes between -1 and 1. The data normalization process minimizes the impact of varying not only pixel_intensity values, reduces training time but helps the model focus to learn meaningful patterns rather than being influenced by scale variance in the dataset.

C. Data Distribution and Quantitative Analysis

A well-structured dataset is the cornerstone of any successful deep learning model. For this study, we carefully curated

TABLE I. Dataset Description for the Construction and Application of Gardens

Class Name	Description
Agriculture	Areas used for farming, including crop fields and orchards, suitable for plant-rich designs.
Beach	Coastal sandy areas, often requiring salt-resistant plants and erosion control measures.
Denseresidential	Urban areas with closely packed houses or apartments, ideal for rooftop or small-space gardens
Forest	Large areas covered by trees, offering inspiration for natural, eco-friendly garden designs.
Golfcourse	Open green spaces maintained for recreational purposes, with efficient irrigation systems.
Mediumresidential	Suburban areas with moderately spaced housing, suitable for private gardens or community spaces.
Parkinglot	Large paved areas used for vehicle parking, often with potential for integrating green spaces.
River	Natural flowing water bodies, influencing designs with water management and riparian vegetation.

a dataset with eight distinct land-use classes: *Agriculture*, *Beach*, *Denseresidential*, *Forest*, *Golfcourse*, *Mediumresidential*, *Parkinglot*, and *River*. Each class is represented by 500 images, ensuring an equal distribution that prevents class imbalance, which is often a major challenge in classification tasks. The dataset is further divided into training and testing sets. Resulting in 400 images for training as well as 100 images for testing. This distribution strategy provides sufficient data for model training while preserving a fair portion for unbiased evaluation as shown in Table II. The dataset is balanced with a total of 4,000 images, distributed as 3,200 training images and 800 testing images. This balanced lays the foundation for fair and consistent model learning.

TABLE II. DATASET OVERVIEW SHOWING THE NUMBER OF IMAGES, TRAINING SET, AND TEST SET DISTRIBUTION ACROSS CLASSES

Class Name	No. of Images	Training Set	Test Set
Agriculture	500	400	100
Beach	500	400	100
Denseresidential	500	400	100
Forest	500	400	100
Golfcourse	500	400	100
Mediumresidential	500	400	100
Parkinglot	500	400	100
River	500	400	100
Total	**4000**	**3200**	**800**

To visualize this distribution, Fig. 1, presents a pie chart that illustrates the equal percentage of images contributed by each class. With each class forming exactly 12.5% of the dataset, the dataset achieves perfect equilibrium, eliminating any inherent bias toward a specific category. This balance is critical to ensure the model generalizes well across all land-use categories and does not overfit to any particular class.

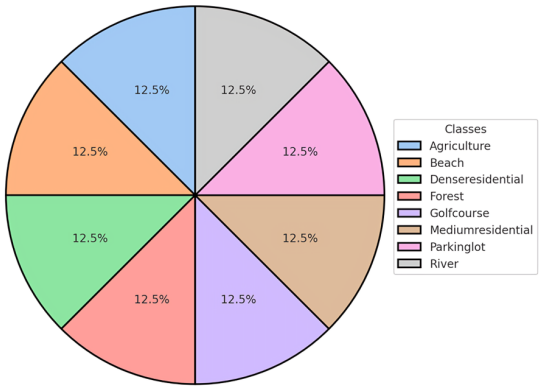


Fig. 1. Class distribution in the dataset, showing balanced contributions from each class.

To further validate this equal distribution, Fig. 2 shows a bar chart displaying the total number of images per class. The uniform height of the bars emphasizes that every class has exactly 500 images, underscoring the equal composition of the dataset. This visual confirmation strengthens confidence in the integrity of the dataset and its suitability for training a robust classification model.

In addition to the overall distribution, it is crucial to examine the segregation of the dataset into training and testing

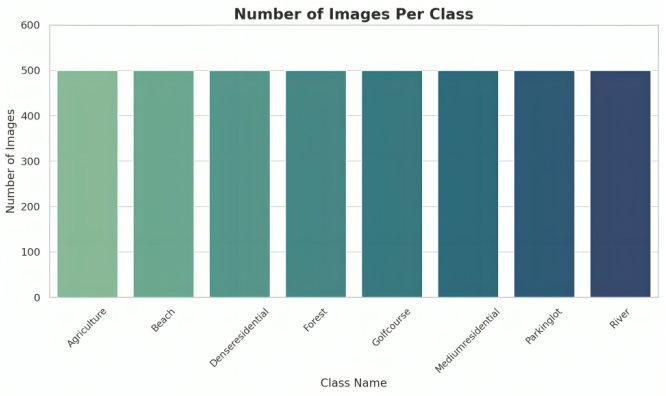


Fig. 2. Total number of images per class.

subsets. Fig. 3 provides a detailed visualization of the training and testing distribution for each class. Training set consists of 400 images for each class in the dataset and 100 images to the testing set which shows the 80% split for training and 20% split for test set. This distribution ensures that the model is trained on a substantial portion of the data while reserving enough samples for an objective evaluation of its performance. The design of this dataset ensures a harmonious blend of diversity and balance. The balance representation of all classes guarantees that the model receives varied inputs, preventing any single class from dominating the learning process. Moreover, the structured division into training and testing subsets aligns with best-practices in machine learning, facilitating reliable and unbiased performance assessment.

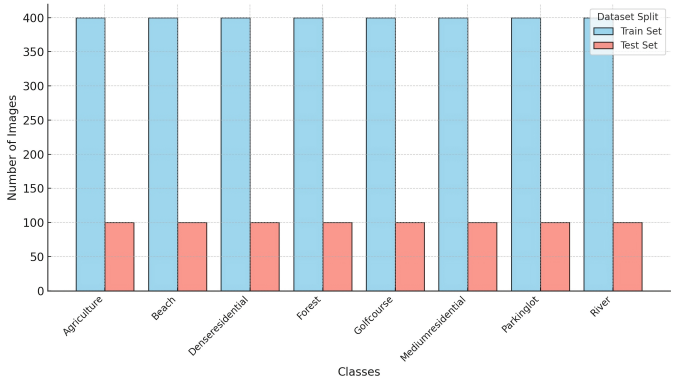


Fig. 3. Total number of images in train and test set.

By combining these quantitative analysis with visual insights, we can conclude that the dataset is well prepared to support the development of an effective classification model. This precise approach of distribution of data improves the reliability of the study and also emplace a solid foundation for future research on land use classification.

D. Proposed Framework

The proposed classification framework uses advanced machine learning to provide an effective and efficient solution. The training and testing workflow is presented in Fig. 4. The system begins by receiving image inputs from the dataset,

which are then divided into training, testing, and validation sets.

Furthermore, the proposed framework for this work is presented in Fig. 5, designed to optimize the construction and application of gardens using satellite image classification. It categorizes images into eight distinct classes Agriculture, Beach, Denseresidential, Forest, Golfcourse, Mediumresidential, Parkinglot, and River enabling precise and efficient landscape planning.

The process begins with preprocessing, where satellite images are resized and normalized to ensure consistency. These preprocessed images are then passed through a custom Convolutional Neural Network (CNN) that extracts key spatial features, such as textures, patterns, and vegetation density. Convolutional and max-pooling layers work together to identify and retain essential features while reducing data complexity. The extracted features are then mapped through fully connected layers, where the softmax activation function ensures accurate classification by assigning probabilities to each class. The framework not only enhances the classification of land instances. But also supports AI_driven decisions for garden construction and sustainability. For instance identifying agricultural regions or *River* landscapes allows for informed garden designs tailored to specific environmental contexts. By integrating both automation as well as precision the proposed framework transforms traditional garden planning into more effective and sustainable process.

IV. Results and Discussion

A. Experimental Setup

The experiments were conducted using intel(R) Core(TM) i5-6500 CPU running at 2.60 GHz, along with 16 GB, RAM on the windows 10 operating system. Anaconda-based python 3.11 environment configured with TensorFlow-and PyTorch.

B. Evaluation Metrics

The performance of the model is evaluated using various metrics such as, accuracy, precision, recall as well as F1-score. Which together offers a detailed understanding of the proposed model classification capabilities. Firstly ,accuracy calculated using Eq. 1, determines the overall correctness of the model by measuring the ratio of correctly predicted instances. True positives (TP) and True negatives (TN) to the total predicted result. Secondly, precision defined in Eq. 2, evaluates the proportion of true positive predictions among all positive predictions.High precision is very important in such cases where minimizing false positives is crucial. Third metric is recall computed using Eq. 3, reflects the models ability to accurately identify all actual positive instances. Lastly, F1-score as presented in Eq. 4, shows the harmonic mean of precision and recall which strikes a balance between precision and recall.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

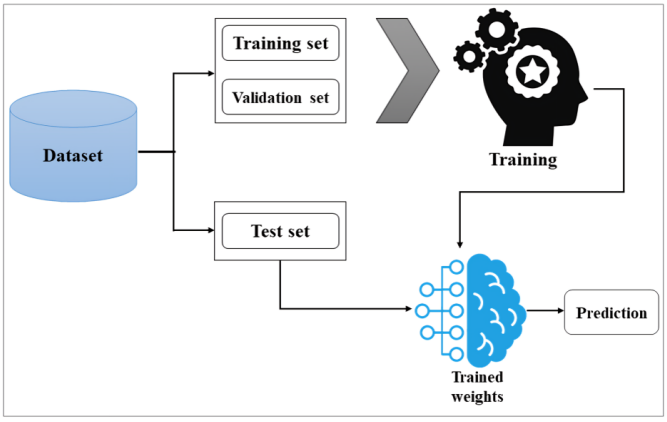


Fig. 4. Workflow for training and evaluating model using the dataset. The dataset is divided into training, validation, and test sets, which are utilized for model evaluation. The trained model then makes predictions, demonstrating the pipeline from data preparation to deployment.

C. Performance Comparison of Pre-trained Models

Performance evaluation of pretrained models is performed to get insights into their robustness in addressing the classification difficulties in the construction as well as application of gardens. These models include include Artificial Neural Network (ANN) [24], VGG16 [25], and Inception V3 [26]. The metrics used for comparison were accuracy, loss, validation accuracy (Val_Accuracy), validation loss (Va_Loss), precision, recall, and F1-score were the metrics used for comparison. which collectively provide a holistic view of each model's performance. Table III, displays the results of the ANN model. It achieved an accuracy of 0.88 with a validation accuracy of 0.83, indicating reasonable performance for a baseline model.

TABLE III. Performance Metrics of ANN, Including Accuracy, Loss, Precision, Recall, and F1-score

Metric	Accuracy	Loss	Val_Accuracy	Val_Loss	Precision	Recall	F1_Score
Values	0.88	0.19	0.83	0.21	0.85	0.82	0.83

However, the higher validation loss (0.21) compared to its training loss (0.19) suggests some overfitting, limiting its ability to generalize effectively to unseen data. The precision, recall, and F1-score of 0.85, 0.82, and 0.83, respectively, further highlight that while ANN performs decently, it falls short in addressing the complexities of the dataset. Table IV, showcases the performance of the VGG16 model, which marked a significant improvement over ANN.

With an accuracy of 0.93 and a validation accuracy of 0.91, VGG16 demonstrated strong generalization capabilities. The low training loss (0.10) and validation loss (0.13) reflect its ability to learn meaningful features from the data efficiently. Precision, recall, and F1-score values of 0.92, 0.91, and 0.91,

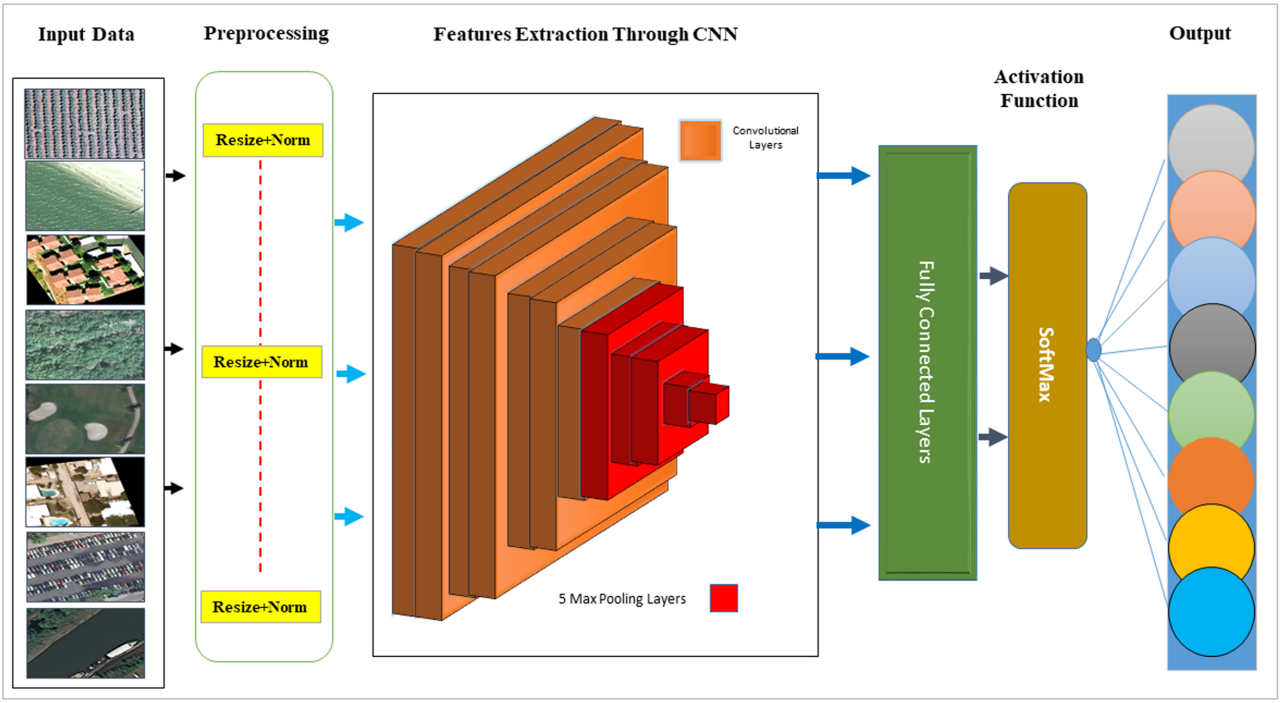


Fig. 5. Proposed framework for optimizing the construction and application of gardens through satellite image classification using a custom CNN, featuring preprocessing, feature extraction, and class-specific predictions for precise landscape planning.

TABLE IV. PERFORMANCE METRICS OF THE VGG16 MODEL, INCLUDING ACCURACY, LOSS, PRECISION, RECALL, AND F1-SCORE

Metric	Accuracy	Loss	Val_Accuracy	Val_Loss	Precision	Recall	F1_Score
Values	0.93	0.10	0.91	0.13	0.92	0.91	0.91

respectively, indicate that VGG16 reliably classifies the data with fewer false positives and negatives, making it well-suited for this task. Table V, presents the results for the Inception V3 model, which outperformed both ANN and VGG16.

TABLE V. PERFORMANCE METRICS OF THE INCEPTIONV3 MODEL, INCLUDING ACCURACY, LOSS, PRECISION, RECALL, AND F1-SCORE

Metric	Accuracy	Loss	Val_Accuracy	Val_Loss	Precision	Recall	F1_Score
Values	0.96	0.11	0.91	0.17	0.92	0.90	0.91

Inception V3 achieved an accuracy of 0.96 and a validation accuracy of 0.91, indicating its robustness in identifying intricate patterns and features within the dataset. While its validation loss (0.17) was slightly higher than that of VGG16, its precision (0.92), recall (0.90), and F1-score (0.91) demonstrate a strong balance between sensitivity and specificity. The superior performance of Inception V3 can be attributed to its advanced architecture, which excels in multi-scale feature extraction. To better illustrate the comparative performance, Fig. 6, presents a bar chart highlighting the accuracy of all three models. The progression from ANN to VGG16 and Inception V3 emphasizes the importance of employing deeper and more sophisticated architectures for tackling complex classification tasks. While ANN serves as a useful baseline, the results of VGG16 and Inception V3 underscore the potential of pretrained models in achieving higher performance levels.

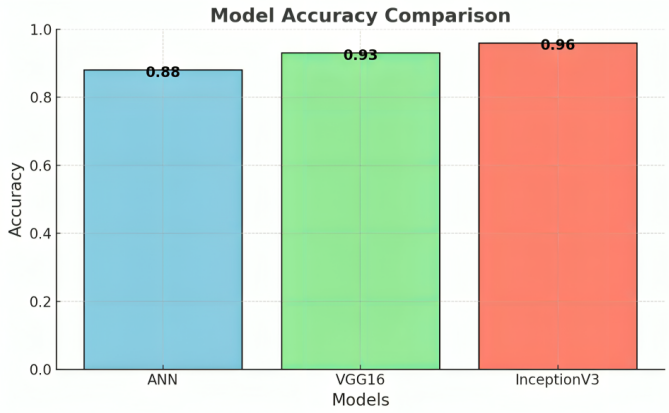


Fig. 6. Comparison of accuracy across pretrained models.

In summary, the performance evaluation highlights the effectiveness of VGG16 and Inception V3 as pretrained models, with Inception V3 emerging as the most accurate. These findings provide a benchmark for understanding the capabilities of pretrained architectures in similar applications, paving the way for further exploration and optimization in the Construction and Application of Gardens through deep learning approaches.

D. Performance Analysis of Proposed Model

The proposed Custom CNN [27] model was evaluated across varying batch sizes to analyze its performance comprehensively, with the results summarized in Table VI. Key metrics such as accuracy, loss, validation accuracy (Val-Accuracy), validation loss (Val-Loss), precision, recall, and F1-score were employed to assess the model's efficacy in handling classification tasks. Fig. 7, depicts the trends in accuracy and loss, while Fig. 8, presents a bar chart showcasing the model's performance across different batch sizes. Starting from the batch size of 4, the proposed model achieved an accuracy of 0.96 as well as validation-accuracy of 0.94. Additionally, minimal loss values of 0.04 for train and 0.03 for validation. These metrics showcases the model's best initial learning and generalization capabilities. Moreover, the model achieved the values for Precision as 0.95, recall (0.93), and F1-score 0.94. Respectively which further depicts its ability to give balanced and reliable results at this structure. At the batch of 8, the proposed model performance improved more and achieved Training accuracy of 0.97 and 0.95 of validation. Training loss and validation loss decreased to 0.03 for training and 0.02 for validation, which indicates more stable learning. Metrics such as Precision and recall rise up to 0.96 and 0.94. While F1-score reached to 0.95 which reflects the models high predictive capabilities and reduced error rate.

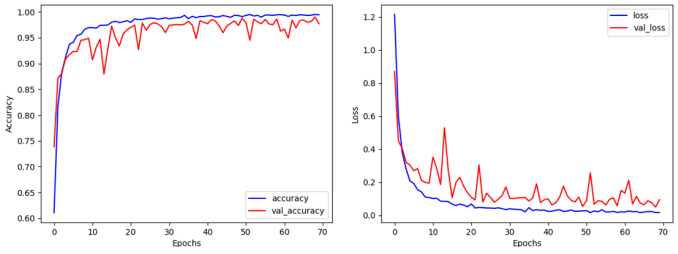


Fig. 7. Accuracy and loss curves of the proposed CNN model at different batch sizes.

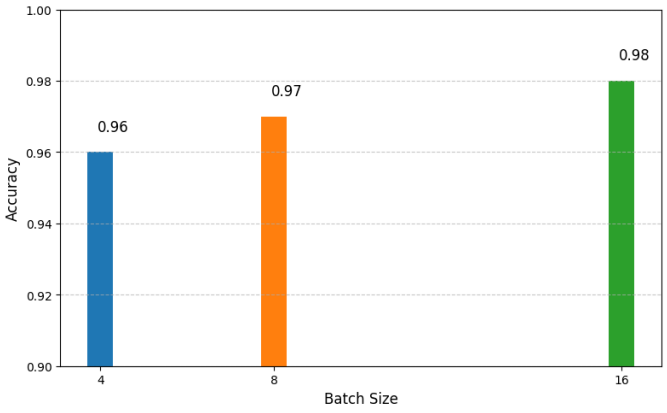


Fig. 8. Comparison of the proposed model across different batch sizes, highlighting accuracy improvements with increasing batch size.

Best results were achieved when the batch-size reached to 16. This is where the model achieved the highest accuracy of Training as 0.98 and for validation as 0.96. Training loss and validation loss were decreased to 0.02 for training and 0.01 for validation. Which depicting the model's best generalization and convergence. Metrics Precision reached at 0.97, recall rise to 0.95, and F1-score maintains its position height at 0.95. Fig. 10 highlights the comparison of metrics among the model and highlights that CNN shows superiority among the others.

TABLE VI. PERFORMANCE METRICS OF THE PROPOSED MODEL FOR DIFFERENT BATCH SIZES

Batch Size	Accuracy	Loss	Val_Accuracy	Val_Loss	Precision	Recall	F1_Score
4	0.96	0.04	0.94	0.03	0.95	0.93	0.94
8	0.97	0.03	0.95	0.02	0.96	0.94	0.95
16	**0.98**	**0.02**	**0.96**	**0.01**	**0.97**	**0.95**	**0.95**

In conclusion, the results shows the effectiveness of the proposed model, with a 16 batch size emerging as the best performer. The proposed model shows high accuracy, minimum loss. And a balanced precision, recall, and F1-score, making it more suitable for applications in the Construction and Application of Gardens. This analysis underscores the robustness of the model and its potential to facilitate sustainable garden design and management. Fig. 9, shows the accuracy of the proposed model with all ML models.

In summary, the comparative analysis of ANN, VGG16, Inception V3, and the proposed Custom CNN demonstrates a clear progression in performance, emphasizing the influence of architectural complexity and feature extraction capabilities. The relatively lower performance of ANN highlights its limitations in capturing complex patterns due to its simpler structure and limited feature learning capacity. In contrast, the balanced performance of VGG16 shows the effectiveness of moderate depth and transfer learning. Inception V3 outperforms both models by leveraging its advanced architecture for multiscale feature extraction. The proposed Custom CNN achieves the highest accuracy, particularly with a batch size of 16, due to its custom design that optimizes learning and generalization for this domain-specific application. These results underscore the importance of selecting appropriate model architectures and hyperparameters to effectively address classification challenges, paving the way for optimized solutions in the Construction and Application of Gardens.

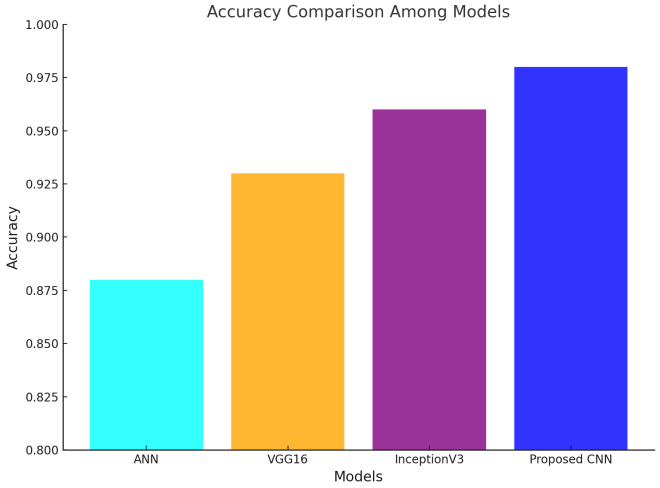


Fig. 9. Comparison accuracy of the proposed model with all ML models.

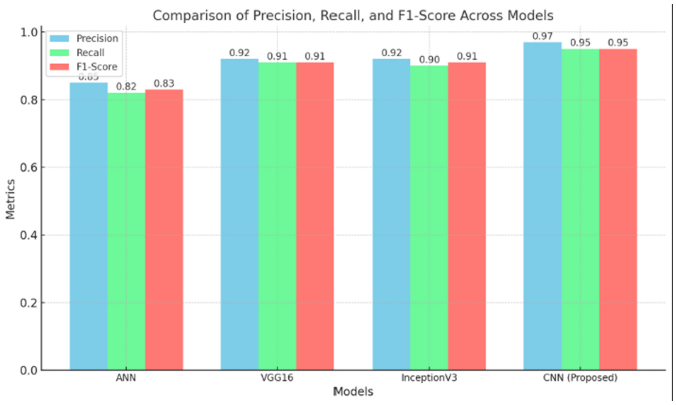


Fig. 10. Comparison accuracy of the proposed model with different models.

V. Conclusion

In conclusion,this research demonstrates the transformative potential of AI in revolutionizing garden design and sustainability through precise environmental classification and analysis. By utilizing advanced machine learning (ML) models, including both pretrained architectures and a custom CNN, this paper highlights the effectiveness of ML in accurately categorizing diverse landscape types, such as Agriculture, Beaches, Forests, and residential areas. The comprehensive evaluation of models, coupled with the use of a robust and diverse dataset, ensures the applicability of the findings across different real-world scenarios, making the work not just theoretical but practically impactful. The importance of this article lies in its ability to address pressing challenges in sustainable garden design by offering a data-driven approach to optimize planning and resource management. The results, supported by detailed performance analysis, reveal the strengths of different models while showcasing the custom CNN's superior capability in achieving high accuracy and efficient processing. The use of graphical analysis further enhances the papers clarity and accessibility, providing actionable insights for researchers and practitioners alike. This work not only sets a strong foundation for integrating AI into environmental and garden applications, but also opens doors for future advancements. The dataset can be expanded to include more complex and varied environments, and the models can be refined to handle real-time applications. By bridging the gap between technology and nature, this article paves the way for innovative, sustainable, and scalable solutions in garden construction and environmental optimization.

However, despite its promising contributions, this study is not without limitations. The dataset focuses mainly on specific landscape types like Agriculture, Beaches, Forests, and Residential areas, which may limit model generalization to more complex or mixed environments. Additionally, the current implementation lacks real-time processing capabilities, which are crucial for dynamic garden management and environmental monitoring. The models may also struggle with classification accuracy under extreme weather, varying light, or seasonal changes. Moreover, human intervention may still be required in complex or ambiguous scenarios to ensure classification accuracy. Addressing these limitations provides a balanced perspective and opens avenues for future research. Enhancing dataset diversity and improving model adaptability to extreme weather, varying light, and seasonal changes can enhance classification accuracy. Optimizing real-time processing capabilities and reducing computational demands will improve usability in dynamic garden management and environmental monitoring. Expanding the geographical scope and exploring edge computing solutions can boost scalability and practical deployment. Additionally, addressing the need for human validation in complex scenarios can refine automation accuracy. By transparently addressing these challenges, this study not only contributes to the academic field but also drives innovation at the intersection of AI and sustainable garden construction. It lays a solid foundation for future research focused on achieving environmental harmony through intelligent design and resource optimization.

Acknowledgment

We extend our heartfelt gratitude to all individuals and institutions that contributed to the completion of this research. Special thanks are directed to the developers and researchers behind the deep learning and machine learning models used in this work, whose innovations have been pivotal in advancing the field of artificial intelligence. We also wish to acknowledge the organizations and institutions that provided access to the datasets employed in this study, enabling us to conduct our experiments on diverse and representative data.

References

[1] W. Jiang and J. Luo, "Graph neural network for traffic forecasting: A survey," *Expert systems with applications*, vol. 207, p. 117921, 2022.

[2] M. S. Ma, *Place and Placelessness: An Ecocritical Approach Towards the Mountain Imagery in Country Music*. PhD thesis, Duquesne University, 2024.

[3] W. Jiang, "Applications of deep learning in stock market prediction: recent progress," *Expert Systems with Applications*, vol. 184, p. 115537, 2021.

[4] M. Di Paola, "Green and smart visions of urban futures," in *Greentopia: Utopian Thought in the Anthropocene*, pp. 207–226, Springer, 2024.

[5] C. X. Hui, G. Dan, S. Alamri, and D. Toghraie, "Greening smart cities: An investigation of the integration of urban natural resources and smart city technologies for promoting environmental sustainability," *Sustainable Cities and Society*, vol. 99, p. 104985, 2023.

[6] Y. Xing, W. Gan, and Q. Chen, "Artificial intelligence in landscape architecture: A survey," *arXiv preprint arXiv:2408.14700*, 2024.

[7] A. A. Zuniga-Teran, C. Staddon, L. de Vito, A. K. Gerlak, S. Ward, Y. Schoeman, A. Hart, and G. Booth, "Challenges of mainstreaming green infrastructure in built environment professions," *Journal of Environmental Planning and Management*, vol. 63, no. 4, pp. 710–732, 2020.

[8] M. Carmona, *Public places urban spaces: The dimensions of urban design*. Routledge, 2021.

[9] R. Stephens, "Green cities artificial intelligence," 2023.

[10] O. B. Akintuyi *et al.*, "Vertical farming in urban environments: a review of architectural integration and food security," *Open Access Research Journal of Biology and Pharmacy*, vol. 10, no. 2, pp. 114–126, 2024.

[11] D. Cembrowska-Lech, A. Krzemińska, T. Miller, A. Nowakowska, C. Adamski, M. Radaczyńska, G. Mikiciuk, and M. Mikiciuk, "An integrated multi-omics and artificial intelligence framework for advance plant phenotyping in horticulture," *Biology*, vol. 12, no. 10, p. 1298, 2023.

[12] V. Patil, J. Patil, A. Kadam, A. R. Patil, D. Mokashi, and G. M. Lonare, "Ai-driven green space optimization for sustainable urban parks: Enhancing biodiversity and resource efficiency," *Library Progress International*, vol. 44, no. 3, pp. 3412–3417, 2024.

[13] Z. Li, B. Chen, S. Wu, M. Su, J. M. Chen, and B. Xu, "Deep learning for urban land use category classification: A review and experimental assessment," *Remote Sensing of Environment*, vol. 311, p. 114290, 2024.

[14] L. Yang, J. Driscol, S. Sarigai, Q. Wu, H. Chen, and C. D. Lippitt, "Google earth engine and artificial intelligence (ai): a comprehensive review," *Remote Sensing*, vol. 14, no. 14, p. 3253, 2022.

[15] X. Tang and W.-j. Chung, "Automated urban landscape design: an ai-driven model for emotion-based layout generation and appraisal," *PeerJ Computer Science*, vol. 10, p. e2426, 2024.

[16] A. S. Rathor, S. Choudhury, A. Sharma, P. Nautiyal, and G. Shah, "Empowering vertical farming through iot and ai-driven technologies: A comprehensive review," *Heliyon*, 2024.

[17] E. Appolloni, F. Orsini, K. Specht, S. Thomaier, E. Sanyé-Mengual, G. Pennisi, and G. Gianquinto, "The global rise of urban rooftop agriculture: A review of worldwide cases," *Journal of Cleaner Production*, vol. 296, p. 126556, 2021.

[18] E. E. K. Senoo, L. Anggraini, J. A. Kumi, B. K. Luna, E. Akansah, H. A. Sulyman, I. Mendonça, and M. Aritsugi, "Iot solutions with artificial intelligence technologies for precision agriculture: definitions, applications, challenges, and opportunities," *Electronics*, vol. 13, no. 10, p. 1894, 2024.

[19] A. Yu, "Environmental projects management (on the example of the "green oasis" urban garden project)," Master's thesis, Lesya Ukrainka Volyn National University, 2024.

[20] K. Nguyen, E. Thaens, D. S. Can, E. Maya Gracia, A. Abdalla, and S. Peeters, "Examining the influence of adaptive ai integration in vr experiences for raising awareness of biodiversity loss," B.S. thesis, Universitat Politècnica de Catalunya, 2024.

[21] M. Javaid, A. Haleem, I. H. Khan, and R. Suman, "Understanding the potential applications of artificial intelligence in agriculture sector," *Advanced Agrochem*, vol. 2, no. 1, pp. 15–30, 2023.

[22] A. Holzinger, A. Saranti, A. Angerschmid, C. O. Retzlaff, A. Gronauer, V. Pejakovic, F. Medel-Jimenez, T. Krexner, C. Gollob, and K. Stampfer, "Digital transformation in smart farm and forest operations needs human-centered ai: challenges and future directions," *Sensors*, vol. 22, no. 8, p. 3043, 2022.

[23] M. Sourek, "Artificial intelligence in architecture and built environment development 2024: A critical review and outlook,"

[24] M. Fadavi Amiri, M. Hosseinzadeh, and S. M. R. Hashemi, "Improving image segmentation using artificial neural networks and evolutionary algorithms," *International Journal of Nonlinear Analysis and Applications*, vol. 15, no. 3, pp. 125–140, 2024.

[25] A. MAYANJA, I. A. OZKAN, and Ş. TAŞDEMİR, "Utilizing transfer learning on landscape image classification using the vgg16 model," in *Proceedings of the International Conference on Advanced Technologies*, vol. 11, pp. 71–76, 2023.

[26] M. N. Khan, S. Das, and J. Liu, "Predicting pedestrian-involved crash severity using inception-v3 deep learning model," *Accident Analysis & Prevention*, vol. 197, p. 107457, 2024.

[27] T. L. Giang, Q. T. Bui, T. D. L. Nguyen, Q. H. Truong, T. T. Phan, H. Nguyen, M. Yasir, K. B. Dang, *et al.*, "Coastal landscape classification using convolutional neural network and remote sensing data in vietnam," *Journal of Environmental Management*, vol. 335, p. 117537, 2023.

Multi-Objective Osprey Optimization Algorithm-Based Resource Allocation in Fog-IoT

Nagarjun E, Dharamendra Chouhan, Dilip Kumar S M
Department of Computer Science and Engineering-University of Visvesvaraya College of Engineering,
Bangalore University, Bengaluru, India

***Abstract*—Fog Computing (FC) paradigm offers significant potential for hosting diverse delay-sensitive Internet of Things (IoT) applications. However, the limited resources of fog devices pose significant challenges for deploying multiple applications, particularly in heterogeneous and dynamic IoT scenarios, due to the absence of effective mechanisms for resource estimation and discovery. An efficient resource allocation strategy is crucial for meeting the Quality of Service (QoS) requirements of IoT applications while enhancing overall system performance. Identifying the optimal allocation strategy for IoT applications with multiple QoS parameters is a complex and computationally intensive challenge, classified as an NP-complete problem. This paper proposes a Multi-Objective Optimization Algorithm (MOOA) for optimal resource allocation using the Osprey Optimization Algorithm (OOA) to efficiently allocate available resources. The proposed algorithm was evaluated against existing approaches, including the Genetic Algorithm (GA) and Particle Swarm Optimization (PSO), under varying task loads ranging from 100 to 500 tasks. The simulation results demonstrate significant performance improvements, including an average reduction in execution time by 12.45% compared to PSO and 22.97% compared to GA, response time by 32.57% compared to GA and 24.45% compared to PSO, and completion time by 44.39% compared to GA and 33.23% compared to PSO. These findings highlight the proposed algorithm's ability to efficiently handle task allocation in dynamic FC environments and its potential to address complex QoS requirements in real-world IoT applications.**

***Keywords*—Fog computing; IoT; resource allocation and reallocation; task allocation**

I. Introduction

The Internet of Things (IoT) has revolutionized data generation, driving advancements in healthcare monitoring, virtual reality, industrial automation, and many other applications that demand reliable, low-latency communication and computational services [1]. These applications often impose stringent QoS requirements, such as minimizing delay and energy consumption, which pose significant challenges for IoT devices [2], [3].

Fog computing (FC) is a computing paradigm placed in the middle of a three tiered architecture that encompasses the cloud, fog nodes (FNs), and IoT devices at the edge. Cisco introduced this layer in 2012 to bring the features and capabilities of cloud computing nearer to data sources and user devices that communicate over the internet [4], [5], [6]. FC is recognized as a key solution to the limitations of cloud computing, particularly for delay-sensitive applications. However, while FC provides a decentralized architecture, the challenge of selecting suitable FNs for application modules with diverse deadline constraints remains a critical issue [7].

Resource allocation in FC entails the distribution of storage, computational, and communication resources among FNs, users, and service providers. These entities often have conflicting objectives and requirements, necessitating mechanisms that balance their interests while optimizing overall system performance [8]. Furthermore, the dynamic nature of FC characterized by fluctuations in resource availability and user demand highlights the need for adaptive allocation strategies. Effective resource allocation and reallocation in FC are particularly crucial for IoT tasks, where metrics such as resource utilization, execution time, response time, and completion time are key considerations [9].

To address these challenges, this study proposes a Multi-Objective Osprey Optimization Algorithm (MOOA) to allocate the appropriate number of fog resources corresponding to fluctuations in IoT task demands.

- Proposed a meta-heuristic MOOA for resource allocation and reallocation in FC for IoT tasks to handle the dynamic and heterogeneous requirements of IoT tasks.
- Implement the proposed algorithm and evaluate its performance against GA and PSO by varying the number of IoT tasks and exploring different parameters.
- The results demonstrate a significant reduction in average response time, execution time and completion, showcasing the efficiency of our approach.

The rest of this paper is organized as follows: Section II provides a detailed background on the problem of resource allocation improvement in FC environments, highlighting the drawbacks of existing methods. Section III presents the system model, problem statement, and objectives. Section IV describes the proposed algorithm and its key features. In Section V, we present our simulation and experimental results, comparing the performance of our algorithm with existing approaches. Finally, Section VI concludes the paper and discusses future research directions in this area.

II. Related Works

Hussain *et al.* [10] developed the Resource Aware Prioritized Task Scheduling (RAPTS) approach for heterogeneous FC environments. The primary objective of RAPTS is to ensure the execution of tasks with strict deadlines while optimizing response time, cost, and makespan, and enhancing resource utilization within the fog layer. This approach was implemented in iFogSim and evaluated based on various performance metrics, including response time, resource utilization, task deadline compliance, cost, and makespan. Comparative analysis

with advanced fog schedulers like RACE (CFP) and RACE (FOP) demonstrated that RAPTS achieved improvements of up to 29%, 53%, 15%, 11%, and 43% in resource utilization, response time, makespan, cost, and task deadline adherence, respectively. However, completion time and execution time are not addressed.

Arshed *et al.* [11] introduced the Resource Aware Cost-Efficient Scheduler (RACE) to allocate incoming application modules to fog devices. The scheduler manages to optimize resource utilization at the fog layer by minimizing the monetary cost of using cloud resources, and reduces the application execution time and bandwidth usage. It incorporates two key algorithms: the ModuleScheduler, which classifies application modules based on their computational and bandwidth requirements, and the CompareModule, which determines their placement. Simulation results indicate that RACE outperforms traditional cloud placement strategies and baseline algorithms in most scenarios. However, the approach does not specifically address response time, completion time, and execution time.

Khan *et al.* [12] introduced an improved Ripple-Induced Whale Optimization Algorithm (RWOA) for scheduling independent tasks in fog-cloud environments. The method leverages ripple effects to refine suboptimal solutions, aiming to minimize makespan and energy consumption while enhancing throughput. Despite its effectiveness, the approach does not explicitly consider metrics such as response time, completion time, and execution time. Chafi *et al.* [13] proposed a novel algorithm based on Particle Swarm Optimization (PSO) to optimize energy consumption and workflow time in heterogeneous FC environments. By utilizing the collective behavior of particles, the algorithm effectively explores the solution space and adapts to the dynamic and unpredictable nature of FC resources. Simulation results demonstrate notable enhancements in workflow completion time, energy efficiency, and resource utilization. However, the approach does not explicitly consider metrics such as response time, and execution time.

Bandopadhyay *et al.* [8] proposed a Game-Theoretic Resource Allocation and Dynamic Pricing Mechanism in FC (GTRADPMFC). The model incorporates non-cooperative competition among FNs for resource allocation and employs dynamic pricing mechanisms to promote efficient resource usage. Through theoretical evaluation and simulation experiments, the study demonstrated that GTRADPMFC enhances both resource efficiency and the overall performance of FC systems. Furthermore, the paper outlines methods to manage scenarios with insufficient data samples and provides flexibility for users who cannot meet specific completion time requirements. GTRADPMFC optimizes resource allocation by establishing pricing strategies while accounting for potential delays in task completion. The research also includes simulations, convergence analyses, complexity assessments, and guarantees for optimization. However, the model does not explicitly address performance metrics such as response time and execution time.

Mokni *et al.* [14] proposed decision-making phase that analyzes data generated by IoT devices to identify the optimal offloading strategy. To accomplish this, the TOF-NSGAII approach was introduced, aiming to reduce both energy consumption and latency during the offloading of IoT tasks in a fog-cloud environment. Each IoT device transmits tasks with distinct attributes, such as input data size, and the algorithm assigns these tasks to specific virtual machines to optimize resource utilization. By utilizing the combined resources of fog and CC, the TOF-NSGAII approach effectively meets its goals of minimizing energy usage and latency, as validated by experimental results. However, the model does not explicitly address performance metrics such as response time and execution time.

Raissouli *et al.* [15] developed an improved version of NSGA II, namely, reinforcement weighted probabilistic NSGA II, which uses weighted probabilistic mutation. This algorithm replaces random mutation with probabilistic mutation to enhance exploration of the solution space. This method uses domain-specific knowledge to improve convergence and solution quality, resulting in improved energy efficiency and reduced delay compared to traditional NSGA II and other evolutionary algorithms. However, the model does not explicitly address performance metrics such as response time, completion time and execution time.

The studies described above show that various efforts have been made to solve resource allocation and task scheduling problems in FC environments. However, the efficiency of algorithms in handling complex and dynamic FC scenarios across varying conditions has yet to be thoroughly investigated. Additionally, existing algorithms do not explicitly consider key performance metrics such as response time, completion time, and execution time. In contrast, our proposed approach addresses these limitations by incorporating these critical metrics to enhance the effectiveness of resource allocation and reallocation in FC.

III. System Model

An IoT network consisting of IoT devices and FNs is considered, where the FNs vary in energy and processing power capabilities, making task scheduling a challenging problem. In this setup, we have I IoT devices and F heterogeneous FNs, represented by sets $I = \{1, 2, 3, \ldots, I\}$ and $F = \{1, 2, 3, \ldots, F\}$, respectively. The IoT devices collect data from a range of sensors and transmit it to the fog layer. It is assumed that the IoT devices offload tasks directly to the FNs without any local processing. Each FN is equipped with an agent tasked with receiving and processing data from the IoT devices. Task scheduling is assumed to occur in a distributed manner, with each FN acting as a scheduler upon receiving data from IoT devices. A task is represented as $T_t(S_t, C_t)$, where S_t is the task size in KB and C_t is the task complexity in Million Instructions (MI). Since the end devices often lack sufficient computational resources to process the tasks, the tasks are offloaded to nearby FNs for processing. The commonly-used symbols are delineated in Table I.

A. Problem Statement and Objectives

The problem is defined as finding optimal resource allocation for IoT tasks in a FC environment. Each fog node serves as a resource for processing tasks, and the goal is to allocate these tasks efficiently with the following objectives:

1) Objectives: The objectives of the proposed work are as follows:

- Minimize completion time for processing IoT tasks.
- Minimize response time for IoT tasks.
- Minimize execution time for IoT tasks.
- Maximize the resource utilization.

TABLE I. NOTATION TABLE FOR THE MULTI-OBJECTIVE OSPREY OPTIMIZATION ALGORITHM (MOOA)

Notation	Description
T	Set of IoT tasks, where $T = \{t_1, t_2, \ldots, t_N\}$.
FN	Set of Fog Nodes (FNs), where $FN = \{fn_1, fn_2, \ldots, fn_M\}$.
N	Population size (number of osprey solutions in the search space).
T_{max}	Total number of iterations for the optimization process.
R	Population matrix representing resource allocation solutions, where $X = \{R_1, R_2, \ldots, R_N\}$.
RP_i	A single solution in the population, representing task-to-node allocations.
$r_{i,j}$	Position of the j^{th} dimension of solution R_i, indicating resource allocation for task j.
$r_{i,j}^{P1}$	Updated position of the j^{th} dimension for osprey i during optimization.
$r_{i,j}^{P2}$	Updated position of the j^{th} dimension for osprey i during optimization.
lb_j, ub_j	Lower and upper bounds for the j^{th} dimension of the solution space.
$r_{i,j}$	A random number generated for the j^{th} dimension of osprey i.
$F(X_i)$	Fitness value of solution X_i, considering multiple objectives.
f	Objective function
X^*	Best solution obtained after optimization.
t_{new}	Newly arriving IoT task that requires resource reallocation.
fn_{fluct}	Fog node with fluctuating resource availability due to dynamic conditions.

IV. PROPOSED WORK

In this section, the details of the proposed Multi-Objective Optimization Algorithm (MOOA) are presented. The diagram illustrating the proposed architecture is presented in Fig. 1.

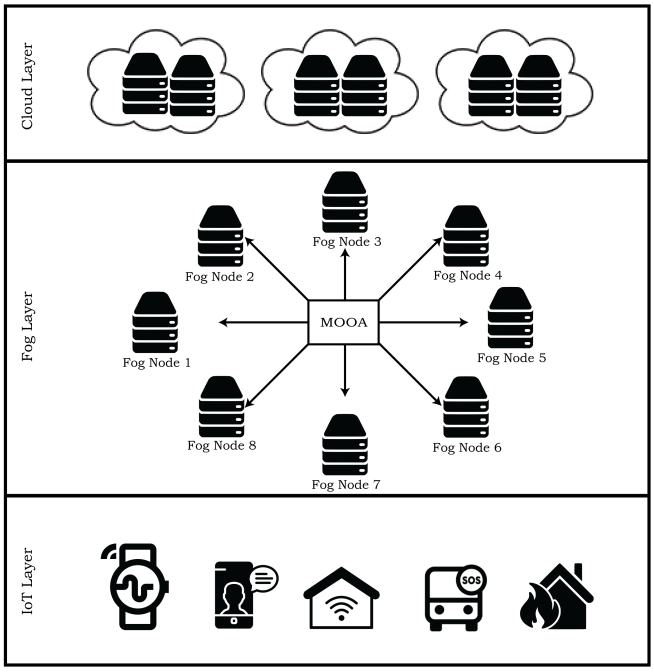


Fig. 1. Proposed MOOA architecture.

A. OOA Overview

The Osprey Optimization Algorithm (OOA) is inspired by the hunting and behavior patterns of the osprey, a bird of prey that primarily feeds on fish. Known for its sharp vision and specialized hunting strategy, the osprey locates fish underwater from a height of 10 to 40 meters before diving to capture them. After catching its prey, the osprey takes it to a safe location to consume it. This natural behavior, marked by precision and efficiency, forms the basis of the OOA, which applies these characteristics to create an effective optimization technique [16], [17]. In the context of FC, the OOA is adapted to allocate and reallocate resources efficiently among tasks and FNs. The OOA is designed to optimize multiple objectives simultaneously, such as minimizing response time, execution time, completion time, and balancing the load among FNs. The algorithm operates in three key phases: initialization, iterative optimization, and updating.

B. Initialization Phase

In the initialization phase, a population of N ospreys is randomly generated. Each osprey in the population represents a candidate solution, where its position encodes a potential allocation of resources. The matrix representation of the population can be modeled as shown in Eq. 1. The position of the i-th osprey in the j-th dimension is represented by $r_{i,j}$, where $r_{i,j}$ corresponds to a problem variable related to resource allocation (e.g. CPU, RAM and Bandwidth). The initial positions are within the predefined bounds of the problem variables.

$$R = \begin{bmatrix} R_1 \\ \vdots \\ R_i \\ \vdots \\ R_N \end{bmatrix}_{N\times m} = \begin{bmatrix} r_{1,1} & \cdots & r_{1,j} & \cdots & r_{1,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ r_{i,1} & \cdots & r_{i,j} & \cdots & r_{i,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ r_{N,1} & \cdots & r_{N,j} & \cdots & r_{N,m} \end{bmatrix}_{N\times m}, \tag{1}$$

Where N is the number of candidate solutions (ospreys), m is the number of tasks to be allocated, $r_{i,j}$ represents the allocation of resources from a FN to the j-th task in the i-th candidate solution. At the beginning of the algorithm, the initial allocation matrix is generated randomly to ensure diversity. The positions of the ospreys in the resource allocation search space are initialized using Eq. 2:

$$r_{i,j} = lb_j + r_{ij} \cdot (ub_j - lb_j), i = 1, 2, \ldots, N, j = 1, 2, \ldots, m, \tag{2}$$

Where lb_j and ub_j are the lower and upper bounds of resource availability for the j-th task, $r_{i,j}$ is a random number uniformly distributed in the range $[0, 1]$, ensuring a diverse initial allocation.

C. Objective Function

The main goal of task allocation in a FC environment is to minimize the time needed for task completion, response, and execution. The values obtained from evaluating the objective function for the given problem from Eq. (4) to (10) can be expressed as a vector, as outlined in Eq. (3).

$$F = \begin{bmatrix} F_1 \\ \vdots \\ F_i \\ \vdots \\ F_N \end{bmatrix}_{N\times 1} = \begin{bmatrix} F(R_1) \\ \vdots \\ F(R_i) \\ \vdots \\ F(R_N) \end{bmatrix}_{N\times 1}, \quad (3)$$

1) Execution Time: The execution time $E_{i,j}$ of $task_i$ on FN_j can be expressed as:

$$E_{i,j} = \frac{T_{i,j}}{C_{i,j}} \quad (4)$$

Where $T_{i,j}$ is task length for $task_i$ on FN_j (number of instructions). $CC_{i,j}$ is computational capacity of FN_j, calculated as:

$$CC_{i,j} = P_{i,j} \times M_{i,j} \quad (5)$$

Where $P_{i,j}$ is number of processing elements (PEs) in FN_j. $M_{i,j}$ is MIPS rate of FN_j.

2) Response Time: The response time $RT_{i,j}$ for $task_i$ on FN_j is calculated as:

$$RT_{i,j} = F_{i,j} - S_{i,j} \quad (6)$$

Where $F_{i,j}$ is finish time of $task_i$ on FN_j. $S_{i,j}$ is start time of $task_i$ on FN_j.
The finish time $F_{i,j}$ is given by:

$$F_{i,j} = R_{i,j} + E_{i,j} \quad (7)$$

3) Completion Time: The completion time $CT_{i,j}$ for $task_i$ on FN_j is calculated as:

$$CT_{i,j} = F_{i,j} + C_{T_{i,j}} \quad (8)$$

Where $C_{T_{i,j}}$ is communication time for $task_i$ on FN_j. Thus, the completion time $CT_{i,j}$ is:

$$CT_{i,j} = (RT_{i,j} + E_{i,j}) + C_{T_{i,j}} \quad (9)$$

The objective function value is evaluated at each position of the osprey:

$$F_i = [F_1(E), F_2(RT), F_3(CT)]^T \quad (10)$$

D. Phase 1: Position Identification and Resource Allocation

In fog resource allocation, resource updates can mimic osprey hunting behavior. Just as ospreys identify and target fish underwater, the algorithm identifies better resource positions in the search space to enhance exploration and escape local optima. For each resource, the positions of other resources in the search space that have better objective function values are considered optimal targets. These targets are analogous to the "fish" in the search space. The set of optimal resources for each resource is defined in Eq. (11).

$$RP_i = \{R_k \mid k \in \{1, 2, \ldots, N\} \wedge R_k < F_i\} \cup \{R_{\text{best}}\} \quad (11)$$

Here, R_k represents the resources with better objective values, and R_{best} is the best resource found so far. This approach helps guide the allocation process towards more efficient solutions in FC.

$$r_{ij}^{P1} = r_{ij} + r_{ij} \cdot (SF_{ij} - I_{ij} \cdot r_{ij}), \quad (12a)$$

$$r_{ij}^{P_1} = \begin{cases} r_{ij}^{P_1}, lb_j \le r_{ij}^{P_1} \le ub_j; \\ lb_j, r_{ij}^{P_1} < lb_j; \\ ub_j, r_{ij}^{P_1} > ub_j. \end{cases} \quad (12b)$$

The updated resources R_i are refined based on their fitness values:

$$R_i = \begin{cases} R_i^{P1}, F_i^{P_1} < F_i; \\ R_i, \text{ else }, \end{cases} \quad (13)$$

This approach ensures efficient resource allocation and optimizes performance in FC systems.

E. Phase 2: Identifying the Suitable Resources

After identifying a suitable resource node, the FC system allocates tasks to the selected node for processing in an efficient and resource-aware manner. This phase models the process of refining resource allocation decisions to ensure that the tasks are processed in a way that minimizes costs and optimizes system performance. The refinement of allocation decisions involves making small adjustments to the assigned resources, enhancing the exploitation capabilities of the resource allocation algorithm, and converging toward better solutions near the currently identified optimal nodes.

In the design of the FC allocation algorithm, this behavior is simulated by calculating a new potential resource allocation for each task based on predefined criteria using Eq. (14a), (14b). If the new allocation improves the value of the objective function (e.g. response time, completion time, execution time, or improving resource utilization), the algorithm updates the allocation to this new configuration according to Eq. (15). This iterative refinement ensures more efficient exploitation of available resources and promotes convergence toward the most optimal allocation.

$$r_{i,j}^{p_2} = r_{i,j} + \frac{lb_j + r \cdot (ub_j - lb_j)}{t}, i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, m, t = 1, 2, \ldots, T \quad (14a)$$

After updating the position, the new position is constrained within the bounds:

$$r_{i,j}^{P2} = \begin{cases} r_{i,j}^{P2}, lb_j \le r_{i,j}^{P2} \le ub_j; \\ lb_j, r_{i,j}^{P2} < lb_j; \\ ub_j, r_{i,j}^{P2} > ub_j, \end{cases} \quad (14b)$$

$$R_i = \begin{cases} R_i^{P2}, F_i^{P2} < F_i; \\ R_i, \text{ else }, \end{cases} \quad (15)$$

else, the osprey remains at its current position.

Algorithm 1: MOOA for Resource Allocation and Reallocation in FC for IoT Tasks

Input: Task set $T = \{t_1, t_2, ..., t_N\}$, Fog Node set $FN = \{fn_1, fn_2, ..., fn_M\}$, T_{max}, N, Resource constraints (e.g., CPU, MIPS, memory, energy per unit), f.
Output: Optimized resource allocation and reallocation solutions for IoT tasks in the fog network.
/* Phase 1: Initial Resource Allocation */
Step 1: Initialize the Population;
Generate initial population matrix with random resource allocation positions for each task to FNs using using (1) and (2).
Step 2: Evaluate Initial Fitness;
foreach $X_i \in X$ **do**
  Compute fitness for each solution using objective function f by (3)
for $t \leftarrow 1$ **to** T_{max} **do**
  foreach *osprey* $i \in \{1, 2, ..., N\}$ **do**
    foreach *task-to-node mapping dimension* $j \in \{1, 2, ..., m\}$ **do**
      /* Phase 1: Position identification and Resource Allocation */
      /* Calculate New Position for Resource Allocation */
      Generate random number $r_{i,j}$;
      Update position for osprey i using (11).
      Calculate new position $x_{P1_{i,j}}$ using (12a).
      Check boundary conditions for $x_{P1_{i,j}}$ using (12b).
      Update R_i using (13).
    /* Phase 2: Identifying the suitable resources */
    /* Evaluate New Fitness for the Updated Allocation */
    Compute new fitness by using (3).
    Calculate new position $x_{P2_{i,j}}$ using (14a).
    Check boundary conditions for $x_{P2_{i,j}}$ using (14b).
    /* Replace if New Solution is Better */
    Update R_i using (15).
Output: Initial resource allocation solution X^*;
/* Phase 2: Resource Reallocation */
Step 3: Monitor Task and Resource Fluctuations;
Collect real-time information about task arrivals, resource availability, and execution progress in FNs;
Step 4: Reallocate Resources Dynamically;
foreach *new task* t_{new} *or fluctuating resource* fn_{fluct} **do**
  foreach *solution* $X_i \in X$ **do**
    foreach *task-to-node mapping dimension* j **do**
      /* Phase 1: Position identification and Resource Allocation */
      /* Recompute Positions for Reallocation */
      Update resource mapping based on osprey behavior and fluctuating resource constraints by using (11).
      Calculate new position $x_{P1_{i,j}}$ using (12a).
      Check boundary conditions for $x_{P1_{i,j}}$ using (12b).
      Update R_i using (13).
    /* Phase 2: Identifying the suitable resources */
    /* Reevaluate Fitness for Reallocation */
    Compute by using (3).
    Calculate new position $x_{P2_{i,j}}$ using (14a).
    Check boundary conditions for $x_{P2_{i,j}}$ using (14b).
    /* Replace Solution if Reallocation is Better */
    Update R_i using (15).
Output: Final optimized solution X^* for resource allocation and reallocation;

F. Termination Criterion

The optimization continues until a termination criterion is met, such as a maximum number of iterations T_{max} or a satisfactory solution is found.

G. Iterative Process, Flowchart, and Algorithm of MOOA

The MOOA for resource allocation and reallocation in FC for IoT tasks is designed to handle the dynamic and heterogeneous requirements of IoT tasks. The steps for implementing the MOOA are illustrated as the flowchart in Fig. 2, and their

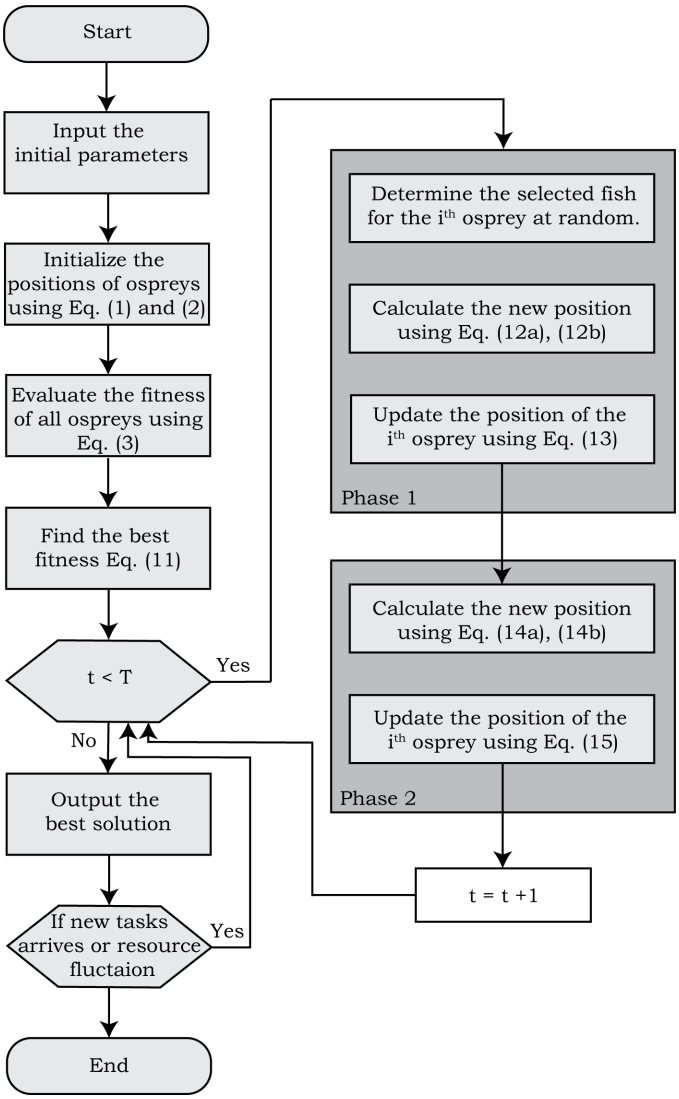


Fig. 2. Flowchart of MOOA.

procedural details are outlined in Algorithm I as pseudocode.

Phase 1: Initial Resource Allocation: The algorithm begins by generating an initial population matrix X, where each solution represents a potential mapping of tasks to fog nodes. The positions in the matrix are initialized randomly within the defined bounds for resource constraints (e.g., CPU, MIPS, memory). Each solution is evaluated using multi-objective fitness functions mentioned in the Eq. (10). For each iteration t of the optimization process, every osprey (solution) adjusts its position by considering random factors ($r_{i,j}$) and the bounds of resource constraints (lb_j, ub_j). The new position is calculated for each dimension (task-to-node mapping), ensuring the solution stays within the defined bounds. The fitness of the updated position is evaluated for all objective functions. If the updated position improves the fitness of the solution, it replaces the existing solution. This iterative process continues until the maximum number of iterations ($T_{\max}$) is reached, producing the initial resource allocation solution X^*.

Phase 2: Dynamic Resource Reallocation: In the second phase, the algorithm monitors the fog network in real-time to account for fluctuations in task arrivals and resource availability. Information about new tasks, resource constraints, and ongoing execution is collected. For every new task or fluctuating resource, the algorithm adjusts the resource mapping dynamically. Each solution in the population is reevaluated by recalculating the task-to-node mappings based on osprey behavior, incorporating updated resource constraints. The positions are updated using random factors ($r_{i,j}$) while ensuring they remain within the bounds of resource availability. The fitness of each updated solution is computed, and if the reallocation improves the solution, it replaces the current one. The algorithm outputs the optimized resource allocation and reallocation solution X^*, balancing the objectives of equation (10). This ensures efficient and adaptive resource management in FC environments for IoT tasks.

V. Simulation Results and Discussion

This section presents the evaluations conducted to demonstrate the effectiveness of the proposed approach. The experiments were performed using the iFogSim2 [18] simulation platform, and performance of the proposed MOOA algorithm was compared with existing algorithms, including GA [19] and PSO [20]. The simulations were run on a Windows 11 system equipped with an Intel(R) Core(TM) i5-9300H CPU operating at 2.40 GHz. The experimental parameters are related to various IoT tasks, the fog and cloud computing environment, and the configuration of the MOOA algorithm. Details of the experimental settings are provided in Table II.

TABLE II. Simulation Parameters

Parameters	**Values**
Simulation tool	iFogSim2
System Architecture	x86
OS	Linux
VMM	Xen
No. of FNs	50
RAM	4,000 (MB)
Storage	1000000
BW	10000
GA Parameters	
Simulation parameters	Value
Number of iterations	60
Population size	10
Mutation probability	0.2s
PSO Parameters	
Number of particles	30
Iterations	100
Inertia constant	0.85
Cognitive constant	1
Social constant	2
MOOA Parameters	
Number of ospreys	30
Iterations	100
Experiment 1	
Purpose	Heterogenous task
Data input parameters	100, 200, 300, 400, 500
FN parameters	50
Experiment 2	
Purpose	Heterogenous nodes
Data input parameters	200
FN parameters	100, 200, 300, 400, 500

A. Experiment 1

Fig. 3 compares the execution time for varying task loads (100 to 500 tasks) among the proposed algorithm, GA, and

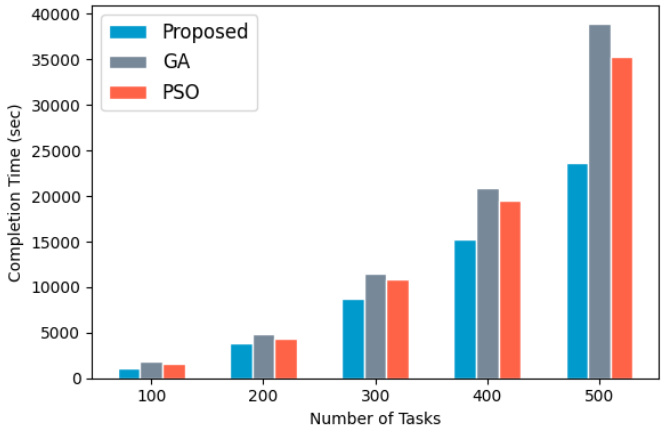


Fig. 3. Execution time of tasks with 50 FNs.

PSO. As expected, execution time increases with the number of tasks. The proposed algorithm achieved a significant reduction in execution time, showing an average improvement of 12.45% over PSO and 22.97% over GA. This underscores the proposed algorithm's ability to handle tasks efficiently in FC environments.

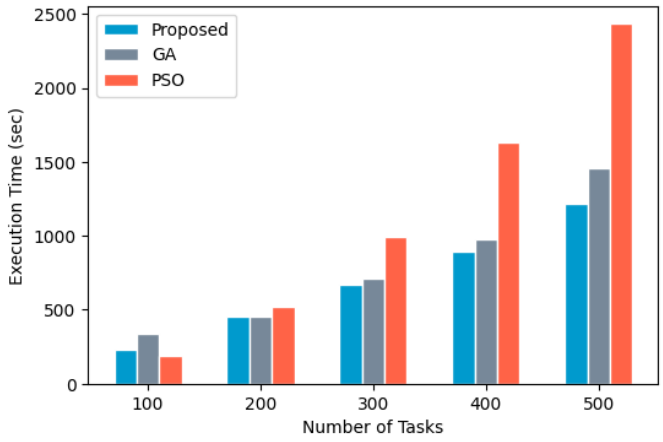


Fig. 4. Response time of tasks with 50 FNs.

Fig. 4 illustrates the response time for different task loads. The proposed algorithm consistently outperformed GA and PSO, achieving an average response time improvement of 32.57% compared to GA and 24.45% compared to PSO. This highlights the robustness of the proposed algorithm in minimizing delay and enhancing performance.

Fig. 5 depicts the completion time comparison across task loads. The proposed algorithm demonstrated superior performance, with an average reduction of 44.39% compared to GA and 33.23% compared to PSO. This confirms the proposed algorithm's scalability and effectiveness in optimizing computational processes in FC environments.

B. Experiment 2

Fig. 6 compares the execution time for varying FNs (10 to 50 nodes) among the proposed algorithm, PSO and GA.

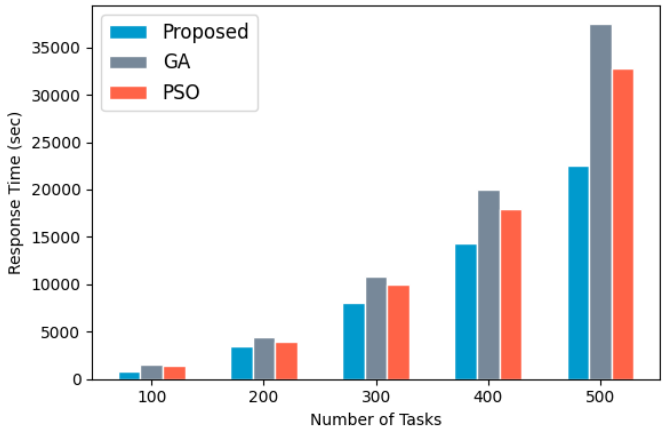


Fig. 5. Completion time of tasks with 50 FNs.

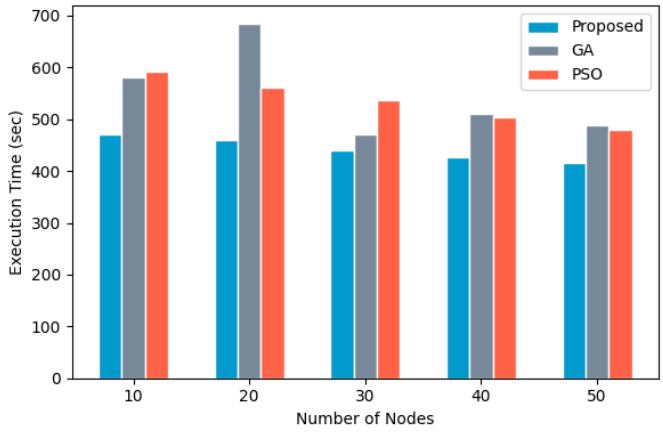


Fig. 6. Execution time of 200 tasks with different FNs.

As expected, execution time decreases with the number of FNs. The proposed algorithm achieved a significant reduction in execution time, showing an average improvement of 17.16% over PSO and 18.02% over GA. This underscores the proposed algorithm's ability to handle tasks efficiently in FC environments.

Fig. 7 illustrates the response time comparison across for varying FNs (10 to 50 nodes). The proposed algorithm consistently outperformed GA and PSO, achieving an average response time improvement of 39.65% compared to GA and 17.57% compared to PSO. This highlights the robustness of the proposed algorithm in minimizing delay and enhancing performance. Fig. 8 depicts the completion time comparison across for varying FNs (10 to 50 nodes). The proposed algorithm demonstrated superior performance, with an average reduction of 36.44% compared to GA and 16.58% compared to PSO. This confirms the proposed algorithm's scalability and effectiveness in optimizing computational processes in FC environments.

VI. Conclusion

In this paper, we proposed a MOOA aimed at optimizing resource allocation and reallocation in FC environments for

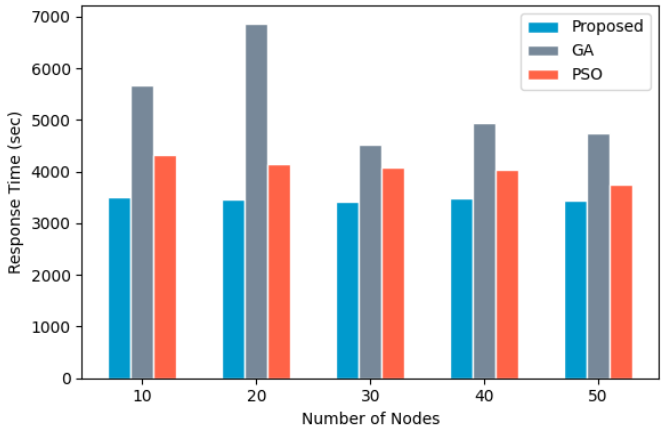


Fig. 7. Response time of 200 tasks with different FNs.

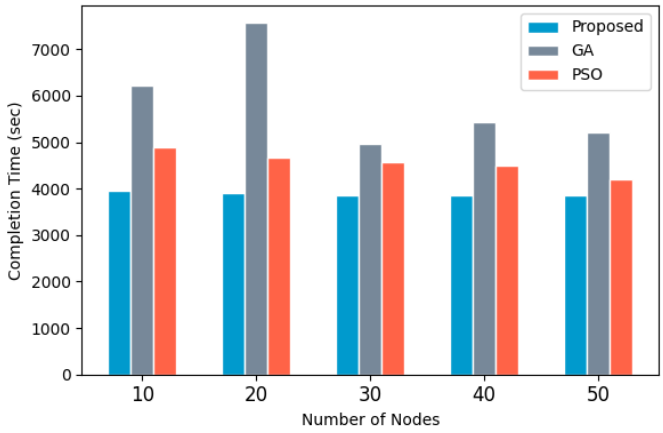


Fig. 8. Completion time of 200 tasks with different FNs.

IoT tasks. By optimizing multiple objectives such as completion time, response time, and execution time, MOOA provides an efficient solution for managing fog resources in dynamic and resource constrained environments. The simulation results demonstrated the effectiveness of the proposed algorithm in handling tasks efficiently in FC environments. The algorithm achieved a notable reduction in execution time, with an average improvement of 12.45% over PSO and 22.97% over GA. Moreover, MOOA consistently outperformed GA and PSO in terms of response time, achieving average improvements of 32.57% and 24.45%, respectively. Analysis of completion time further demonstrated the algorithm's superior performance, with an average reduction of 44.39% compared to GA and 33.23% compared to PSO. These results underline the robustness and scalability of MOOA, highlighting its effectiveness in optimizing task allocation and computational processes in FC environments. Future work will focus on extending this approach to incorporate energy efficiency, fault tolerance, and multi-objective optimization to address diverse QoS requirements in dynamic FC environments.

REFERENCES

[1] T. Arpitha, D. Chouhan, and J. Shreyas, "An efficient aco-inspired multi-path routing for source location privacy with dynamic phantom node selection scheme in iot environments," *Soft Computing*, pp. 1–18, 2024.

[2] N. Srinidhi, E. Nagarjun, and S. Dilip Kumar, "Hybrid algorithm for efficient node and path in opportunistic iot network," *Journal of Information Technology Management*, vol. 13, pp. 68–91, 2021.

[3] N. Srinidhi, E. Nagarjun, J. Shreyas, S. Dilip Kumar, and D. Chouhan, "Ensuring fault tolerant connectivity in iot networks," in *Computer Communication, Networking and IoT: Proceedings of ICICC 2020*. Springer, 2021, pp. 391–400.

[4] H. K. Apat, B. Sahoo, V. Goswami, and R. K. Barik, "A hybrid meta-heuristic algorithm for multi-objective iot service placement in fog computing environments," *Decision Analytics Journal*, vol. 10, p. 100379, 2024.

[5] M. Zolghadri, P. Asghari, S. E. Dashti, and A. Hedayati, "Resource allocation in fog–cloud environments: State of the art," *Journal of Network and Computer Applications*, vol. 227, p. 103891, 2024.

[6] F. U. Khan, I. A. Shah, S. Jan, S. Ahmad, and T. Whangbo, "Machine learning-based resource management in fog computing: A systematic literature review," *Sensors*, vol. 25, no. 3, 2025. [Online]. Available: https://www.mdpi.com/1424-8220/25/3/687

[7] E. Nagarjun, D. Chouhan, I. Zabiulla, and S. D. Kumar, "Efficient resource provisioning in fog computing using agent-based contract net protocol: A smart healthcare case study," in *2024 First International Conference on Software, Systems and Information Technology (SSITCON)*. IEEE, 2024, pp. 1–6.

[8] A. Bandopadhyay, S. Swain, R. Singh, P. Sarkar, S. Bhattacharyya, and L. Mrsic, "Game-theoretic resource allocation and dynamic pricing mechanism in fog computing," *IEEE access*, 2024.

[9] D. Zhao, Q. Zou, and M. Boshkani Zadeh, "A qos-aware iot service placement mechanism in fog computing based on open-source development model," *Journal of Grid Computing*, vol. 20, no. 2, p. 12, 2022.

[10] M. Hussain, S. Nabi, and M. Hussain, "Rapts: resource aware prioritized task scheduling technique in heterogeneous fog computing environment," *Cluster Computing*, pp. 1–25, 2024.

[11] J. U. Arshed and M. Ahmed, "Race: resource aware cost-efficient scheduler for cloud fog environment," *IEEE Access*, vol. 9, pp. 65 688–65 701, 2021.

[12] Z. A. Khan and I. A. Aziz, "Ripple-induced whale optimization algorithm for independent tasks scheduling on fog computing," *IEEE Access*, 2024.

[13] S.-E. Chafi, Y. Balboul, M. Fattah, S. Mazer, and M. El Bekkali, "Novel pso-based algorithm for workflow time and energy optimization in a heterogeneous fog computing environment," *IEEE Access*, 2024.

[14] I. Mokni and S. Yassa, "A multi-objective approach for optimizing iot applications offloading in fog–cloud environments with nsga-ii," *The Journal of Supercomputing*, pp. 1–39, 2024.

[15] H. Raissouli, S. B. Belhaouari, and A. A. B. Ariffin, "Rwp-nsga ii: Reinforcement weighted probabilistic nsga ii for workload allocation in fog and internet of things environment," *International Journal of Distributed Sensor Networks*, vol. 2024, no. 1, p. 7645953, 2024.

[16] M. Dehghani and P. Trojovskỳ, "Osprey optimization algorithm: A new bio-inspired metaheuristic algorithm for solving engineering optimization problems," *Frontiers in Mechanical Engineering*, vol. 8, p. 1126450, 2023.

[17] F. Wei, X. Shi, and Y. Feng, "Improved osprey optimization algorithm based on two-color complementary mechanism for global optimization and engineering problems," *Biomimetics*, vol. 9, no. 8, p. 486, 2024.

[18] R. Mahmud, S. Pallewatta, M. Goudarzi, and R. Buyya, "Ifogsim2: An extended ifogsim simulator for mobility, clustering, and microservice management in edge and fog computing environments," *Journal of Systems and Software*, vol. 190, p. 111351, 2022.

[19] K. H. K. Reddy, A. K. Luhach, B. Pradhan, J. K. Dash, and D. S. Roy, "A genetic algorithm for energy efficient fog layer resource management in context-aware smart cities," *Sustainable Cities and Society*, vol. 63, p. 102428, 2020.

[20] I. M. Jabour and H. Al-Libawy, "An optimized approach for efficient-power and low-latency fog environment based on the pso algorithm," in *2021 2nd Information Technology To Enhance e-learning and Other Application (IT-ELA)*. IEEE, 2021, pp. 52–57.

Leveraging Deep Semantics for Sparse Recommender Systems (LDS-SRS)

Adel Alkhalil
Department of Software Engineering,
College of Computer Science and Engineering,
University of Ha'il, Ha'il, 81481, Saudi Arabia

***Abstract*—RS (Recommender Systems) provide personalized suggestions to the user(s) by filtering through vast amounts of similar data, including media content, e-commerce platforms, and social networks. Traditional recommendation system (RS) methods encounter significant challenges. Collaborative Filtering (CF) is hindered by the lack of sufficient user-product engagement data, while CBF (Content Based Filtering) depends extensively on feature extraction techniques in order to describe the items, which requires an understanding of both content contextual and semantic relevance of the information. To address the sparsity issue, various matrix factorization methods have been developed, often incorporating pre-processed auxiliary information. However, existing feature extraction techniques generally fail to capture both the semantic richness and topic-level insights of textual data. This paper introduces a novel hybrid recommendation system called Topic-Driven Semantic Hybridization for Sparse Recommender Systems (LDS-SRS). The model leverages the semantic features from item descriptions and incorporates topic-specific data to effectively tackle the challenges posed by data sparsity. By extracting embeddings that capture the deep semantics of textual content—such as reviews, summaries, comments, and narratives—and embedding them into Probabilistic Matrix Factorization (PMF), the framework significantly alleviates data sparsity. The LDS-SRS framework is also computationally efficient, offering low deployment time and complexity. Experimental evaluations conducted on publicly available datasets, such as AIV (Amazon Instant Video) and Movielens (1 Million & 10 Million), demonstrate the exceptional ability of the method to handle sparse user-to-item ratings, outperforming existing leading methods. The proposed system effectively addresses data sparsity by integrating embeddings that encapsulate the deep textual semantics content, including summaries, comment(s), and narratives, within PMF (Probabilistic Matrix Factorization). The LDS-SRS framework is also highly efficient, characterized by minimal deployment time and low computational complexity. Experimental evaluations conducted on publicly available MovieLens (1 Million and 10 Million) and AIV (Amazon Instant Video) benchmark datasets demonstrate the framework's exceptional ability to handle sparse user-item ratings, surpassing existing advanced methods.**

***Keywords*—LDA-2-Vec technique; content representation; topic-based modeling; probabilistic matrix decomposition**

I. Introduction

RSs are integral to e-commerce, offering tailored product suggestions for various items, including movies, books, clothing, and news. In recent years, their applications have expanded to areas such as social media platforms, websites, and articles. Leading Fortune 500 companies like Facebook, Netflix, and eBay have created proprietary RSs to predict and cater to customer preferences. The success of these organizations in the business market is significantly influenced by the effectiveness of their RSs. For instance, Netflix suggests movies based on individual user preferences, while Amazon and eBay recommend related products to customers, and social media platforms display relevant pages and advertisements to users. As a result, it is evident that recommendation systems play a crucial part in driving the financial growth of these companies [1], [2].

The principles underlying RSs are broadly categorized into CF, CB, and Hybrid Recommendation Methods [3], [4] combine multiple filtering techniques, such as Collaborative Filtering (CF), to derive user-item ratings based on historical metadata, analysing individual preferences and behaviours. It recommends new products by identifying similarities between users based on their past preferences, without relying on the specific details of the items' content [5], [6]. Conversely, Content-Based (CB) filtering analyzes item descriptions, utilizing their attributes and features to match them with user profiles. By analyzing items liked by the user, CB generates a similarity matrix to recommend the most relevant new items. This method heavily relies on item descriptions and user profiles for making accurate recommendations. This approach depends largely on item descriptions and user profiles to provide precise recommendations [7], [8]. Hybrid filtering, which integrates both CB and CF methods, utilizes user preferences along with content semantics to improve the effectiveness of recommender systems.

Collaborative Filtering (CF) is widely considered a powerful model for developing recommendation systems, as it predicts ratings by examining user-item rating matrices [1]. CF can be categorized into two main approaches: Memory Based and Model Based techniques [9], [10]. Memory-based methods concentrate on identifying similarities among users or products to find potential neighbors, using these similarities to predict ratings. However, this approach is often hindered by issues like limited data availability and the cold start problem [3], [11]. In contrast, Model-Based Collaborative Filtering utilizes trained techniques, including decision trees, clustering methods [12], [13], Bayesian models [15], latent factor models [14], and dimensionality reduction methods [16]. While these approaches can be highly effective, their implementation and upkeep demand substantial effort and computational resources, particularly due to the necessity of frequent parameter optimization [17]. Among these, Matrix Factorization (MF) stands out for its balance of scalability and precision [11]. When working with sparse or large-scale datasets, Matrix Factorisation (MF) may perform poorly, perhaps producing less accurate

predictions. MF breaks down the user product engagement grid to latent features that uncover underlying rating patterns [4].

The rising volume of online users interaction and available products has exacerbated sparse ratings problem. For Collaborative Filtering (CF)-based recommendation systems to function effectively, comprehensive user rating histories are crucial. However, upon the registration of new users, the absence of historical data leads to less accurate recommendations. Consequently, the sparsity of ratings and the lack of detailed semantics in item descriptions hinder the accuracy of predictions and recommendations [1], [18]. Trust-based collaborative filtering methods have been explored to mitigate these issues by leveraging user trust relationships to enhance recommendations [22]. To address these challenges, several methods have been introduced that integrate both user feedback data and supplementary details, including user preferences (e.g., likes, interests, and social media activity) and item specifics (e.g. reviews, summaries, and plots) [19] - [20].

To address the limitations of traditional CF and hybrid methods in handling sparse recommendation scenarios, we choose to propose the LDS-SRS framework. This method enhances PMF by embedding deep semantic and topic-driven representations, ensuring more accurate recommendations even in sparse datasets. Unlike neural network-based models, which demand high computational resources, LDS-SRS efficiently integrates on textual knowledge while maintaining low computational complexity.

This research seeks to address the sparsity problem by extracting the rich semantics from textual item descriptions. The proposed method utilizes an embedding model to analyze auxiliary item data, capturing semantic information enhanced with topic-related details. These embeddings are integrated into the document latent factors of Probabilistic Matrix Factorization (PMF). By employing PMF as the CF technique, experimental outcomes demonstrate the model's effectiveness. The main contributions of the research study include:

- The introduced framework, LDS-SRS, integrates the detailed semantics of textuaThe main contributions of the research study include:l item data with topic-specific insights.
- The derived embeddings serve as input for memory-based collaborative recommendation methods to enhance the precision of predicted ratings.
- A comparative analysis demonstrates that LDS-SRS outperforms existing paradigms on MovieLens (1 Million and 10 Million) and AIV (Amazon Instant Video) benchmark datasets

The structure of this paper is outlined as follows: Section II reviews the relevant literature, while Section III details the proposed model for the recommendation system. Section IV provides a comprehensive analysis of the results and performance metrics, and Section V wraps up the study with concluding remarks.

II. Literature Review

The functionality of recommendation systems (RS) is significantly affected by the cold start issue and the scarcity of user-item rating data. To enhance the accuracy of rating predictions and the quality of recommendations, research has focused on CB(Content Based) and CF(Collaborative Filtering) techniques. One major challenge, the cold start problem, is addressed by applying text analysis techniques to extract valuable information from both user and item data, allowing the system to effectively incorporate new users or items. Additionally, the CB approach [21] helps mitigate sparsity by utilizing item-specific features, making it easier to manage new items.

These techniques use low-rank factorization to approximate the user item interaction grid, improving the prediction of products. Matrix Factorization techniques break down the user item grid into lower-dimensional embeddings for both users & products. Additionally, some systems enhance collaborative recommenders by incorporating trust-based information [46]. The Topic-MF model [23] integrates biased matrix factorization techniques to address sparsity by combining rating data with topic-related information derived from user reviews. Since text reviews provide more semantic depth than ratings, they offer richer insights into user and item characteristics. Functional Matrix Factorization (FMF) uses interview-based tools to build user profiles [24], while the Latent Dirichlet Allocation (LDA) model is employed to extract topic-specific details from item descriptions.

ALS (Alternating Least Squares) [25] and SGD (Stochastic Gradient Descent) [4] are well-established optimization techniques frequently used for training Matrix Factorization models. In [26], two alternative multiplicative update methods for Non-negative Matrix Factorization (NMF) were proposed, each with unique update rules for the multiplicative factors. The method presented in [27] integrates Weighted Singular Value Decomposition (WSVD) with a linear regression model, where each latent factor is assigned a weight parameter. Furthermore, [28] introduced a cosine similarity-based Matrix Factorization (CosMF) model, aimed at addressing sparsity issues without the need for additional data processing. However, sparse rating data often hampers the effective training of user and product vectors because of the lack of supplementary data. This model substitutes dot products with cosine similarity for users and products with limited data, helping to mitigate the adverse effects of missing auxiliary data.

Probabilistic Matrix Factorization (PMF) [29] outperforms SVD in recommendation tasks. Recently, advanced generalizations of PMF, such as Generalized PMF [30], ConvMF [32], and Bayesian PMF [31], have been introduced. While the CB model [21] addresses sparsity by utilizing item features, it struggles with effectively accommodating new users. Hybrid RS models were introduced to integrate CF with user or product content information [20], [32], [33]. For example, in [34], item features were generated using a Stacked Denoising AutoEncoder (SDAE) trained on online item descriptions. These features were integrated into the timeSVD++ CF model, which uses a weighted bag-of-words approach but fails to capture semantic word similarities. Approaches for document representation, including Latent Dirichlet Allocation (LDA) and Stacked Denoising Autoencoders (SDAE), have been utilized to extract content features from supplementary sources like reviews, synopses, or abstracts [35], [36], [20], [37]. Wang et al. combined Probabilistic Matrix Factorization (PMF) and SDAE to enhance the accuracy of latent models for rating

predictions [20]. In contrast, Collaborative Deep Learning (CDL) adopted a more straightforward collaborative filtering approach, concentrating on generating top-N recommendations.

Convolutional Neural Networks (CNNs) are widely applied in Digital Image Processing (DIP) and Computer Vision (CV) tasks [38]. Additionally, CNNs have shown considerable promise in Natural Language Processing [39], [40], [41] and information retrieval applications. CNNs effectively capture contextual features from images or textual descriptions by employing subsampling, shared weights, and receptive fields [38]. CNNs can incorporate word embeddings, such as Word2Vec, to extract contextual word features. ConvMF [32] integrates CNN architecture, specifically designed for NLP, with PMF. Similarly, [42] utilized a pre-trained embedding model with CNNs, which were then integrated into PMF. While these models capture the contextual characteristics of item textual data, they often fail to incorporate deep semantics. Furthermore, these models may struggle with negative values in latent representations, potentially leading to the degradation of information during rating prediction [49]. In [51], a hybrid recommendation system was introduced that utilizes content embeddings to predict scores for cold start products. The $HRS-CE$ model generates word embeddings from item descriptions and uses them to build user profiles for recommending similar items. [52] enhanced Non-Negative Matrix Factorization (NMF) by incorporating contextual item information through a sophisticated embedding model that captures semantic meaning as well as additional contextual data [53], [54].

In contrast to traditional recommendation system (RS) models, the proposed LDS-SRS model adopts a hybrid approach that captures comprehensive content embeddings of products and integrates them with CF (collaborative filtering) methods for predicting missing ratings. While extracting content features, both the semantic and topic-related elements of item descriptions are recognized and incorporated into the hidden representations of the PMF(Probabilistic Matrix Factorization) model. The LDS-SRS model does not rely on neural networks for natural language processing tasks, nor does it require the iterative updating of item embeddings. As a result, it offers greater computational, temporal, and memory efficiency compared to other RS algorithms.

III. Proposed Methodology

A recommender system is presented that combines deep semantic insights with topic-specific information to predict item ratings. The model combines "local" features extracted from dense vectors with "global" representations based on topic information, aligning them within a unified space. These contextual embeddings are subsequently utilized to initialize the item representation factors within the framework of PMF.

This section outlines the both the mathematical artifacts/formulation and system's design. Fig. 1 shows the suggested paradigm architecture. Experiments are conducted using the MovieLens dataset, where each film is treated as a product, with its corresponding plot acting as the product description.

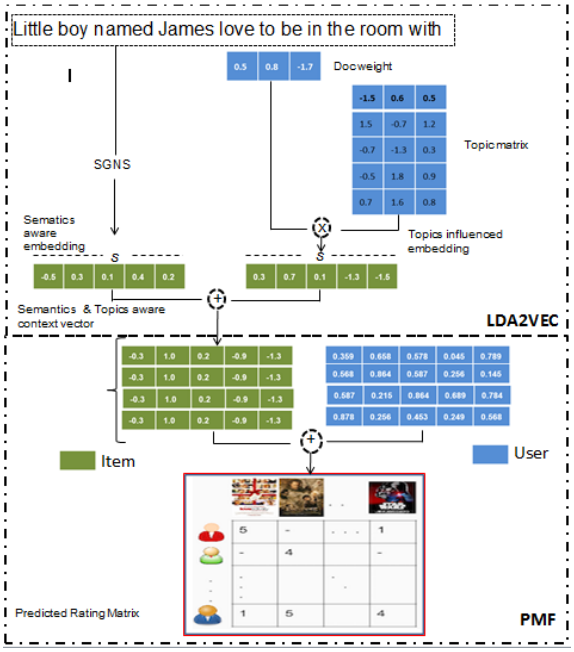


Fig. 1. Proposed framework architecture.

A. Mathematical Modeling

Let $U = [u_1, u_2, u_3, \ldots, u_m]$ denote the set of users, and $I = [i_1, i_2, i_3, \ldots, i_n]$ denote the set of products. The user product rating grid is represented as a two-dimensional array, as follows:

$$R_{ui} \in [0.5, 1, 2, 3, 4, 5]^{m \times n}$$

Here m: Total Users No.
n: Items No.
R_{ui}: Ratings of the Item(i) by the User's(u).

The proposed LDS-SRS approach aims to estimate ratings $\hat{R}_{ui}$ for items (i) that are unrated in the rating matrix, based on the existing ratings. For collaborative filtering, probabilistic matrix factorization (PMF) is applied. PMF decomposes the matrix $R^{m \times n}$ into three separate components: $U \in R^{s \times m}$, $I \in R^{s \times n}$, and $\Omega \in R^{s \times s}$. Thus,

$$RU \times \Omega \times I \tag{1}$$

In this scenario, $R^{m \times n}$ refers to the $(m \times n)$ matrix of user-item ratings, where U represents the latent factors associated with users and I represents those associated with items. Additionally, Ω is a diagonal matrix that represents the hidden patterns within R, while s represents the count of latent factors contained in R.

In order to address the problem of data sparsity, LDS-SRS integrates embeddings containing semantic and topic-related information from text into the probabilistic matrix factorization (PMF) model. The aim is to learn latent representations for users and products, $U \in R^{s \times m}$ & $I \in R^{s \times n}$, in a way that their multiplication $(U^T I)$ closely matches the user-item rating grid R [29]. These latent models are refined by minimizing a

cost function ℓ, which includes the squared difference between actual and predicted ratings, along with regularization terms.

$$\ell = \sum_{u}^{m}\sum_{i}^{n} A_{ui}(r_{ui} - u_u^T i_i)^2 + \alpha \sum_{u}^{m} \|u_u\|^2 + \beta \sum_{i}^{n} \|i_i\|^2 \quad (2)$$

In this context, A_{ui} is defined as a function that returns 1 if user *u* has rated item *i* and 0 otherwise. The performance of the model is evaluated by computing the Root Mean Square Error (RMSE) on the test dataset. A Gaussian noise-based probabilistic linear framework is utilized. The probability distribution based on the observed ratings is expressed as follows:

$$\rho(R|U, I, \sigma^2) = \prod_{u}^{m}\prod_{i}^{n} \eta(r_{ui}|u_u^T i_i, \sigma^2)^{A_{ij}} \quad (3)$$

In this context, $\eta(x|\mu, \sigma^2)$ denotes the probability density function of a Gaussian distribution characterized by a mean value of μ and a variance of σ^2. A spherical Gaussian prior, centered at zero, is applied to the latent vector of the user, as per the equation below:

$$\rho(U|0, \sigma_U^2) = \prod_{u}^{m} \eta(u_u|0, \sigma_U^2 A)$$

In contrast to the user latent method, our approach suggests deriving the probabilistic representation of the item's latent vector from a features paradigm based on the D_i (product description), which includes a Gaussian noise component represented by a gamma variable.

$$i_i = \textit{Embeddings}(D_i) + \gamma_i$$

$$\gamma_i \sim \eta(0, \sigma_I^2 A)$$

The item latent model's distribution is then expressed as:

$$\rho(I|D_i, \sigma_I^2) = \prod_{i}^{n} \eta(i_i|\textit{Embeddings}(D_i), \sigma_I^2 A)$$

Here, D_i represents the description document linked to item *(i)*. The context vector derived from the dense embedding characterizes the Gaussian distribution for the item, with the variance being determined by the Gaussian noise. This variance is essential for initializing the item latent representation in the PMF framework.

In the proposed paradigm, an embedding is created specifically for movie plots. This embedding effectively captures the local semantic features of each plot, while also incorporating broader, topic-related information. Let G represent the movie plot (or item description), consisting of l words, denoted as $w_1, w_2, w_3, \ldots, w_l$.

$$G \xleftarrow[\text{Plot Description}]{} \{w_1, w_2, w_3, ...w_l\}$$

A collection C is constructed from the textual descriptions of movie plots, encompassing all the movies (denoted as G), and can be expressed as:

$$C \xleftarrow[\text{Corpus Generation}]{} \{G_1, G_2, G_3, ...G_n\}$$

In which $C \in \{w_1, w_2, w_3, ...w_l, w_{l+1}, ...w_t\}$, where t denotes the total word count in the entire corpus C.

The Word2Vec (w2v) model employs the Skipgram Negative Sampling (SGNS) method on the corpus C to produce dense word embeddings with s dimensions. This paradigm captures local semantic relationships within a plot by computing the probabilistic weights of words in relation to a reference word. It *locally* predicts the surrounding words based on the given pivot word in the context of a movie plot.

$$P(P_{TW}|P_{PW})$$

In this context, P_{TW} represents the likelihood of the intended word, while P_{PW} denotes the likelihood of the pivot word within the model. The Word2Vec model generates enriched real-valued embeddings that capture both semantic and syntactic relationships within the plots. These compact representations offer increased flexibility but are less interpretable. In contrast, Latent Dirichlet Allocation (LDA) is applied to the dataset to extract topic-specific information from the text. LDA generates a sparse probability vector, providing better interpretability and a clearer understanding of the data. Operating on a *global* scale, LDA constructs a unified plot vector that summarizes topic-related features, which are later used for word prediction within the plot.

$$P(P_{TW}|P_{Topics})$$

Here, P_{Topics} represents the probability of the topic information. The enhanced embedding model combines the dense representations produced by Word2Vec (w2v) with the interpretability provided by Latent Dirichlet Allocation (LDA). This method combines the w2v embeddings of a plot with sparse vectors produced by LDA to approximate a multinomial distribution across underlying word categories. The resulting representation is subsequently passed through a probabilistic model to assign distinct topics to a specific set of movies or items.

$$P(P_{TW}|P_{PW} + P_{Topics}) \quad (4)$$

Eq. 5 is composed of two components. The first component, $\pounds_{SGNS}$, adheres to the standard word2vec (w2v) methodology, aiming to maximize the likelihood of target words and negative samples in relation to the surrounding word representation. The computation is performed using the context vector C_v, the target word vector P_{TW}, the pivot word vector P_{PW}, and the vector for negative (irrelevant) words NS_v. The second component, $\pounds_{LDA}$, introduces a Dirichlet likelihood term that is linked to the document weights. The complete structure of the suggested recommendation model is depicted in Fig. 2. The overall objective function is the combination of the SGNS loss component, augmented by the Dirichlet distribution term that is applied to the document weights.

$$\pounds = \pounds_{SGNS} + \pounds_{LDA} \quad (5)$$

$$\pounds_{SGNS} = \log(C_v|P_{TW}) + \log(-C_v|NS_v) \quad (6)$$

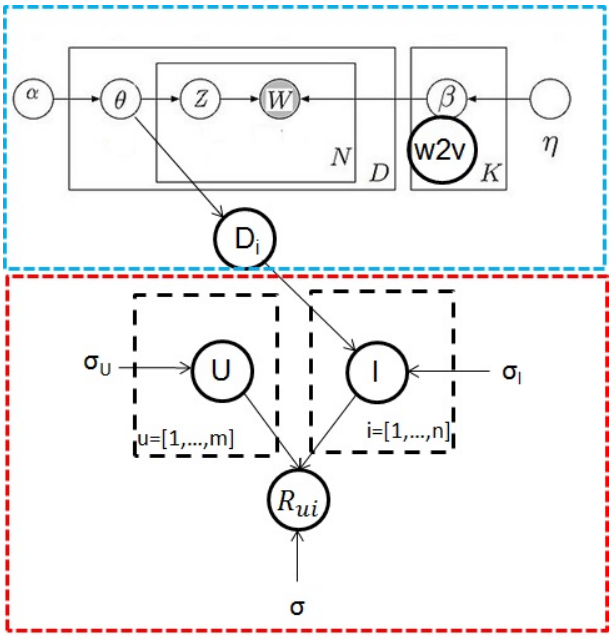


Fig. 2. Proposed paradigm with LDA2VEC embedded in W2V for generating ratings.

Ultimately, by following the outlined procedure, the embedding model processes movie plots and generates a corresponding latent vector for each one. This vector encapsulates the deep semantic context of the plot, while also capturing topic-specific information unique to each movie.

$$LV_i = \textit{Embeddings}(D_i)$$

Here, D_i refers to the description of the *i*th film, while LV_i denotes the latent vector associated with the *i*th film.

B. Optimization

In order to optimize the latent representations for users and products, a posteriori estimation is carried out as follows:

$$\max_{U,I} \rho(U, I|R, \boldsymbol{D}, \sigma^2, \sigma_U^2, \sigma_I^2) = \max_{U,I}[\rho(R|U, I, \sigma^2) \rho(U|\sigma_U^2)\rho(I|D, \sigma_I^2)] \tag{7}$$

By applying the negative logarithm to Equation 7, we obtain:

$$\ln(U, I) = \sum_{u=1}^{m}\sum_{i=1}^{n} \frac{A_{ui}}{2}(r_{ui} - u_u^T i_i)^2 + \frac{\alpha}{2}\sum_{u=1}^{m}\|u_u\|^2 + \frac{\beta}{2}\sum_{i=1}^{n}\|i_i - \textit{Embeddings}(D_i)\|^2 \tag{8}$$

Gradient descent is employed to update the latent representations for users (U) & items (I), iteratively optimizing each model while holding the other variables constant. The main objective is to identify the local minima of the OF (Objective Function) by differentiating Eq. 8 with respect to u_u and i_i in a closed-form expression. Consequently, the latent representations for both the user and the item are computed as follows:

$$u_u = \frac{IR_u}{II_u I^T + \alpha I_s} \tag{9}$$

$$i_i = \frac{UR_i + \beta \textit{Embeddings}(D_i)}{UI_i U^T + \beta I_s} \tag{10}$$

In this context, I_u refers to a diagonal matrix with diagonal elements $I_{ui}, i = 1, \ldots, n$, while R_u represents a vector containing $(r_{ui})_{i=1}^n$ for the user u. Similarly, for the item *(i)*, I_i and R_i are specified in the same way as I_u and R_u.

IV. Outcomes and Performance Assessment

This section provides an evaluation and discussion of the proposed model's performance. It starts with an in-depth description of the dataset and its processing, followed by a thorough explanation of the experimental setup. The results are then presented, analyzed, and visually displayed, with comparisons drawn to leading methods in the field.

A. Dataset Pre-processing

The Movie-Lens 1 Million corpus [43] contains 1,000,209 evaluations $R_{ui} \in \{0.5, 1, 2, 3, 4, 5\}$ for approximately 3,900 movies, rated by 6,040 users. The MovieLens 10 Million dataset includes 10,000,054 evaluations across 10,681 movies (items), contributed by 71,567 users. Additionally, the Amazon Instant Video (AIV) dataset features evaluations for 15,149 items provided by 29,757 users, with scores $R_{ui} \in \{1, 2, 3, 4, 5\}$ ranging from 1 to 5.

Table I provides a summary of the statistics for the datasets used.

TABLE I. Dataset Characteristics

Dataset Name	Total Ratings	User Count	Item Count	Rating Scale	Density (%)
AIV	1,351,88	29,757	15,149	[1 - 5]	0.030
1M	1,000,209	6,040	3,900	[0.5 - 5]	1.431
10M	1,000,005,4	71,567	10,681	[1 - 5]	4.641

To predict ratings for items, additional information such as movie plots is required. However, not all movies in the MovieLens open-source datasets have corresponding plots. To address this, the original datasets were expanded to ensure that the experiments could be carried out effectively. Movie plots were obtained using the OMDB API[1]. A Python script was written to query the OMDB database using each movie's ID, from which the corresponding plot was automatically retrieved. The gathered data underwent several preprocessing steps, including text cleaning and removal of stopwords. Stopwords, which are frequent yet low-information words (e.g. "a", "an", "the", "that"), were eliminated to enhance the quality of text analysis. The following preprocessing steps were performed on the movie plot data: 1) Plot lengths were restricted to a maximum of 200 words. 2) The Word2Vec (w2v) model was

[1]http://www.omdbapi.com

utilized with a sliding window of size 1 to capture contextual relationships among words in the dataset. 3) Subscripts were removed from the text as a precautionary measure. 4) The tokens were generated for the text corpus. 5) LDA2VEC methods were implemented and applied to the tokens of the corpus text using a pre-trained w2v model based on the same dataset [44]. 6) The vector dimensionality was configured to 100. Furthermore, movies without plot summaries were excluded from the dataset to maintain accuracy and reliability in the results.

Harnessing the detailed semantic insights encapsulated in the latent vectors produced by our embedding model, we utilized this information to initialize the item latent factors within the PMF. This step plays a crucial role in integrating the LDA2VEC model with the PMF, enabling an effective fusion of item descriptions and ratings data.

B. Evaluation Metrics

The LDS-SRS model's performance is assessed by employing a ten fold cross validation approach. The overall prediction metric is calculated by averaging the results across all 10 iterations. The training dataset contains a User-Product grid that includes known feedback, while the testing dataset consists of user-product pairs for which the ratings must be predicted. These predictions are generated by multiplying the matrices *U* and *I*. The RMSE, a commonly employed cost function in traditional rating prediction models, is computed as follows:

$$RMSE = \sqrt{\frac{\sum_{u,i}^{U,I}(R_{ui} - \hat{R})^2}{Numberof Ratings}}$$

Along with the previously discussed metrics, Precision and Recall were also employed to assess the effectiveness of the proposed paradigm. Precision represents the fraction of accurate recommendations among the overall recommendations generated, while Recall@K signifies the fraction of correct recommendations among all relevant items. To evaluate Precision and Recall, items were categorized into a pair of groups based on their assigned ratings: non-relevant (ratings 1-3) and pertinent (ratings 4-5). The items in the datasets were then categorized into those that were predicted by the model and those that were not. Precision and Recall for this approach are calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

Here,

TP: True +ive (An item is accurately identified as relevant)

FP: False +ive (An item is incorrectly identified as relevant when it is not)

FN: False -ive (An item selected as negative but is actually positive)

The RMSE of the suggested model demonstrates a decrease in inaccuracies on the test set. RMSE is a widely used metric in recommendation systems as it quantifies the deviation of predicted ratings from actual user ratings, making it a reliable measure of prediction accuracy. Additionally, Precision and Recall are employed to assess the model's ability to recommend relevant items, with Precision reflecting the proportion of correctly recommended items and Recall indicating the proportion of relevant items successfully retrieved. These validation measures are essential in evaluating the effectiveness of LDS-SRS in mitigating sparsity while maintaining high-quality recommendations.

C. Outcomes & Evaluation

The effectiveness of the LDS-SRS model was assessed using datasets from *MovieLens 1M*, *10M*, and AIV. Multiple experiments were carried out to examine the model's convergence, with RMSE being the main metric for validation in each case. The model's performance was evaluated under various conditions, where *s* represents the size of the user-product latent factors. Table II shows the RMSE values for the suggested paradigm with $K = 50/100$ for the both the Movielens and AIV datasets. Fig. 3 demonstrates the model's convergence throughout the iterations for $K = 50/100$. The smallest value of RMSE was obtained when the value of $K = 100$, suggesting that the model performs better in capturing detailed and precise information for both user and item embeddings. For other values of K, the RMSE increased, indicating a decline in the accuracy of movie representations. The plots provide a clear visualization of the convergence trends for both the training and testing datasets.

TABLE II. EFFECTIVENESS OF THE LDS-SRS FOR K@50/100

Paradigm	**Dataset (DS)**	**Assessment (RMSE)**	
		K@50	K@100
Suggested	1M	0.857	0.851
	10M	0.789	0.782
	Amazon Instant Video	1.101	1.083

Table III compares the prediction Effectiveness of the suggested paradigm for various vector sizes generated by the content embedding model. The findings show that the model achieves the best performance when the vector size is configured to 100. On the other hand, increasing the vector size beyond 100 does not enhance performance, as the model effectively captures the required semantic and topic information with a vector size of 100. Further increases in vector size result in a drop in performance, as larger vectors become sparse, leading to an uneven distribution of topic information. Fig. 4 shows the model's convergence for different content vector sizes over iterations. In addition, Fig. 5 illustrates the improvement in item recommendations, with both Precision and Recall steadily increasing. The Precision and Recall at 10 values for our paradigm are computed and compared with those of other leading models from existing literature.

Table IV presents the prediction errors of the proposed model alongside several other recommender systems [33], [20], and [32]. In [33], Wang et al. incorporated LDA into PMF, achieving RMSE values of 0.8969 on the '1M' collection, 0.8275 on the '10M collection', and 1.549 on the 'AIV' collection. In [20], SDAE was employed to learn item features within PMF, resulting in RMSE values of 0.8879 for '1M',

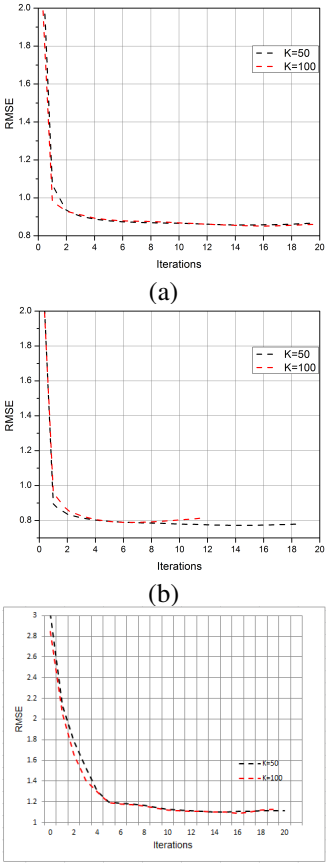


Fig. 3. LDS-SRS Performance with $K = 50$ and $K = 100$.

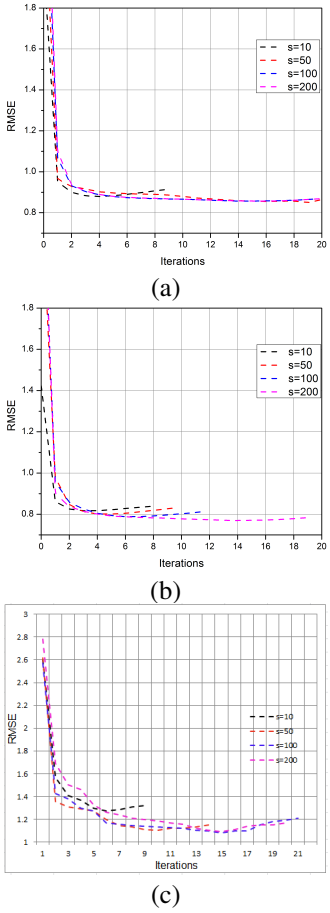


Fig. 4. LDS-SRS performance with different embedding sizes.

TABLE III. EVALUATION OF THE PROPOSED MODEL WITH VARYING VECTOR SIZES

Paradigm	Dataset (Ds)	Assessment (RMSE)			
		Size=10	Size=50	Size=100	Size=200
Suggested	1M	0.879	0.868	**0.856**	0.857
	10M	0.818	0.801	**0.782**	0.780
	Amazon Instant Video	1.12	1.119	**1.083**	1.094

0.8186 for '10M', and 1.3594 for 'AIV'. More recently, [32] combined contextual item information with PMF, enhancing RMSE to 0.853 for '1M', 0.795 for '10M', and 1.133 for 'AIV'. In [42], CNN was integrated with PMF to account for item statistics and Gaussian noise, yielding RMSE values of 0.847 for '1M', 0.784 for '10M', and 1.101 for 'AIV'. [45] introduced imputed data from similar neighbors into SVD, resulting in an RMSE of 0.850 for '1M'. [48] employed data clustering based on user/item similarity for recommender systems, achieving an RMSE of 1.0878 for '1M'. Lastly, [50] classified items and users into three distinct categories and influenced them with a Bhattacharya coefficient, obtaining an RMSE of 1.7000 for '1M'.

Table IV clearly demonstrates that the proposed model surpasses other recommender systems on the MovieLens and AIV datasets. Specifically, it achieves RMSE values of 0.846, 0.779, and 1.083 for the '1 Million', '10 Million', and 'AIV (Amazon Instant Video)' datasets, respectively, within a sparse user to item matrix, highlighting its superior performance.

Its runtime for rating predictions is significantly faster compared to other models.

TABLE IV. SUGGESTED PARADIGM ANALYSIS WITH HYBRID MODELS

Paradigm	ML-1 M	ML-10 M	AIV
CTR [33]	0.8969	0.8275	1.549
CDL [20]	0.8879	0.8186	1.3594
CMF [32]	0.853	0.795	1.133
CMF+ [32]	0.854	0.793	1.1279
RCMF [42]	0.847	0.784	1.101
RN [47]	0.863	0.807	1.105
ISVD [45]	0.850	-	-
GAGE [48]	1.0878	-	-
NCBR [50]	1.7000	-	-
Suggested	**0.846**	**0.779**	**1.083**

D. Complexity Analysis

The suggested paradigm demonstrates considerably lower computational complexity in comparison to other models. Unlike approaches like [32], which necessitate retraining the entire network and updating weights at each iteration, the proposed model operates without relying on neural networks.

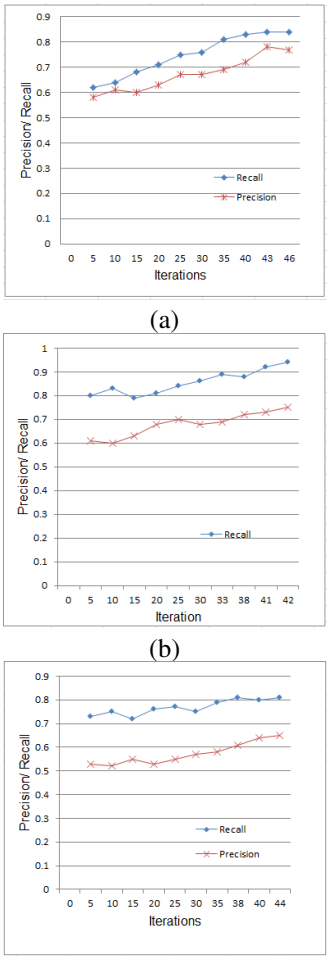


Fig. 5. Precision and recall for the proposed model.

This absence of neural network dependence leads to greater efficiency in its performance.

As shown in Table IV, our proposed LDS-SRS framework achieves a lower RMSE compared to CTR (0.8969 vs. 0.846), CDL (0.8879 vs. 0.846), and CMF (0.853 vs. 0.846) on the MovieLens 1M dataset. The key advantage of LDS-SRS lies in its ability to embed semantic and topic-driven representations within PMF, enhancing recommendation accuracy even in sparse user-item matrices. Unlike deep hybrid models that require high computational overhead, LDS-SRS efficiently integrates contextual knowledge while maintaining a lower computational cost.

In the outlined approach, movie plot vectors are computed a single time using a sophisticated content embedding model, with a computational complexity of $O(UIr + r^3)$, where $r = \min(U, I)$. These vectors are employed to set the item-specific latent factors within the PMF framework and are not subject to recalculation in subsequent iterations. The iterative updates for the latent factors U and I are performed with a complexity of $O(A^2\hat{R} + A^3U + A^3I)$, where $\hat{R}$ corresponds to the recorded ratings within the predicted feedback matrix, which are updated at every epoch. Consequently, the overall computational cost per iteration for the LDS-SRS model is $O(A^2\hat{R} + A^3U + A^3I) + UIr + r^3$, making it substantially more efficient compared to traditional recommender systems.

The proposed method exhibits outstanding efficiency with respect to time complexity. It employs a model for embedding content that integrates item semantics with topic information using LDA2VEC, which facilitates faster convergence of the collaborative filtering (CF) technique and improves overall accuracy. In contrast to neural network-based models that necessitate recalculating the content model in each iteration, this approach eliminates redundant computations, thus enhancing its efficiency.

E. Discussion

The results of this study demonstrate the effectiveness of LDS-SRS in mitigating the impact of data sparsity in recommender systems. By integrating deep semantic and topic-aware embeddings into PMF, our method enhances user-item interactions, leading to better rating predictions. Compared to existing hybrid recommendation approaches, our framework achieves higher accuracy while maintaining computational efficiency. This is particularly relevant for real-world applications where user data is often incomplete or sparse. Additionally, these findings align with prior research in matrix factorization-based recommendation (e.g. [20], [32], [42]), reinforcing the importance of leveraging textual semantics in recommendation tasks. Future work could further improve this approach by incorporating adaptive topic modeling techniques for dynamic recommendation scenarios.

V. Conclusion

This enhanced proposed approach and embedding model are incorporated into PMF to address the typical challenge of sparsity in recommender systems (RS). LDS-SRS predicts user ratings by utilizing item descriptions, with item representations generated through the embedding model that captures both semantic and topic data. These document representations are subsequently integrated into underlying factors of PMF, enhancing the accuracy of rating predictions. Our model was trained and evaluated using datasets from MovieLens and AIV, with movie plots serving as item descriptions. It can be adapted to other domains, such as e-commerce platforms like Amazon, Alibaba, and eBay, where product descriptions or user reviews can be processed using the enhanced LDA2VEC model and integrated into PMF for recommendation generation. It can be extended to other areas, including e-commerce platforms like Amazon, Alibaba, and eBay, where product descriptions or user reviews can be processed using the enhanced LDA2VEC model and integrated into PMF for recommendation generation. The integration of deep semantics and topic information significantly enhances the model's capability to provide more accurate rating predictions.

References

[1] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,"*IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734-749, Jun. 2005.

[2] P.G. Campos, F. Dez and I. Cantador, "Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols," *User Modeling and User-Adapted Interaction*, vol. 24, no. 1-2, pp. 67-119, Feb. 2005.

[3] F. Ricci, L. Rokach, and B. Shapira, "Introduction to recommender systems handbook", *In Recommender systems handbook,* New York, NY, USA: Springer, pp. 1-35, 2011.

[4] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer,* vol. 42, no. 8, Aug. 2009.

[5] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms", *In Proceedings of the 10th international conference on World Wide Web,* pp. 285-295, Apr 2001.

[6] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances in artificial intelligence.,* vol. 2009, pp. 4, Jan. 2009.

[7] J. Wang, A. P. de Vries, and M. J. T. Reinders, "Unifying user-based and item-based collaborative filtering approaches by similarity fusion," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.,* pp. 501-508, 2006.

[8] M. Balabanovi, and Y. Shoham, "Fab: content-based, collaborative recommendation", *Communications of the ACM,* vol. 40, no. 3, pp.66-72, 1997.

[9] G. Chen, F. Wang, and C. Zhang, Collaborative Filtering Using Orthogonal Nonnegative Matrix Tri-factorization, Information Processing and Management: an International Journal, 3, 368-379 (2009).

[10] K. Christidis and G. Mentzas, A topic-based recommender system for electronic marketplace platforms, Expert Systems with Applications, 11, 4370-4379 (2013).

[11] Aghdam, M.H., Analoui, M., Kabiri, P.: A novel non-negative matrix factorization method for recommender systems. Appl. Math. Inf. Sci. 9(5), 2721 (2015)

[12] T. Hofmann and J. Puzicha, Latent class models for collaborative filtering, Proceedings of the 16th international joint conference on Artificial intelligence, San Francisco, CA, USA, 688-693 (1999).

[13] L.H. Ungar and D.P. Foster, Clustering methods for collaborative filtering, Proceedings of the Workshop on Recommendation Systems, Menlo Park, CA, 1-16 (1998).

[14] J. Canny, Collaborative filtering with privacy via factor analysis, Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, 238-245 (2002).

[15] J.S. Breese, D. Heckerman, and C. Kadie, Empirical analysis of predictive algorithms for collaborative filtering, Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, San Francisco, CA, USA, 43-52 (1998).

[16] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, Eigentaste: a constant time collaborative filtering algorithm, Information Retrieval, 2, 133-151 (2001).

[17] G.R. Xue et al., Scalable collaborative filtering using cluster-based smoothing, Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, 114-121 (2005).

[18] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, , 22(1):5–53, Jan. 2004.

[19] S. Li, J. Kawale, and Y. Fu. Deep collaborative filtering via marginalized denoising auto-encoder. In , *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 811–820, New York, NY, USA, 2015. ACM

[20] H. Wang, N. Wang, and D.-Y. Yeung. Collaborative deep learning for recommender systems. *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,*, KDD '15, pages 1235–1244, New York, NY, USA, 2015. ACM.

[21] P. Lops, M. de Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends". In *F. Ricci, L. Rokach, B. Shapira, P. Kantor (Eds.), Recommender systems handbook,* pp. 73105. 2011

[22] G. Guo, J. Zhang and D. Thalmann, "Merging trust in collaborative filtering to alleviate data sparsity and cold start", *In Knowledge-Based Systems,* vol. 57, pp. 57-68, 2014.

[23] Y. Bao, H. Fang and J. Zhang, "TopicMF: Simultaneously Exploiting Ratings and Reviews for Recommendation", In*AAAI*, Vol. 14, pp. 2-8, July. 2014.

[24] K. Zhou, S. H. Yang, and H. Zha, "Functional matrix factorizations for cold-start recommendation", In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 315-324, Jul. 2011.

[25] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan, "Large-scale parallel collaborative filtering for the netflix prize", *Lecture Notes in Computer Science,* vol. 5034, pp. 337-348, Jun. 2008.

[26] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing 13 (Proc. NIPS*2000)*. MIT Press, 2001.

[27] Hung-Hsuan Chen, Weighted-SVD: Matrix Factorization with Weights on the Latent Factors, *arXiv:1710.00482*, 2017.

[28] Wen, H., Ding, G., Liu, C., Wang, J.: Matrix factorization meets cosine similarity: addressing sparsity problem in collaborative filtering recommender system. In: Chen, L., Jia, Y., Sellis, T., Liu, G. (eds.) APWeb 2014. LNCS, vol. 8709, pp. 306–317

[29] A. Mnih, and R. R. Salakhutdinov, "Probabilistic matrix factorization", In *Advances in neural information processing systems,* pp. 1257-1264, 2008.

[30] H. Shan, and A. Banerjee, "Generalized probabilistic matrix factorizations for collaborative filtering", In *Data Mining (ICDM), 2010 IEEE 10th International Conference on,* pp. 1025-1030, Dec. 2010.

[31] R. Salakhutdinov, and A. Mnih, "Bayesian probabilistic matrix factorization using Markov chain Monte Carlo", In *Proceedings of the 25th international conference on Machine learning,* pp. 880-887, Jul. 2008.

[32] Kim, D, Park, C., Oh, J., Lee, S., Yu, H.: Convolutional matrix factorization for document context-aware recommendation. In: *Proceedings of the 10th ACM Conference on Recommender Systems*, pp. 233–240. ACM (2016).

[33] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* KDD '11, pages 448–456. ACM Press, August 2011.

[34] Jian Wei, Jianhua He, Kai Chen, Yi Zhou, Zuoyin Tang, Collaborative Filtering and Deep Learning Based Recommendation System For Cold Start Items, *Expert Systems With Applications*, KDD '15, pages 1235–1244, New York, NY, USA, 2015. ACM.

[35] G. Ling, M. R. Lyu, and I. King. Ratings meet reviews, a combined approach to recommend. In *Proceedings of the 8th ACM Conference on Recommender Systems, RecSys '14*, pages 105–112, New York, NY, USA, 2014. ACM.

[36] J. McAuley and J. Leskovec. Hidden factors and hidden topics: Understanding rating dimensions with review text. *In Proceedings of the 7th ACM Conference on Recommender Systems, RecSys '13*, pages 165–172, New York, NY, USA, 2013. ACM

[37] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. *In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ' 11, pages 448–456, ACM Press, August 2011.

[38] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Intelligent Signal Processing,* pages 306–351. IEEE Press, 2001.

[39] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research (JMLR),* 12:2493–2537, Nov. 2011.

[40] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), June 2014.

[41] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Empirical Methods in Natural Language Processing (EMNLP),* pages 1746–1751, 2014.

[42] Donghyun Kim, Chanyoung Park, Jinoh Oh, and Hwanjo Yu. 2017. Deep Hybrid Recommender Systems via Exploiting Document Context and Statistics of Items. Information Sciences (2017).

[43] Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 4, Article 19 (December 2015), 19 pages.

[44] C, Moody.: 2016. Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec, arXiv:1605.02019v1 [cs.CL] 6 May 2016

[45] X. Yuan, L. Han, S. Qian, G. Xu, and H. Yan, "Singular value decomposition based recommendation using imputed data," *Knowledge-Based Systems, 2018*

[46] Farah Saleem, Naima Iltaf, Hammad Afzal and Mobeena Shahzad. "Using Trust in Collaborative Filtering for Recommendations." *In 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pp. 214-222. IEEE, 2019.

[47] H. Wang, F. Zhang, J. Wang, M. Zhao, W. Li, X. Xie and M. Guo. *RippleNet: Propagating User Preferences on the Knowledge Graph for Recommender Systems* ACM International Conference on Information and Knowledge Management October 22 to 26, 2018, Torino, Italy. ACM, New York, NY, USA, 10 pages.

[48] T. Mohammadpoura, A. M. Bidgolia, R Enayatifarb and H S Javadi "Efficient clustering in collaborative filtering recommender system: Hybrid method based on genetic algorithm and gravitational emulation local search algorithm" *Genomics, https://doi.org/10.1016/j.ygeno.2019.01.001* 2019

[49] Anwar, Fahad, Naima Iltaf, Hammad Afzal, and Haider Abbas. "A Deep Learning Framework to Predict Rating for Cold Start Item Using Item Metadata." *In 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pp. 313-319. IEEE, 2019.

[50] Bag, S., Kumar, S., Awasthi, A., and Tiwari, M. K. (2019). "A noise correction-based approach to support a recommender system in a highly sparse rating environment." *Decision Support Systems, 118, 46–57. doi:10.1016/j.dss.2019.01.001* (2019)

[51] F Anwaar, N Iltaf, H Afzal, and R Nawaz (2018). "HRS-CE: a hybrid framework to integrate content embeddings in recommender systems for cold start items" *Journal of Computational Science 2018)*. doi:10.1016/j.jocs.2018.09.008

[52] Z Khan, N Iltaf, H Afzal and H Abbas "Enriching Non-negative Matrix Factorization with Contextual Embeddings for Recommender Systems" *Neurocomputing (2019)* doi: https://doi.org/10.1016/j.neucom.2019.09.080

[53] Xiangyong Liu, Guojun Wang, and Md Zakirul Alam Bhuiyan, "The Strength of Dithering in Recommender System," *Proc. of the IEEE 14th International Symposium on Pervasive Systems, Algorithms, and Networks (IEEE I-SPAN 2017)*, Exeter, UK, Jun 23-27, 2017

[54] Abdullah Al Omar, Rabeya Bosri, Mohammad Shahriar Rahman, Nasima Begum and Md Zakirul Alam Bhuiyan, "Towards Privacy-preserving Recommender System with Blockchains," *Proc. of the 5th International Conference on Dependability in Sensor, Cloud, and Big Data Systems and Applications (DependSys 2019)*, Guangzhou, China, November 12-15, 2019

A Deep Learning Approach for Nepali Image Captioning and Speech Generation

Sagar Sharma, Samikshya Chapagain, Sachin Acharya, Sanjeeb Prasad Panday*
Department of Electronics and Computer Engineering-Pulchowk Campus-Institute of Engineering,
Tribhuvan University, Lalitpur, Nepal

***Abstract*—This article introduces a novel approach for Image-to-Speech generation that aims in converting images into textual captions along with spoken descriptions in Nepali Language using deep learning techniques. By leveraging computer vision and natural language processing, the system analyzes images, extracts features, generates human-readable captions, and produces intelligible speech output. The experimentation utilizes state-of-the-art transformer architecture for image caption generation complemented by ResNet and EfficientNet as feature extractors. BLEU score is used as an evaluation metric for generated captions. The BLEU scores obtained for BLEU-1, BLEU-2, BLEU-3, and BLEU-4 n-grams are 0.4852, 0.2952, 0.181, and 0.113, respectively. Pretrained HifiGaN(vocoder) and Tacotorn2 are used for text to speech synthesis. The proposed approach contributes to the underexplored domain of Nepali-language AI applications, aiming to improve accessibility and technological inclusivity for the Nepali-speaking population.**

***Keywords*—Image captioning; speech generation; image-to-speech generation; deep learning; BLEU score; HiFiGaN; TTS**

I. Introduction

Image caption generation is a crucial task within the realms of Computer Vision and Natural Language Processing, entailing the creation of textual descriptions corresponding to images. This process serves a variety of purposes, from aiding visually impaired individuals to indexing images and facilitating image-based search engines. It is a pivotal aspect of scene understanding in Computer Vision, demanding not only the identification of objects within an image but also their coherent expression in natural language. While traditional object detection and classification tasks exist, caption generation poses a more intricate challenge, necessitating the fusion of Computer Vision techniques for content comprehension and feature extraction with Natural Language Processing methods for sequential description generation.

Recent advancements have demonstrated the effectiveness of transformer-based models over conventional CNN-RNN architectures for image captioning. Research in languages such as Hindi and Bengali has achieved notable success using transformer networks. However, Nepali remains an underexplored language in this domain due to the scarcity of labeled datasets. This study bridges this gap by employing a deep learning-based image-to-speech generation approach, integrating image captioning with speech synthesis to provide a holistic multimodal AI system.

The paper is organized as follows: Section II reviews related works, discussing existing methodologies and their limitations. Section III introduces the proposed approach, detailing model architecture and dataset preparation. Section IV explains the methodology in detail, followed by Section V, which describes the experimental setup. Section VI presents the results, while Section VII provides a discussion. Finally, Section VIII concludes with key insights and future research directions.

*Corresponding authors.

II. Related Works

The field of image captioning has witnessed significant progress through various related works. One notable contribution is the "Show and Tell" [1] model, which introduced a framework for image captioning using a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs). This approach first encodes the image using a CNN to extract visual features and then generates a caption using an RNN. It laid the foundation for subsequent advancements in image captioning by demonstrating the feasibility of combining deep-learning models to generate captions for images.

Several studies have delved into image captioning across languages, highlighting the superiority of transformer-based models over conventional CNN-RNN architectures. Mishra et al. (2021)[2] presented a transformer-based encoder-decoder architecture for Hindi image captioning, emphasizing the drawbacks of RNN models. They obtained substantial BLEU scores, marking the efficacy of the transformer in generating accurate captions. Similarly, Ami et al. (2020) [3] explored Bengali image captioning using CNN and transformer networks, demonstrating improved performance compared to earlier models. These studies utilized pre-trained CNN models like ResNet-101, InceptionV3, and Xception, along with various datasets such as BanglaLekha and Flickr8k, showcasing enhanced BLEU and METEOR scores.

Image caption generation in various languages has seen significant progress, particularly in languages like English, Hindi, and Bengali. However, Nepali, due to limited datasets and research focus, remains an understudied language in this domain. Adhikari and Ghimire (2019)[4] introduced Nepali image captioning using encoder-decoder models, but with limited performance owing to the use of traditional CNN-RNN architectures. Addressing this gap, this study by Subedi and Bal (2022) [5] pioneers Nepali image captioning utilizing a CNN-Transformer model, leveraging the strengths of transformer networks and their effectiveness in handling NLP tasks.

Tacotron [6], which combined an encoder-decoder framework with a sequence-to-sequence model to generate high-quality speech. Another influential contribution is WaveNet [7]

which introduced a deep generative model capable of generating high-fidelity audio waveforms. Building upon Tacotron developed Tacotron 2 [6] in 2018, incorporating a WaveNet vocoder to enhance the naturalness and quality of synthesized speech. Deep Voice, a series of models introduced by Arik et al. [8] in 2017, employed deep convolutional and recurrent neural networks for speech synthesis. Transformer TTS [9] proposed in 2019, harnessed the Transformer architecture originally used for machine translation and showcased excellent results in generating natural and expressive speech.

In the landscape of image description, especially in the context of Nepali speech, a noticeable dearth of research exists. While textual image captioning has received some attention in the Nepali language, the adaptation of these methodologies to cater to speech-based image descriptions in Nepali remains largely unexplored. This gap in the literature signifies an opportunity for innovative research and development. In light of this research void, the present project assumes significance by actively addressing the under representation of Nepali speech-based image description.

The absence of sufficient research and publications in the domain of Nepali speech-based image description underscores the critical need for initiatives. By recognizing and addressing this gap in the literature, this study stands as a pivotal endeavor to advance image understanding technologies tailored to Nepali speech, promising to improve accessibility and technological inclusivity for the Nepali-speaking population.

III. Proposed Work

This article emphasizes not only in the image caption generation but extends the model for generating the fluent and natural speech for the generated textual captions in Nepali Language. For understanding the visual content of the images, various state of the art architectures of Convolutional Neural Network are used for the image captioning tasks. Architectures such as VGGNet, EfficientNet, ResNet, InceptionNet have produced remarkable results which extracts various features from the image and thus generates a feature map that captures the important representations of the image.

Further for the caption generation various RNN architectures are used because of its robustness in sequential language tasks. The use of Long Short-Term Memory and Gated Recurrent Units have been overshadowed by the transformer architecture because of the attention mechanism that enables to selectively attend the different parts of the input data, assigning varying degrees of importance to different elements.

The Pretained Text to Speech [10] serves as a crucial element in the model pipeline, providing synthesized speech outputs based on textual inputs. Specifically, Tacotron2, a state-of-the-art Text to Speech (TTS) model is employed, which has been pre-trained on a diverse dataset encompassing a wide range of linguistic features and speech patterns. By integrating a pre-trained TTS system into the framework, the ability to generate high-quality spectrograms from input text is caplitalized, thereby streamlining the process of speech synthesis.

Moreover, the utilization of a pre-trained TTS system offers several advantages, including reduced training time and resource requirements, as well as enhanced synthesis quality owing to the model's extensive training on large-scale datasets. By incorporating Tacotron2 into the research, the image-to-speech generation pipeline is benefited from cutting-edge TTS technology, thereby enabling to achieve superior performance in generating natural and intelligible speech outputs for Nepali language inputs [11] to [16].

IV. Methodology

The methodology involves a combination of image processing, feature extraction, and natural language processing. Image processing and feature extraction includes the utilization of architecture of ResNet and EfficientNet. The extracted features are further used in the transformer architecture for predicting captions in Nepali Language. BLEU score metric has been used as an evaluation metric for n-grams prediction up to 4-grams. Pretrained text to speech model has been utilized to convert the textual captions to Nepali speech.

A. Dataset Preparation

Flickr8k dataset is used in the research which is openly available in the Internet. The dataset consists of real life images specially human and animal activities and each image consisting of four or five English annotated captions. The English captions are further annotated using Google Translation API into Nepali captions.

B. Image Preprocessing

The images are normalized by resizing them into a consistent size and resolution. Augmentation techniques such as rotation, flipping and blurring is used in order to generalize the images for better feature extraction.

C. CNN Feature Extraction

Convolutional Neural Network architecture ResNet and EfficientNet are used for high level visual feature extraction. EfficientNetB7 and ResNet50 architecture is used which is fine-tuned for Flicker8k dataset using pretrained weight from the ImageNet dataset and the particular fine-tuned weights are utilized. The extracted features capture meaningful visual representations.

D. Caption Generation

The captions are tokenized and vectorized as required before training. The transformer architecture chosen so as to utilize its ability of attention mechanism effectively. The model thus learns to associate the visual features with Nepali captions during training. In sharp contrast to the orthodox single layered transformer implementation, this article introduces the utilization of fine tuned nepali embedding layers. Similarly, layered transformer model is also used for caption generation.

E. Evaluation

The performance of the trained model is evaluated using the BLEU (Bilingual Evaluation Understudy) metric. The metric evaluates the predicted caption by comparing it with the reference captions on the basis of the n-grams similarity, where the sequential groups of words such as unigrams (single words), bigrams (two word sequence), trigrams (three word sequence) and so on.

F. Speech Synthesis

The speech synthesis model incorporates Tacotron which is a text-to-speech (TTS) model that generates mel-spectrograms from input text. HiFi-GAN is a high-fidelity generative adversarial network designed for audio waveform generation. Tacotron provides accurate and expressive mel-spectrograms, while HiFi-GAN transforms them into high-fidelity speech waveforms.The combined approach allows for fine-grained control over synthesized speech and produces natural-sounding output.

The overall workflow of the proposed methodology is illustrated in Fig. 1.

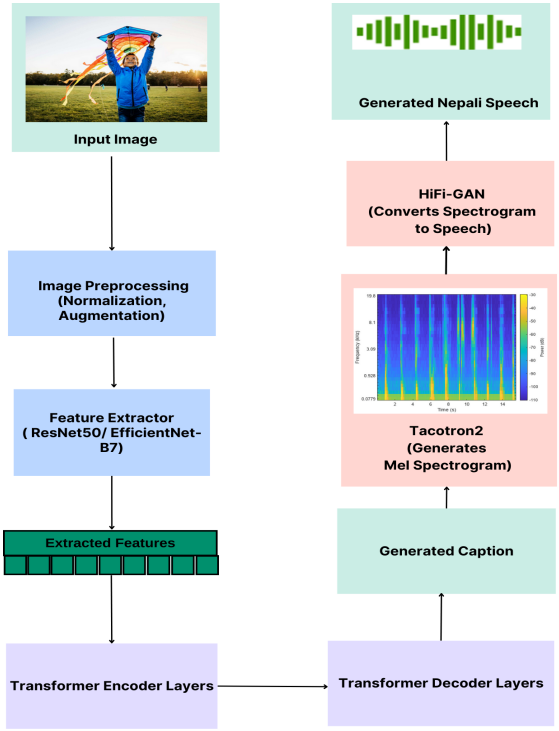


Fig. 1. Proposed methodology workflow.

V. Experimental Setup

A. Datasets

Flickr 8k dataset is utilized for training of image captioning transformer model. The dataset consist of 8k images mostly representing human, dogs and their activities. The data shuffled using random seed 42 produced best accuracy. The original dataset consists of four or five English captions for each image. The Google Language Translation API is used for the corresponding conversion into Nepali captions. Poorly translated captions is removed by manual examinations. The dataset is divided into training and validation set as shown in Table I below:

TABLE I. Split of Dataset

Dataset	Training Split	Validation Split
Flickr8k	6.4k	1.6k

B. Model Architectures

Model uses transformer based encoder decoder architecture for image caption generation. Model utilizes one of the two CNN architectures for feature extraction: ResNet50 and EfficientNetB7. The output of feature extractor is passed to the transformer encoder for further processing differentiating from the previous works which was primarily based on the single layered transformer model with the consideration of decoder part only. The layered transformer and learned embedding models are utilized as two new models for image caption generation. Both layered transformers and learned embedding uses EfficientNetB7 as the feature extractor. EfficientNetB7 uses less computation power and produces similar accuracy. For the complex and computationally expensive models like layered transformer and learned embeddings, EfficientNetB7 was implemented to reduce the overall computation requirements. Fig. 2 illustrates the architecture of single-layered transformer.

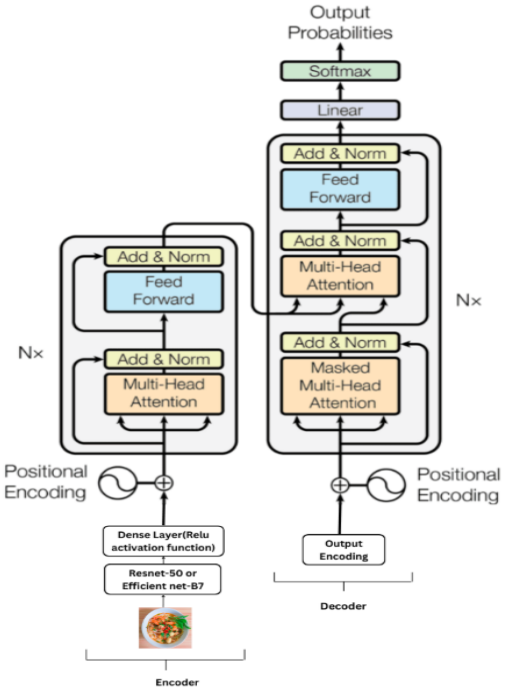


Fig. 2. Single layer transformer architecture.

1) Feature extractor: CNN architectures Resnet50 and EfficientNetB7 are utilized for extracting spatial features from images. Pretrained Imagenet weights are utilized for the extraction of image features. The augmentation techniques such as Random Flip, Random Rotation and Random Zoom are implemented for increasing the diversity of the training images. Classification layer is removed and output of second last layer is utilized as an input for the transformer encoder layer.

2) Transformer encoder: In order to take advantage of attention mechanism in the encoded image, encoder layer of transformer is utilized. The output of the feature extractor is passed to the positional embedding layer. The positional encoded features are normalized using layer normalization.

The output is then passed to the Dense layer with ReLU activation function. The extracted embeddings are passed to the multihead self attention layer. Introduction of dense layer with non linear activation function (ReLU) enables the encoder to learn complex images features and attention layer learns multiple embedding dimensions. The self attention in the encoded image features allows model to attend the important features of the image for generation of distinct captions.

3) Decoder: The transformer decoder layer is utilized for decoding the image caption. It generates caption using one word at each timestep. The previously generated words are passed as an input to the decoder. Vectorizer vectorizes each word. Vectorized word is passed to the positional encoding layer of the transformer. Masking is used to prevent the transformer to attend the future output. Masked output is then passed to the multihead self attention layer which highlights only the important part of input data while diminising the effect of the rest. The attended decoder input and the input from encoder are passed through the attention layer called encoder decoder attention. This layer builds the relationship between extracted features from encoder and the attention output from decoder. This relationship along with residual connection is finally used for predicting the Nepali caption. The whole architecture is trained using backpropagation techniques.

C. Compared Methods

The single layered transformer model with ResNet-50 as feature extractor is compared with pretrained embedding model and layered transformer model.

1) Pretrained embedding model: The pretrained embedding model utilizes the pretrained nepali word2vec text embedding. This pre-trained Word2Vec model had 300-dimensional vectors for more than 0.5 million Nepali words and phrases. About 20% of the embeddings from our dataset was missing in the pretrained embedding. Fine tuning of the word2vec model is performed using window size as 5. The model was trained for 10 epochs and fine tuned embedding were than utilized in the transformer model for image caption generation.

Every word is vectorized using text vectorizer. The vectorized word is mapped to 300 dimension word embedding. The embedding layer is trained along with other parameters of transformer decoder. Word embedding were utilized for finding relation between words in the training caption using already trained embedding from large corpus.

2) Layered transformer model: Single layer transformer is stacked in both encoder and decoder side to obtain multilayered transformer model. Three encoder layers and three decoder layers were stacked to search for intricate relationship between features. The complexity of the model goes on increasing from single layered transformer model. Single layered model is least computationally expensive while layered transformer model require highest computation as compared to other two models.

D. Evaluation Metrices

The performance of the trained model is evaluated using the BLEU (Bilingual Evaluation Understudy) metric.The metric evaluates the predicted caption by comparing it with the reference captions on the basis of the n-grams similarity, where the sequential groups of words such as unigrams (single words), bigrams(two word sequence), trigrams(three word sequence) and so on.

Four BLEU scores (BLEU-1, BLEU-2, BLEU-3, and BLEU4) are typically calculated in the context of image captioning. These scores evaluate merely matching grams of a certain order, such as single words (1-gram) or word pairs (2-gram or bigram), and so forth. BLEU score ranges from 0 to 1 where a score between 0.6 to 0.7 is considered to be the best achievable result but at the same time, a score between 0.3 to 0.4 is considered an understandably good translation and a score greater than 0.4 is considered high-quality translation.

The evaluation of the image-to-caption model's performance also incorporated the utilization of the sentence BLEU score. Unlike corpus BLEU, which uses the word level tokenization, sentence BLEU is modified to find BLEU Score using character-level tokenization to evaluate the quality of each caption independently.This sentence-level evaluation metric offers a more detailed assessment of caption quality. Considering the absence of lemmatizer for the nepali word corpus, the word level tokenization was not able to predict the generalizing ability of the model. For example, dog and dogs can be lemmatized easily for English corpus but due to complicated structure of Nepali corpus, root words could not be extracted properly. Such complication were removed using character level tokenization which does not significantly penalizes the prediction of similar words other than the root word such as dog and dogs in Nepali. Notably, the sentence BLEU method provided metrics that closely aligned with the benchmark evaluations.

E. Text-to-Speech

The implementation of a pretrained Text-to-Speech (TTS) model within the image-to-speech conversion framework was informed by a scholarly paper detailing Nepali Text-to-Speech synthesis using Tacotron2 for melspectrogram generation. The research method delineated in the paper involves preprocessing and tokenization of Nepali text, which is subsequently fed into a Tacotron2 model to generate melspectrograms. This model, initially trained on an established Nepali dataset, is further refined through fine-tuning on a proprietary dataset to enhance performance and adaptability to language-specific nuances. Employing incremental learning techniques ensures continual updates to the model, enabling it to accommodate evolving linguistic contexts effectively. The melspectrograms generated by the Tacotron2 model are then processed using HiFiGAN and WaveGlow vocoders to produce synthesized speech. Post-processing methods are subsequently applied to refine the output and improve its naturalness. Notably, qualitative evaluation of the synthesized speech yielded a commendable Mean Opinion Score for naturalness, signifying the efficacy of the employed approach. By leveraging the insights gleaned from this scholarly work, the integration of the pretrained TTS model into the image-to-speech conversion system has significantly enhanced the quality and naturalness of the synthesized speech output, particularly for the Nepali language.

VI. Results

The Flickr8k dataset consists of 8k images with 4-5 captions for each images. Same dataset was trained on three different models. The dataset was split into training set and validation set. 6.4k dataset are used for training set and remaining 1.6k are used for validation. The validation dataset is used as test set for generating caption and calculation of BLEU scores after the models are successfully trained. The model parameters used in the experiments are summarized in Table II.

TABLE II. Model Parameters

Parameters	Values
Batch Size	32
Embedding Dimension	512
Feed Forward Dimension	2048
Image Size	(224, 224)
Learning Rate	10^{-4}
Loss Function	Sparse Categorical Cross Entropy
Number of Heads	3
Optimizer	Adam
Sequence Length	25
Vocab Size	15,000

The corpus BLEU score is calculated for different model architectures, with the results presented in Table III. Fig. 3 displays the same data in bar chart form for better comparison.

TABLE III. Corpus BLEU Score

Model	Encoder	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Layered Transformer	EfficientNetB7	0.381	0.176	0.0664	0.02145
Learned Embeddings	EfficientNetB7	0.475	0.269	0.124	0.050
Single Layered Transformer	ResNet-50	0.446	0.268	0.184	0.139

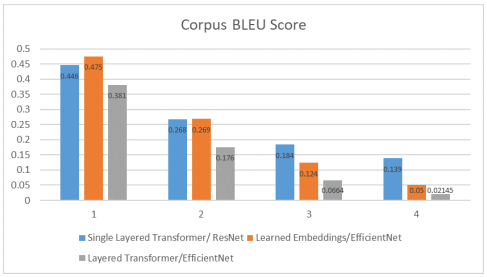


Fig. 3. Corpus BLEU score.

Similarly, the sentence-level BLEU score is calculated, providing a more granular and detailed evaluation of the generated captions. The scores are presented in Table IV, and the bar graph is shown in Fig. 4.

TABLE IV. Sentence BLEU Score

Model	Encoder	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Single Layered Transformer	ResNet-50	0.5223	0.3858	0.3013	0.2462
Learned Embeddings	EfficientNetB7	0.5217	0.3950	0.3171	0.2651
Layered Transformer	EfficientNetB7	0.4960	0.3528	0.2662	0.2125

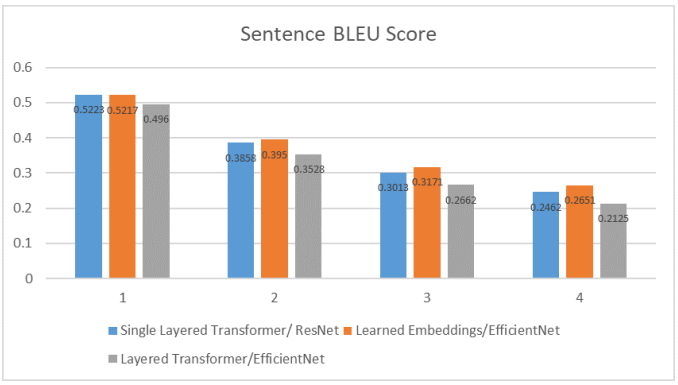


Fig. 4. Sentence BLEU score.

Fig. 5 to 8 present the generated Nepali captions alongside their English equivalents, demonstrating the Single Layered Transformer model's effectiveness in generating captions for different images.

Fig. 5. Equivalent English caption: A guy in a blue shirt is ready to kick a football.

Fig. 6. Equivalent English caption: Two dogs are running on the grass.

Fig. 7. Equivalent English caption: A big bird has landed on the water.

Fig. 8. Equivalent English caption: A dog is running on the field.

VII. Discussion

BLEU scores of layered transformer model is observed to be significantly lower than other models. This is found to be because of the overfitting issue as the increase in layer significantly increased the learnable parameters thus the feature extracted and the size of dataset used were not enough to overcome overfitting issue.

Result obtained by using learnable pretrained embedding layer worked slightly better than other models. The intricate relationship between the words are captured well using embedding layer. This is indicated by improvement in BLEU-2, BLEU-3, BLEU-4 scores. However in case of word level tokenization, BLEU-1 for learned embedding is greatest among all models but the relationship between more than two words could not be generalized well which is marked by low BLEU-3 and BLEU-4 scores.

Single layered transformer model with Resnet-50 as feature extractor is performing at similar accuracy when compared to learned embeddings. Since the complexity of model is low and the dataset is also sparse with only 8k images, training a model like single layered transformer with less learnable parameter performed remarkably well. However, Pretrained embedding could be preferable if abound dataset is available for image captioning.

The results highlight the model's ability to capture diverse scenarios with remarkable detail. In Fig. 5, it not only identifies a person but also recognizes the blue color of the shirt, highlighting its capability to capture intricate visual features. Similarly, in Fig. 6, the model accurately detects two dogs running on the grass, effectively distinguishing multiple objects and their actions. Fig. 7 showcases its ability to recognize both a large bird and the presence of water, indicating its understanding of different elements within a scene. Likewise, in Fig. 8, the model successfully identifies a dog running on a field, emphasizing its proficiency in capturing both subject and activity. These examples indicate the model's effectiveness in generating descriptive and contextually rich captions.

VIII. Conclusion

This study explored a deep learning-based approach for image captioning and speech synthesis in Nepali. Transformer-based architectures were employed for caption generation, while state-of-the-art text-to-speech models were utilized to produce spoken descriptions. Evaluation using BLEU scores demonstrated that a single-layered transformer model with ResNet-50 as the feature extractor performed competitively, while learned embeddings offered slight improvements in capturing linguistic relationships.

One of the primary challenges identified was the limited availability of high-quality Nepali image-caption datasets, which impacted model performance. Additionally, the complex morphological structure of Nepali presented difficulties in tokenization and evaluation, emphasizing the need for more refined linguistic processing techniques. Addressing these challenges in future research could lead to improvements in both caption quality and speech synthesis accuracy.

The findings contribute to the underexplored domain of Nepali-language AI applications, highlighting the potential for advancements in image-to-speech technology for low-resource languages. Future research can focus on dataset expansion, the integration of more advanced language models, and enhancements in text-to-speech synthesis to better capture the nuances of Nepali pronunciation and grammar.

References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," 2014, arXiv:1411.4555.

[2] S. K. Mishra, R. Dhir, S. Saha, P. Bhattacharyya, and A. K. Singh, "Image captioning in Hindi language using transformer networks," *Computers & Electrical Engineering*, vol. 92, p. 107114, 2021, doi:10.1016/j.compeleceng.2021.107114.

[3] A. S. Ami, M. Humaira, M. A. R. K. Jim, S. Paul, and F. M. Shah, "Bengali Image Captioning with Visual Attention," in *Proc. 23rd International Conference on Computer and Information Technology (ICCIT)*, 2020, pp. 1-5, doi:10.1109/ICCIT51783.2020.9392709.

[4] A. Adhikari and S. Ghimire, "Nepali Image Captioning," in *Proc. Artificial Intelligence for Transforming Business and Society (AITB)*, 2019, vol. 1, pp. 1-6, doi:10.1109/AITB48515.2019.8947436.

[5] B. Subedi and B. K. Bal, "CNN-Transformer based Encoder-Decoder Model for Nepali Image Captioning," in *Proc. 19th International Conference on Natural Language Processing (ICON)*, New Delhi, India, 2022, pp. 86-91.

[6] Y. Wang *et al.*, "Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model," 2017, arXiv:1703.10135.

[7] J. Shen *et al.*, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," 2017, arXiv:1712.05884.

[8] S. O. Arik *et al.*, "Deep Voice: Real-time Neural Text-to-Speech," 2017, arXiv:1702.07825.

[9] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Neural Speech Synthesis with Transformer Network," 2019, arXiv:1809.08895.

[10] S. Khadka, R. G.C., P. Paudel, R. Shah, and B. Joshi, "Nepali Text-to-Speech Synthesis using Tacotron2 for Melspectrogram Generation," in *Proc. SIGUL 2023*, pp. 73-77, doi:10.21437/SIGUL.2023-16.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2015, arXiv:1512.03385.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," 2020, arXiv:2010.05646.

[14] I. J. Goodfellow *et al.*, "Generative Adversarial Networks," 2014, arXiv:1406.2661.

[15] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2014, arXiv:1409.1556.

[16] L. Srinivasan and D. Sreekanthan, "Image Captioning-A Deep Learning Approach," 2018, [Online]. Available: https://api.semanticscholar.org/CorpusID:85556880.

Knowledge Graph Path-Enhanced RAG for Intelligent Residency Q&A

Jian Zhu, Huajun Zhang*, Jianpeng Da, Hanbing Huang, Chongxin Luo, Xu Peng
School of Automation, Wuhan University of Technology, Wuhan 430070, China

***Abstract*—As the demand for efficient information retrieval in specialized domains continues to rise, vertical domain question-answering systems play an increasingly important role in addressing domain-specific knowledge needs. This paper proposes a retrieval-augmented generation method that integrates path search in knowledge graphs to enhance intelligent question-answering systems for professional information retrieval. The proposed approach leverages fine-tuned large language models to identify entities and extract relations from user queries, combining pruned marker method with a shortest path generation tree algorithm to efficiently retrieve relevant information. The retrieval results are then integrated with user queries using prompt engineering to generate precise and contextually relevant answers. To validate the practicality of the proposed method, this paper develops a knowledge graph encompassing policies, regulations, and social services within the household registration vertical domain. The experimental results within this vertical domain reveal that the proposed method significantly outperforms existing methods in terms of evaluation metrics such as BLEU, ROUGE, and METEOR, achieving improvements exceeding 3%. Furthermore, ablation experiments validate the importance of combining path search algorithms with fine-tuning techniques in enhancing the question-answering performance.**

***Keywords*—Retrieval-augmented generation; path search; knowledge graph; household registration policy vertical field**

I. Introduction

With the rapid development of natural language processing (NLP) technology, intelligent question-answering systems have been widely applied in the field of policy interpretation and are gradually becoming important tools for obtaining information in medical, legal, educational, and other professional fields. These systems understand user questions through semantic analysis and information extraction technology, and generate relevant answers from knowledge bases or domain-specific databases [1]. Compared with traditional keyword matching, intelligent question-answering systems can more accurately capture the semantic intent of the user, thereby improving the accuracy of the answers and user experience [2]. However, intelligent question-answering systems still face notable limitations. For instance, although pre-trained language models exhibit strong performance across various tasks, they remain less effective in addressing queries that demand a high level of domain-specific expertise, such as medical diagnoses or legal code interpretation [3]. Additionally, due to the system's dependence on static knowledge bases, it is difficult to update domain knowledge in real-time, which may lead to delays in answers [4].

The research on intelligent question-answering systems typically unfolds along multiple technical routes, which mainly include rule-based question-answering systems, information retrieval-based question-answering systems, and generation model-based question-answering systems. Rule-based question-answering systems rely on pre-defined rules and templates, matching answers through the parsing of keywords and syntactic structures. They perform well in structured, fixed-pattern questions, but they have limitations in addressing diverse needs and complex issues [5]. Information retrieval-based question-answering systems find answers by searching for relevant documents, utilizing keyword matching and semantic retrieval techniques to quickly locate content related to the user's question [6]. Their advantage lies in their ability to handle large document repositories and providing diverse sources of answers. However, they often struggle to meet the needs for questions that require complex reasoning or the generation of new answers. In comparison, generation model-based systems directly apply natural language generation technology to generate answers, performing even better in handling complex issues [7].

In recent years, Retrieval-Augmented Generation (RAG) technology has attracted widespread attention due to its combination of the advantages of information retrieval and generation models. RAG first finds relevant documents through information retrieval and then generates more accurate and semantically rich answers based on these documents citeyang2019xlnet. Compared to traditional generation models, RAG maintains the flexibility and generation capability of the latter while ensuring the contextual relevance of the answers [8]. RAG models combine a retrieval module based on BERT with a GPT generation module, demonstrating outstanding performance in knowledge-intensive tasks [9]. Research shows that RAG not only improves the accuracy of answers but also effectively reduces the lag caused by outdated information by retrieving the latest documents [10]. RAG has shown significant effects in open-domain question-answering tasks, especially in areas such as law and policy where knowledge updates are frequent, ensuring timeliness of answers through real-time retrieval [11]. Additionally, multimodal RAG systems that combine image and text data further expand the application of the technology and enhance its ability to handle complex tasks[12]. RAG models have significant advantages in handling complex policy-related questions, especially in cross-domain and multi-dimensional issues, significantly enhancing the relevance and depth of the answers [13].

In the context of the current complex policy system and the growing demand for public services, how to efficiently acquire and utilize knowledge of the household registration policy has become one of the core issues of concern for policymakers and public service institutions [14] [15] [16]. As an important part of social governance, the household registration policy have a

wide-ranging impact that covers multiple areas such as social security, education, and employment[17]. Although RAG technology has made significant progress in question-answering systems, it still faces some critical challenges. RAG technology has a high dependency on the quality of the retrieval stage; if the retrieval results include irrelevant documents, the generated answers are often not accurate; in multi-hop reasoning tasks, insufficient information is more likely to lead to incomplete answers [18]. Additionally, when handling long texts, RAG technology often suffers from context loss, leading to answers that lack coherence [19]. Furthermore, RAG technology has a high computational cost, and the system response speed is slower when processing large document libraries, making it difficult to meet the needs of real-time applications[20].

The above challenges have prompted the development of a retrieval-enhanced generation method that incorporates knowledge graph path search, which has been applied to the household registration vertical field. This method combines knowledge graph path search technology with entity recognition performed by a fine-tuned large model, mapping user queries to corresponding entities and relationships in the graph and using path algorithms for inference to precisely extract relevant knowledge information. Finally, the large model is used for natural language generation to provide answers that are contextually appropriate. In addition, in the subsequent application phase, this paper will construct a knowledge graph that covers policies, regulations, and social services in the household registration vertical field, and further discuss its specific implementation and application effects in vertical field question-answering systems. The primary contributions of this paper are as follows.

1) Proposal of an MST-based algorithm for knowledge extraction: This paper presents a novel Minimum Spanning Tree (MST) algorithm based on the shortest-path methodology to facilitate efficient knowledge extraction from graphs. When integrated with a large language model for natural language generation, this approach enables the system to deliver accurate, contextually relevant policy responses to user inquiries. This enhancement significantly improves the intelligence and responsiveness of the question-answering system.

2) Fine-tuned model for entity recognition and relationship extraction: For tasks involving entity recognition and relationship extraction within the household registration domain, this study employs fine-tuning techniques on large models, coupled with the knowledge graph, to achieve robust entity recognition and relationship inference. This method effectively processes complex entities and multi-layered relationships within the policy domain, ensuring that user queries are accurately aligned with pertinent knowledge points in the graph.

3) Development of a specialized knowledge graph for household registration: A comprehensive knowledge graph tailored to the household registration domain has been developed, incorporating policy content, legal regulations, and related social services. The graph systematically organizes the interconnections among various policy clauses and entities, providing a structured foundation for addressing complex policy queries. This framework significantly improves the system's accuracy in addressing issues related to household registration.

The rest of this paper is organized as follows. Section II describes the related work of this paper. Section III introduces the retrieval-enhanced generation technique, utilizing knowledge graph path search to improve query handling. Section IV details the construction of a knowledge graph within the civil affairs domain and demonstrates the application of this technique to enhance information retrieval and response generation within civil affairs contexts. As for Section V, we conduct comprehensive experiments and sufficient analysis. Finally, Section VI provides a summary of the entire paper.

II. Related Work

Knowledge graphs can better capture semantic relationships in complex domains by structuring the representation of entities and their relationships, thus providing more accurate retrieval cues and context information for RAG technology [21]. In the field of knowledge graphs, early research focused mainly on how to effectively extract entities and relationships from text data to build a structured knowledge base. For example, methods based on statistical models can automatically identify and associate relevant entities from multiple data sources, laying the foundation for the construction of knowledge graphs [22]. However, these traditional methods often rely on predefined rules and templates, and their ability to handle complex semantic relationships is relatively limited. To overcome this limitation, research on knowledge graph completion techniques based on deep learning has gained widespread attention in recent years. Graph neural networks (GNNs), in particular, have been widely applied due to their superior performance in handling complex entity relationship graph structures [23]. In addition, multi-task learning techniques have also begun to be introduced into the construction of knowledge graphs to handle entity recognition and relation extraction simultaneously. Compared to single-task learning, multi-task learning can better utilize the correlations between different tasks, thereby improving the overall performance of the model. For example, through a multi-task learning framework, joint training of entity recognition and relation classification has been achieved, effectively improving the accuracy and efficiency of knowledge graph construction [24]. However, since knowledge graphs are a semantic network that dynamically updates, how to achieve efficient knowledge updating and completion remains an important research challenge [25]. At the same time, with the rapid development of large-scale pre-trained language models (LLMs), the integration of knowledge graphs with LLMs has become a new research direction. Knowledge graphs can provide structured knowledge support for LLMs, thereby improving accuracy in specific domain question-answering systems [26]. For example, by combining knowledge graphs with LLMs, the model's explainability and personalized recommendation capabilities can be enhanced [27]. However, since LLMs are essentially a "black box" model, the issue of how to utilize its powerful capabilities while maintaining the transparency and explainability of the system remains a pressing problem [28].

The Retrieval-Augmented Generation (RAG) technique, which integrates knowledge graphs with pre-trained language models, significantly improves the accuracy and consistency of generated content by introducing structured knowledge into the generative process. For example, Yasunaga et al. proposed the QA-GNN model, a framework that combines

language models with knowledge graphs to handle complex question-answering tasks. This model leverages path reasoning within the knowledge graph, thereby enhancing the logical consistency of the generated responses [29]. Similarly, Hu et al. developed the OREO-LM model, which connects pre-trained language models with knowledge graph reasoning modules via a knowledge interaction layer, resulting in improved performance on question-answering tasks [30]. In dialogue generation, advancements have been made by incorporating structured knowledge from graphs. Xu et al., for instance, introduced the DMKCM model, which fuses structured knowledge graphs with unstructured textual data to produce dialogue with greater depth and coherence [31]. In a related development, Kang et al. proposed the SURGE framework, which retrieves relevant subgraphs and employs contrastive learning during generation, enhancing the quality and knowledge consistency of the generated dialogues [32].

Significant strides in dialogue generation have been achieved by integrating knowledge graphs with pre-trained language models. Peng et al. introduced the GODEL model, which incorporates external knowledge during pre-training to improve performance across task-oriented dialogue, conversational question-answering, and open-domain dialogue. This model outperforms existing pre-trained dialogue models, particularly in few-shot fine-tuning scenarios [33]. Yang et al. further advanced the field with the Graph Dialog model, which embeds structural knowledge from graphs into an end-to-end task-oriented dialogue system. By utilizing graph convolutional networks (GCNs), Graph Dialog captures structural information from both dialogue history and knowledge bases, thereby generating more precise responses [34].

III. Retrieval-Augmented Generation Based on Knowledge Graph Path Search

The retrieval-enhanced generation method based on knowledge graph path search includes three modules: entity recognition, path search, and answer generation. In the entity recognition module, a large language model is used to analyze the user's question and extract a set of entities similar to the knowledge graph nodes. In the path search module, the identified entities are linked to knowledge graph nodes, and the shortest path is found using a 2-hop coverage query, generating a tree structure that covers all entity nodes. Finally, in the answer generation module, the results of the path search are input into a large language model along with the user's question. Through prompt engineering, the model generates and returns the final responses. The overall workflow diagram of the retrieval-enhanced generation method based on knowledge graph path search is illustrated in Fig. 1.

A. Entity Recognition Module

In this paper, a large language model is employed to conduct an in-depth analysis of user input queries, extracting a set of entities related to nodes within a knowledge graph. This process leverages advanced natural language processing techniques and harnesses the semantic understanding capabilities of the model, enabling accurate identification of essential entities within user queries. These entities, typically exhibiting strong correlations with knowledge graph nodes, form a crucial foundation for subsequent retrieval and reasoning tasks.

The user query Q is initially received as input, upon which the model conducts multi-level semantic parsing and entity mapping to identify a potential set of entities $E = [E_1, E_2, \ldots, E_n]$. This set is derived from relevant nodes within the knowledge graph, guided by the context of the user query and its implicit semantic cues. Through this entity recognition method, the model not only extracts entities directly pertinent to the user query but also captures potential implicit information, enhancing the alignment between the query and the knowledge graph. This process can be represented as follows:

$$E = LLM\,(Q) \tag{1}$$

Prompt Template for Household Registration Information Extractor:

[Prompt]: You are an expert in household registration information extraction, skilled in analyzing provided sentences and extracting specified household registration items. Please note that the output should only include the household registration items mentioned in the text, and they should be numbered sequentially. The input sentence may correspond to multiple household registration items; do not output any extraneous information. The household registration items include: 1) Proof of legal and stable residence. 2) Correction of place of birth.

[Input Text]: I want to register my parents' household registration under my name. What materials are required?

[Output Keywords]: 1) Parental relocation to child's household registration

Through the entity recognition and retrieval stage, the model effectively minimizes the semantic gap between the user's query and the nodes within the knowledge graph, thereby improving the accuracy of subsequent knowledge graph queries and information reasoning. This process deepens the system's comprehension of the user's query and establishes a more precise entity foundation for path search and information integration during the generation stage. Consequently, it enhances both the quality of the system's responses and the overall user experience.

B. Path Search Module

In this paper, a knowledge graph is constructed for a specialized domain, transforming unstructured or semi-structured texts related to domain knowledge into structured information. This information is stored within a graph structure to provide reliable data support for the subsequent question-answering system. The NEO4J graph database is utilized to store multi-relational data, leveraging Cypher queries for data import and graph data retrieval. Cypher, a descriptive graph query language, is recognized for its simplicity and robust functionality. Furthermore, NEO4J offers the neo4j-import tool, which facilitates the rapid import of large-scale data.

Given the large number of nodes and relationships in the constructed household registration policy knowledge graph, directly executing path searches on the graph proves time-consuming, which adversely affects user experience. Additionally, storing the distances and paths between each pair of nodes offline demands excessive memory space. To address this issue,

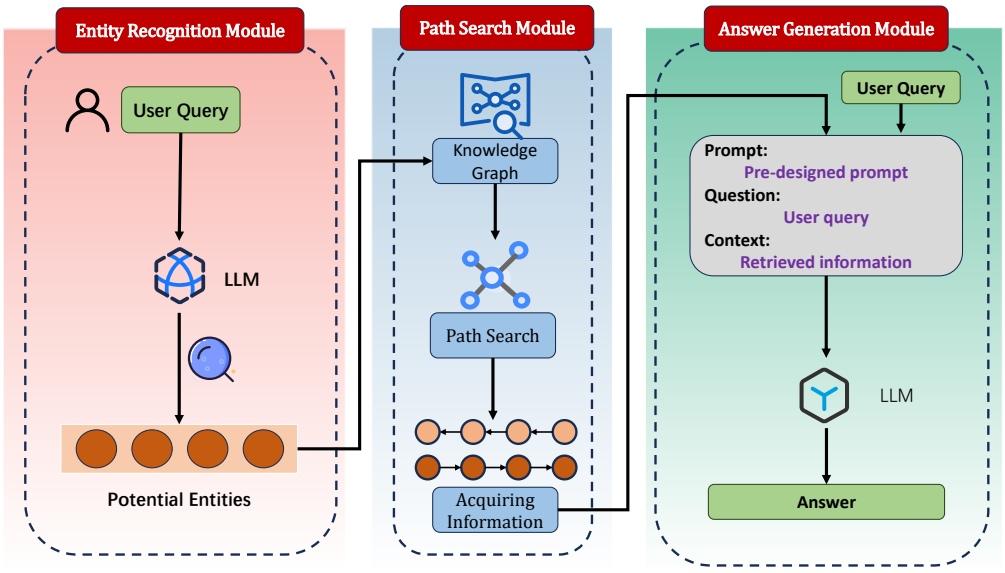


Fig. 1. System workflow diagram.

this paper implements a pruning labeling strategy. During the preprocessing stage, labels are assigned to nodes within the graph, allowing these labels to be utilized at query time to swiftly determine the shortest path between node pairs. Since relationships in the knowledge graph are unweighted, an in-depth analysis was performed on an undirected, unweighted graph. Table I provides explanations of the commonly used symbols in this paper.

1) Definition of shortest distance and shortest path queries: In the knowledge graph, a set of composite labels, denoted as $L(u)$, must be precomputed for each node u. The label set $L(u)$ comprises multiple label pairs, each represented as a triplet (v, d_{uv}, p_{uv}), where v is a node, d_{uv} denotes the shortest distance from node u to node v, and p_{uv} represents the set of all nodes along the shortest path between u and v.The relationships in the knowledge graph are not assigned weights, so the shortest distance between two nodes is represented by the number of nodes in the shortest path between them.

When querying the shortest path between nodes s and t, the shortest distance between s and t is first calculated. To facilitate this, a function $QD(s,t,L)$ is defined, with its specific form as Eq. (2). u represents a common node between s and t. For each common node u, the distance from node s to node t is calculated as $d_{ut} + d_{us}$, with the minimum value among all possible distances considered as the shortest distance.If no common nodes exist between $L(s)$ and$L(t)$, $QD(s,t,L)$ is defined as $+\infty$.

In addition to calculating the shortest distance between nodes s and t,, it is essential to record the path information. To accomplish this, a function $QP(s,t,L)$ is defined to retrieve the shortest path, with its specific definition as Eq. (3). $QP(s,t,L)$returns the path $p_{us} \cup p_{ut}$, representing the shortest path between nodes s and t. This path is constructed by combining the optimal subpaths from node s to the common node u and from u to node t.

For any pair of nodes s and t, the shortest distance between them can be determined by calculating $QD(s,t,L)$, from which the shortest path can then be derived.

2) Generation of composite labels: A composite label generation algorithm based on pruned breadth-first search (BFS) is proposed to optimize the computational efficiency of traditional BFS. Similar to the naive approach, this algorithm performs searches sequentially in the order of vertices $v_1, v_2, \ldots, v_n$. The process begins with an empty index L'_0, and after each pruned BFS iteration from vertex v_k, it updates the index L'_{k-1} to a new index L'_k based on the information gathered. Algorithm 1 outlines the specific steps involved in this composite label generation algorithm.

Algorithm 1 Composite Label Construction Algorithm

$L_k(v) \leftarrow \emptyset$ for all $v \in V$
for $i \leftarrow 1$ **to** n **do**
  $L_i(v) \leftarrow L_{i-1}(v)$ for all $v \in V$
  $distance(v_i) \leftarrow 0$; $distance(v) \leftarrow \infty$ for all $v \in V \setminus \{v_i\}$
  $path(v_i) \leftarrow [v_i]$; $path(v) \leftarrow \emptyset$ for all $v \in V \setminus \{v_i\}$
  $Q \leftarrow$ a queue with only one element v_i
  while Q is not empty **do**
    $u \leftarrow Q$.pull()
    if $d_G(u) < QD(v_i, u, L_{i-1}(v))$ **then**
      $L_i(u) \leftarrow L_{i-1}(v) \cup \{v_i, distance(u), path(u)\}$
      for $w \in$ Neighbors(u) **do**
        $distance(w) \leftarrow distance(u) + 1$
        $path(w) \leftarrow path(u) \cup \{w\}$
        Q.put(w)
      end for
    end if
  end while
end for
return L_n

$$QD(s,t,L) = \min\{d_{ut} + d_{us} | (u, d_{us}, p_{us}) \in L(s), (u, d_{ut}, p_{ut}) \in L(t)\} \quad (2)$$

$$QP(s,t,L) = \operatorname{argmin}\{d_{ut} + d_{us} | (u, d_{us}, p_{us}) \in L(s), (u, d_{ut}, p_{ut}) \in L(t)\} \quad (3)$$

TABLE I. EXPLANATION OF KEY SYMBOLS

Symbol	Description
$G = <V, E>$	Knowledge graph
n	Number of vertices in the knowledge graph
m	Number of edges in the knowledge graph
$N_G(v)$	Neighboring nodes of node v in the knowledge graph
$d_G\ (u, v)$	Shortest distance between nodes u and v in the knowledge graph
$P_G\ (u, v)$	Set of all nodes on the shortest path between nodes u and v in the knowledge graph

The composite label generation algorithm proceeds as follows: Given an input knowledge graph $G =< V, E >$, the algorithm outputs composite labels comprising distance-path pairs for each node in the graph. Initially, the label $L_k(v)$ for each node $v \in V$ is set as an empty set. The algorithm begins at node $v_i$and performs BFS across all nodes. For the starting node v_i, the distance is initialized to 0, and the path path(v_i) is set to $[v_i]$. For all other nodes, distances are initialized to infinity, and their paths are represented as empty sets. The queue Q is initialized to contain only v_i.

During the BFS process, when visiting a vertex u with distance δ and path P, the algorithm checks if the query result $QD\left(v_k, u,\ L'_{k-1}\right) \leq \delta$. If this condition is met, vertex u is pruned, meaning that the label (v_k,δ ,P) is not added to $L'_k(u)$, and traversal of $u\prime s$ adjacent edges is terminated. Conversely, if $QD\left(v_k, u,\ L'_{k-1}\right) > \delta$, the label $L'_k(u)$ is updated to $L'_{k-1}(u) \cup \{(v_k, \delta, P)\}$, and the algorithm proceeds to traverse all adjacent edges of vertex u. For each neighboring vertex w of u, the algorithm updates $w\prime s$ distance and path accordingly and enqueues w into Q.

Fig. 2 presents examples of pruned BFS traversals. Yellow vertices denote the roots, blue vertices denote those which we visited and labeled, green vertices denote those which we visited but pruned, and gray vertices denote those which are already used as roots. The initial pruned BFS, originating from vertex 1, successfully visits all vertices, as shown in Fig. 2a. In the subsequent BFS traversal starting from vertex 2 (Fig. 2b), upon reaching vertex 6, we observe that $QD(2, 6, L'_1) = d_G(2, 1) + d_G(1, 6) = 3 = d_G(2, 6)$, allowing us to prune vertex 6 and avoid traversing any of its outgoing edges. Similarly, vertices 1 and 12 are pruned in this traversal. As additional BFS traversals are performed, the search space progressively diminishes in size, as illustrated in Fig. 2c and Fig. 2d.

The algorithm enhances computational efficiency by employing pruning techniques to minimize redundant traversal, thereby optimizing the process. The final output is the updated label set L_n, which constructs composite labels for each node, embedding both shortest path and corresponding distance information. These composite labels enable rapid determination of the shortest path between any two connected nodes in the graph, eliminating the need for exhaustive graph traversal or complex path-finding operations. This composite labeling mechanism thus provides an efficient approach for querying shortest paths, significantly improving query speed and conserving computational resources.

3) Minimum spanning tree query algorithm based on shortest path.: Given that user queries frequently involve multiple entities, represented as a set of nodes within the graph, the task of identifying the minimum-cost tree that connects these nodes is defined as the minimum spanning tree problem. Algorithm 2 leverages the pre-established composite labels to compute the shortest paths between nodes, thus facilitating the construction of an optimal tree that spans all nodes within the specified set.

Algorithm 2 Minimum Spanning Tree Query Algorithm Based on Shortest Path

```
tree ← ∅
distance(k) ← 0, path_i ← {N_k}
for i ← 1 to k − 1 do
    for j ← i + 1 to k do
        d_G(N_i, N_j) ← QD(N_i, N_j, L_n)
        PG(N_i, N_j) ← QP(N_i, N_j, L_n)
    end for
    distance(i) ← min{d_G(N_i, N_j) | j = 1, 2, ..., k}
    path(i) ← arg min PG(N_i, N_j)   for j = 1, 2, ..., k
end for
Q ← {1, 2, ..., k}
while Q is not empty do
    y ← arg min (distance(m) | m ∈ Q)
    if y ∈ Q then
        tree ← tree ∪ path(y)
        Q.remove(y)
    end if
end while
return tree
```

The minimum spanning tree query algorithm based on shortest path operates as follows: Given input parameters consisting of composite labels L_n and a set of k nodes $N_1, N_2 \ldots N_k$, the algorithm outputs the minimum spanning tree that connects these nodes. Initialization starts by setting the minimum spanning tree, denoted as tree, to an empty set. The initial path distance $distance\,(k)$ for node N_k is set to 0, with the path $path\,(k)$ initialized to N_k. For the remaining nodes $N_1, N_2 \ldots N_{k-1}$, distances are initially set to infinity, and paths are initialized as empty sets.

The algorithm begins with the first node N_1 and performs

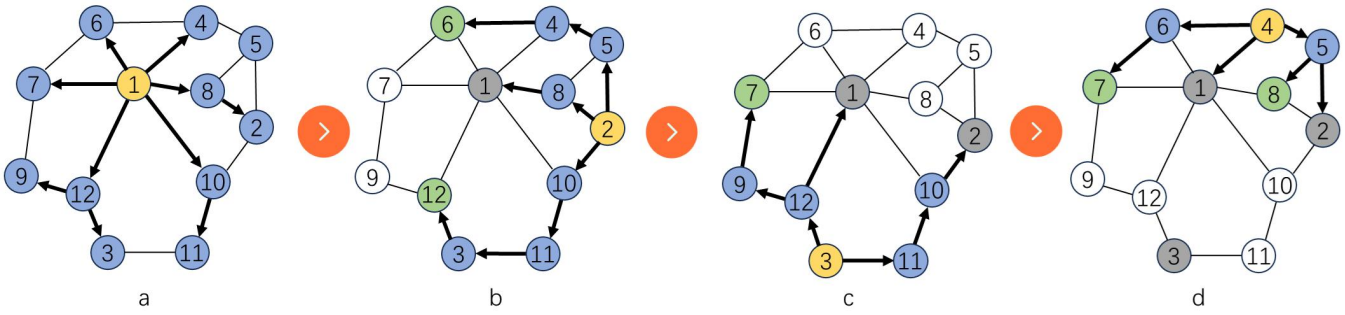


Fig. 2. Examples of pruned BFS traversals.

shortest path queries for each node pair N_i and N_j. Specifically, it calculates the shortest distance $d_G(N_i, N_j)$ and the corresponding path $P_G(N_i, N_j)$ between nodes N_i and N_j using the composite labels L_n. Among all query results, the shortest distance is selected, and the associated shortest path is assigned to $path(i)$. Subsequently, a greedy approach constructs the minimum spanning tree by enqueueing all nodes into Q. In each iteration, the algorithm selects the node y with the smallest path distance from Q, and appends the shortest path $path(y)$ to the minimum spanning tree tree. The node y is then removed from Q, and this process continues until all nodes have been processed.

The algorithm utilizes composite labels to pre-store and compress distance and path information between nodes, enabling rapid retrieval of shortest paths and their corresponding distances through label queries. This approach markedly enhances path computation efficiency by eliminating the need for complex graph traversal each time a shortest path calculation is required. Additionally, the algorithm adopts a greedy strategy to incrementally construct the minimum spanning tree, selecting the next path based on the minimum distance between nodes and adding it to the spanning tree. This greedy selection ensures that each chosen path is optimal at its respective step, thus guaranteeing the overall optimality of the final spanning tree. This method not only simplifies the process of constructing the spanning tree but also maximizes the algorithm's efficiency.

The algorithm guarantees the accuracy of query results while maintaining low time complexity. By minimizing redundant path calculations and optimizing the construction process, it effectively processes large-scale graph data in a short time frame, making it especially suitable for constructing minimum spanning trees in complex networks with numerous nodes and edges. By integrating the efficient query mechanism of composite labels with the benefits of a greedy strategy, the algorithm offers a rapid and precise solution for minimum spanning tree construction.

4) Approximation Ratio Analysis: As the Steiner tree problem is a classic NP-hard problem in graph theory and combinatorial optimization, obtaining an optimal solution for large graphs with extensive terminal node sets demands considerable computational resources and may be infeasible within a reasonable timeframe. To assess and compare the effectiveness of approximation algorithms, the approximation ratio is commonly employed as a metric. This ratio quantifies the discrepancy between the solution produced by an approximation algorithm and the true optimal solution. It is typically defined as Eq. (4).

$$r = \frac{w(T_{appopt})}{w(T_{opt})} \tag{4}$$

In this context, T_{opt} represents the optimal solution to the Steiner tree problem, T_{appopt} denotes the spanning tree obtained using the minimum spanning tree query algorithm based on the shortest path, and $w(T)$ indicates the total weight of tree T.

For a query on the node set $N = \{N_1, N_2 \dots N_K\}$, the algorithm constructs the tree by iteratively adding the shortest paths between nodes. In each iteration, a shortest path is appended to the spanning tree. Suppose that, in each greedy selection, the node $pair(N_i, N_j)$ is chosen, and the shortest path $P_G(N_i, N_j)$ with length $d_G(N_i, N_j)$ is identified between them. Consequently, the weight of the spanning tree generated by the approximation algorithm can be expressed as the sum of all shortest path lengths.

$$w(T_{appopt}) = \sum_{edges} d_G(N_i, N_j) \tag{5}$$

The optimal Steiner tree T_{opt} is the spanning tree with the minimum total weight that connects the specified node set N. This tree may incorporate non-terminal nodes to reduce the connection costs between terminal nodes, and therefore, its total weight $w(T_{opt})$ represents the minimum achievable weight for a spanning tree in the graph.

$$w(T_{opt}) = \min_{tree T \supseteq N} w(T) \tag{6}$$

The upper bound of the approximation ratio for the classical greedy approximation algorithm for the Steiner tree problem is $2\left(1 - \frac{1}{k}\right)$. When k is small, the shortest paths between terminal node pairs closely align with the optimal solution, causing the approximation ratio to approach 1. However, as the number of terminal nodes k and the number of paths between terminal nodes increase, the algorithm increasingly overlooks

the introduction of Steiner points, leading to a gradual rise in the approximation ratio. Nonetheless, the upper bound remains capped at 2, as Eq. (7).

$$r = \frac{w\left(T_{appopt}\right)}{w\left(\ T_{opt}\right)} \leq 2 \tag{7}$$

The above analysis of the approximation ratio suggests that for smaller node sets, the approximation algorithm can closely approach the optimal solution, offering reliable performance guarantees. Consequently, the algorithm demonstrates a degree of effectiveness in addressing large-scale Steiner tree problems.

5) Acquiring graph information via the minimum spanning tree: In this paper, acquiring neighborhood information from the minimum spanning tree aims to establish foundational data support for an advanced question-answering system, thereby enhancing the system's capacity to address complex queries. Specifically, a first-order neighborhood query is performed for each node in the minimum spanning tree, retrieving all directly connected neighboring nodes and their respective relationships to construct each node's local subgraph. Through this approach, the minimum spanning tree provides a streamlined representation of the graph, maintaining essential shortest connections between nodes while eliminating redundant information. By systematically extracting neighborhood information for each node within the minimum spanning tree, the algorithm comprehensively captures the graph's local structural characteristics. This approach ensures the acquisition of complete local information for each node, supplying the necessary contextual and semantic relationship data to support the question-answering system.

C. Answer Generation Module

Given that path algorithms often yield highly redundant information, directly presenting this data to the user can result in information overload and unstructured content, diminishing the overall user experience. To mitigate this issue, we incorporate prompt engineering using a large language model. By designing tailored prompts, we integrate the user's query with relevant information retrieved from external knowledge bases, thereby generating responses that are more targeted, accurate, and contextually relevant.

We have developed a method to convert triples from a knowledge graph (i.e. "entity-relation-entity") into a textual format that is suitable for input into large language models. For each triple, we generate a complete sentence by incorporating connecting words such as "of" and "is". For example, the triple (A, relation, B) is transformed into the sentence "The relation of A is B". This conversion technique not only preserves the structured information from the graph but also facilitates its seamless integration into the LLM input, thereby guiding the model to generate more accurate and coherent responses.

Specifically, the large language model leverages prompt engineering alongside its robust pre-trained generation capabilities to guide the response generation process. Initially, we combine the user's input query Q with a set of relevant retrieved information $[I_1, I_1, \ldots, I_n]$ to construct a prompt, denoted as $Prompt(Q, [I_1, I_1, \ldots, I_n]$. This prompt is then input into the generative model (LLM), which, utilizing both the prompt and the external information, produces a contextually appropriate and precise answer Answer. This process can be represented as Eq. (8).The specific prompts are as follows:

Prompt Template for Response Generator:

[Prompt]: You are a text summarization expert. Please provide answers based solely on the content of the text and conversation records, without fabrication. Pay attention to the distinctions between "approval department", "acceptance department", and "review department", and avoid unnecessary information.

[Question]: What is the application method for replacing (or reissuing) a resident household register?

[Text information]: The acceptance department for replacing (or reissuing) a resident household register is the local police station. The application methods and processing timelines for replacing (or reissuing) a resident household register are: in-person at the service window (1 working day) or online (3 working days). The application conditions for replacing (or reissuing) a resident household register are as follows: if the "Resident Household Register" is damaged or lost and needs to be replaced or reissued, the head of the household can apply for a reissued register for the entire household, while household members can apply for a register that includes only the cover page and their own page.

$$Answer = LLM\left(Prompt\left(Q, [I_1, I_2, \ldots, I_n]\right)\right) \tag{8}$$

Through effective prompt design, the large language model is guided to integrate internal knowledge with external information sources. By leveraging the model's inherent language generation capabilities alongside external knowledge bases, the generated content aligns more closely with the context of the user's query while filtering out redundant information and retaining essential content. This approach ensures both contextual coherence and factual accuracy in the responses, thereby enhancing the user experience.

IV. Construction of the Knowledge Graph in the Household Registration Vertical Domain

This paper presents the construction of a knowledge graph specific to the household registration domain. It transforms unstructured or semi-structured texts, such as policies and regulations related to household registration, into structured information stored in a graph format. This structured representation offers reliable data support for the subsequent question-answering system.

The construction process of the household registration policy knowledge graph encompasses several key steps, including knowledge acquisition, knowledge integration, and knowledge import. The knowledge acquisition phase is primarily based on the Wuhan Household Registration Business Processing Guide(hereafter referred to as Guide) and a question-answer corpus. Through text data preprocessing and information extraction, event triplets reflecting household registration business conditions, procedures, and required materials are generated. In the knowledge integration phase, synonym resolution

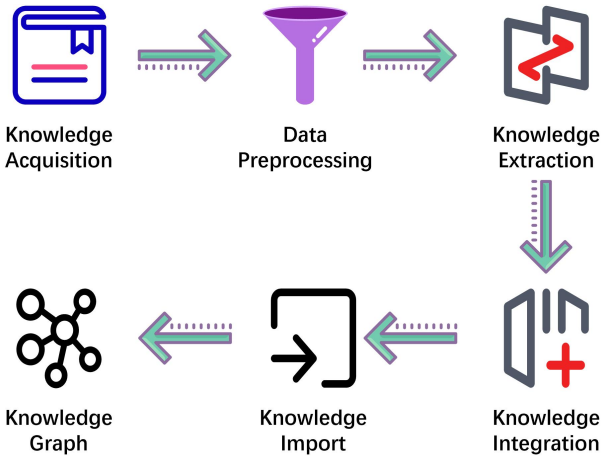


Fig. 3. Flowchart of knowledge graph.

and duplicate entity handling are conducted to ensure data accuracy and consistency. Finally, during the knowledge import phase, the NEO4J graph database is employed to efficiently import the triplet data using Cypher queries and the neo4j-import tool, successfully constructing the knowledge graph that underpins the household registration policy question-answering system. The construction process of the household registration policy knowledge graph is illustrated in Fig. 3.

A. Knowledge Acquisition and Information Extraction

Due to significant regional differences in household registration policies across China, the requirements for handling the same household registration matter can vary considerably from city to city. Therefore, the knowledge sources for the question-answering system must strictly adhere to local household registration policies, and the knowledge graph for the household registration domain must be closed and independent. This paper utilizes the household registration policies of Wuhan, Hubei Province, as the data source to construct the knowledge graph for the household registration domain. The text data primarily derives from the Guide, published by the Wuhan Public Security Bureau, along with a question-answer corpus compiled by household registration center staff. The Guide adopts a semi-structured data format and provides detailed information on 127 household registration services, categorized into six major types and 28 subtypes, covering the conditions for processing these services in Wuhan. It also summarizes numerous domain-specific terms related to household registration and concisely explains how to handle various household registration matters across different scenarios. The question-answer corpus consists of 1,603 question-answer pairs, with all questions being real queries collected manually, closely corresponding to the household registration services described in the Guide.

Based on the textual characteristics of the aforementioned knowledge sources, this paper preprocesses the text data with a focus on household registration events. Specific content for each household registration matter is extracted from the semi-structured tables, resulting in the formation of event triplets. Subsequently, the question-answer texts in the corpus are aligned with the household registration event content, facilitating the construction of entity relationships within the household registration domain. These relationships primarily include item names, application conditions, required materials, processing procedures, application methods, and processing time limits. The specific steps for information extraction are outlined below, with an extraction example illustrated in Fig. 4.

- Given that the Guide is presented in a semi-structured table format, the subject labels and headers of the table can be directly extracted as relationships and attributes. The household registration service item names are identified as head entities, while specific content such as processing locations, procedures, required materials, requirements, application methods, and time limits are extracted as tail entities. Together, these elements constitute the basic event triplets for household registration services.
- To provide supplementary explanations for household registration events under different contextual conditions, the household registration service item names are utilized as head entities, with the supplementary explanations serving as tail entities, thus forming contextual condition triplets for household registration events.
- The questions from the question-answer corpus are manually annotated and mapped to the entities established in steps (1) and (2), resulting in the formation of question triplets for household registration events.

B. Knowledge Integration and Knowledge Import

Merging the data from the two knowledge sources generates a substantial dataset for the knowledge graph; however, numerous duplicate and synonymous entities exist within the two databases. Directly importing these entities would result in redundancy within the knowledge graph. To ensure the accuracy of the question-answering system, this paper integrates the knowledge graphs from both sources by calculating the similarity of entity names. Following this knowledge integration, a total of 1,566 entities across 15 categories and 1,662 relationships across 16 categories were extracted from the documents. The various household registration service contents from the Guide were then imported into the NEO4J database using Cypher queries and the neo4j-import tool.

V. Construction of the Knowledge Graph in the Household Registration Vertical Domain

A. Entity Recognition Experiment

Large language models are trained on extensive text corpora, typically comprising billions of parameters. However, directly applying a large language model to entity recognition tasks often results in suboptimal performance, and conventional fine-tuning methods may induce catastrophic forgetting. The LoRA (Low-Rank Adaptation) efficient fine-tuning technique addresses this issue by employing low-rank decomposition to reduce the number of parameters during training. This

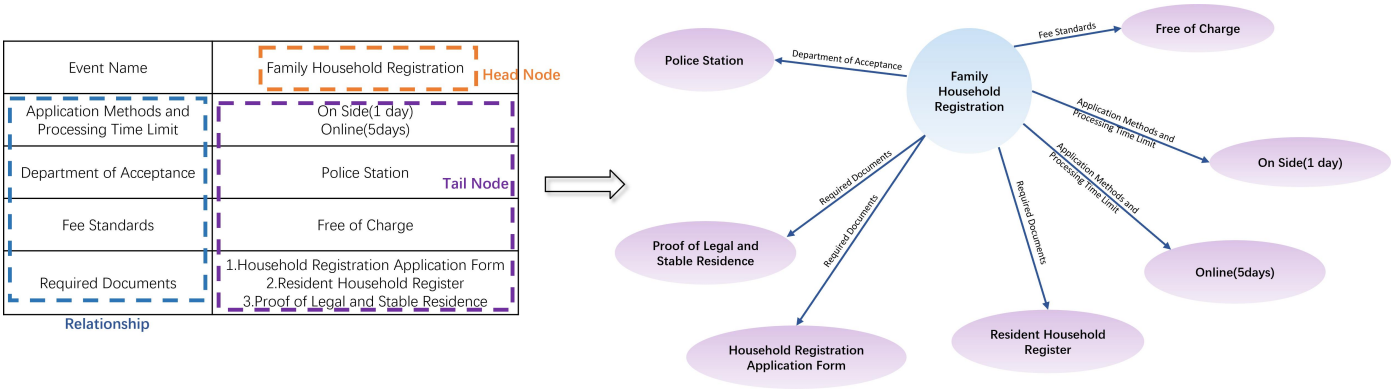


Fig. 4. Example diagram for information extraction.

TABLE II. TEMPLATE OF QUESTIONS FOR VARIOUS CONSULTATION ITEMS

Consultation Items	Template of Questions
Application Methods and Processing Procedures	What application methods are available for processing {item}?
Application Methods and Processing Procedures	What are the procedures for processing {item}?
Precautions	What precautions should be taken when processing {item}?
Application Requirements	What requirements do I need to meet to handle {item}?
Required Documents	What documents are needed to process {item}?
Accepting Department	Where is the accepting department for {item}?
Review Department	Which department handles the review of {item}?
Approval Department	Which department is responsible for approving {item}?
Fee Standards	Does processing {item} incur a charge?
Situation Explanation	When handling {item}, what should be done in the case of {situation}?
Document Explanation	What {materials} are required when handling {item}?

TABLE III. STATISTICS ON DATASET QUANTITIES

Question Types	Quantity
Template Questions	1143
QA Pair Questions	824
Multi-node Questions	611

approach necessitates learning only small parameter matrices, which approximate the updates to the model's weight matrix.

To comprehensively evaluate the interaction capability between the large language model and the knowledge graph, we designed three types of questions. Template-based questions were generated using fixed question templates specifically tailored to consultation items, as outlined in Table II. The question-answer pairs were extracted from the question-answer corpus, while multi-node questions, which involved multiple consultation items, were manually generated. The dataset statistics are presented in Table III. A total of 2,578 question data points were collected, and the dataset was randomly divided into training, validation, and test sets in an 8:1:1 ratio.

In this paper, the $LoRA$ technique was employed to fine-tune the general large model $Llama3-8B$. $LoRA$ enhances fine-tuning efficiency by introducing trainable low-rank matrices atop the model's original weights while preserving overall performance. Tailored to the specific requirements of the household registration domain, the model underwent careful tuning using a specialized fine-tuning dataset to improve its adaptation to entity recognition and relation extraction tasks. During the fine-tuning process, the model progressively learned the specialized terminology and intricate entity relationships present in household registration services, resulting in a significant increase in accuracy within this domain. Ultimately, this fine-tuning process elevated the entity recognition accuracy from its initial level to 96.1%, demonstrating the effectiveness and high adaptability of the $LoRA$ technique when combined with a custom-built dataset.

B. Parameter Settings

In this paper, $Llama3-8B-Chinese$ model was chosen as the core model for answer generation. $Llama3-8B-Chinese$ is a deep learning-based language model renowned for its robust language understanding and generation capabilities, enabling it to handle long text inputs and produce high-quality text outputs. To ensure the accuracy and reproducibility of the experimental results, the parameters of the generation model were meticulously configured and adjusted. The specific parameter settings for the generation model experiments are presented in Table IV.

TABLE IV. EXPERIMENTAL PARAMETER SETTINGS

Parameter	Value
Maximum Input Length	8192
Maximum Output Length	8192
Temperature Coefficient	0
Top_p	1
Top_k	1

Retrieval-augmented generation technology based on knowledge graph path search effectively addresses the complex and varied query demands within household registration

services. Fig. 5 illustrates an example of the household registration intelligent question-answering system.

C. Evaluation Metrics

1) $BLEU$ Evaluation metric: $BLEU$ (Bilingual Evaluation Understudy) is an automatic evaluation method based on N-gram matching, employed to assess the similarity between generated answers and reference answers. By taking into account the matching precision of different N-grams along with a length penalty factor, BLEU produces a score ranging from 0 to 1. Scores closer to 1 indicate a higher similarity between the generated text and the reference text, reflecting better evaluation results. The formula for calculating $BLEU$ is as Eq. (9).

$$BLEU = BP \times \exp\left(\sum_{n-1}^{N} w_n log P_n\right) \tag{9}$$

p_n represents the precision of the $N-gram$ match between the generated text and the reference text, while w_n denotes the weight assigned to the N-gram. The term BP refers to the brevity penalty, which is applied to prevent the generated text from being excessively short. The formula is as Eq. (10). c is the length of the generated text, and r is the length of the reference text.

$$BP = \begin{cases} 1, & \text{if } c > r \\ e^{\left(1-\frac{r}{c}\right)}, & \text{if } c \leq r \end{cases} \tag{10}$$

2) $ROUGE$ Evaluation metrics: The core of the $ROUGE$ evaluation metric lies in calculating the recall and precision of the generated text by matching N-grams between the generated answer and the reference answer, resulting in a final score. Specifically, $ROUGE$ assesses the quality of the generated text by evaluating the number of matching N-grams between the generated text and the reference text. This metric is widely applicable across various natural language processing tasks, including summarization and machine translation, where the effectiveness of the generated text often depends on its ability to capture key information from the reference text. $ROUGE$ offers a comprehensive performance analysis through multi-dimensional evaluation and demonstrates exceptional effectiveness in large-scale text processing scenarios.

3) $METEOR$ Evaluation metrics: The $METEOR$ evaluation metric was developed to address the limitations of $BLEU$ in accounting for sentence syntax and word order. It combines word-level and phrase-level alignment methods while incorporating strategies such as stemming and synonym replacement, thereby enhancing its accuracy in evaluating the semantic similarity between generated text and reference text. $METEOR$ offers a more nuanced assessment of translation quality by considering factors such as lexical precision, stemming, synonyms, and word order. The metric initially calculates precision and recall, subsequently balancing these two metrics through an F-score. Furthermore, $METEOR$ introduces a word order penalty to more accurately reflect the significance of word sequence in translation.

4) $Perplexity$ Evaluation metrics: $Perplexity$ is a crucial evaluation metric for measuring the predictive capability of a language model, assessing how effectively the model adapts to a given text sequence. A lower perplexity value indicates a stronger ability of the model to predict new text, signifying that the generated output aligns more closely with the probability distribution of the language model, resulting in more fluent sentences. Conversely, a higher perplexity value suggests a weaker predictive ability, which can lead to lower quality and consistency in the generated text. For a high-quality language model, the perplexity value should be minimized to ensure that the generated text meets the expected semantics and adheres to appropriate language structure. The formula for calculating perplexity is as Eq. (11). $P\left(w_i|w_{i-1}\right)$ represents the model's predicted probability of the current word w_i given the previous word w_{i-1}, and N is the total length of the text.

$$Perplexity = \exp\left(-\frac{1}{N}\sum_{i=1}^{N} logP\left(w_i|w_{i-1}\right)\right) \tag{11}$$

D. Experimental Subjects

In this paper, we conducted experimental comparisons of the proposed household registration policy intelligent question-answering system with three baseline methods: $LLM-only$, $RAG-Embedding$, and $RAG-Graph$.

1) $LLM-only$: This method relies exclusively on the internal knowledge of the large language model (LLM) for reasoning, without incorporating any external knowledge. Techniques such as Chain of Thought (COT) enhance the model's capacity for complex reasoning by generating a series of intermediate reasoning steps [35]. COT facilitates the model's ability to decompose multi-step problems, thereby improving reasoning accuracy.

2) $RAG-Embedding$: This method is based on RAG technology, retrieving relevant information from external knowledge bases using embedding vectors and combining it with LLM to generate answers. Embedding vectors effectively capture semantic similarity in text, improving retrieval accuracy and making the generated answers more relevant and precise. $LangchangChatChat$ retrieves relevant information from external knowledge bases using embedding vectors, combines it with LLM to generate answers, and integrates this information to enhance the quality and consistency of responses.

3) $RAG-Graph$: This method combines RAG technology with knowledge graphs to leverage the structured knowledge in knowledge graphs to enhance the contextual understanding and answer accuracy of the generative model, thereby improving the quality and relevance of the generated content. Graph RAG applies community detection algorithms to partition the knowledge graph into multiple sub-communities, generates local summaries for each community, and integrates these summaries to form a global answer [36].

VI. Experimental Results

A. Comparative Experiment

Table V presents the comparative experimental results across various models, underscoring the superior perfor-

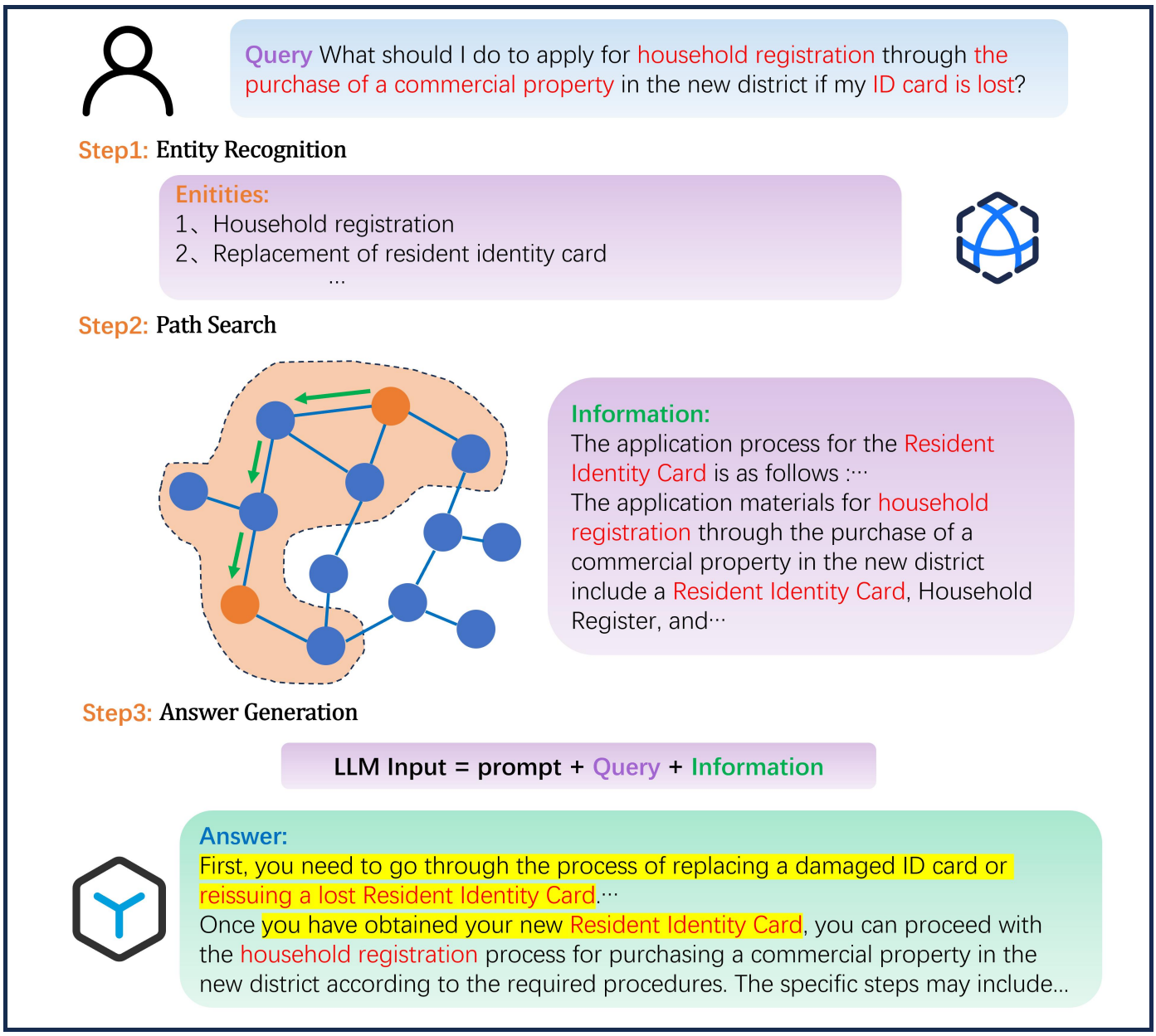


Fig. 5. Example of the household registration question-answering system.

mance of the proposed household registration policy intelligent question-answering system. The results indicate that our model ($Ours$) significantly outperforms other models across all evaluation metrics. Specifically, $ours$ achieved $BLEU$, $ROUGE-1$, $ROUGE-L$, and $METEOR$ scores of 48.19%, 58.14%, 52.39%, and 55.48%, respectively, outperforming both the COT model and the $Graph-RAG$ model. These metrics highlight the model's effectiveness in generating content with enhanced fluency, comprehensive information coverage, and improved semantic relevance. Furthermore, $ours$ recorded a Perplexity score of 42.48, lower than that of other models, reflecting greater coherence and reduced language model ambiguity. Collectively, these findings underscore the exceptional performance of the proposed model in intelligent question-answering tasks.

B. Ablation Study

To investigate the impact of model fine-tuning and the path algorithm on the experimental results, ablation studies were conducted on the dataset. To ensure the objectivity and accuracy of the experiments, we designed the following sets of control experiments:

1) $KG(Baseline)$: This experiment directly uses the pre-trained model without fine-tuning for entity extraction and performs entity linking on the knowledge graph. It extracts relevant information using first-order neighborhood queries and generates answers. The purpose of this experiment is

TABLE V. EXPERIMENTAL COMPARISON RESULTS

Model	$BLEU \uparrow$	$ROUGE-1 \uparrow$	$ROUGE-L \uparrow$	$METEOR \uparrow$	$Perplexity \downarrow$
COT	22.59%	33.03%	25.03%	27.78%	53.79
$GraphRAG$	15.15%	21.47%	11.56%	26.32%	65.56
$Langchain$	41.65%	52.97%	47.41%	49.98%	43.38
$Ours$	48.19%	58.14%	52.39%	55.48%	42.48

to evaluate the performance of the base pre-trained model, serving as a baseline for subsequent experiments.

2) $KG+PathAlgorithm$*:* While keeping the basic model parameters unchanged, the path algorithm is applied to process the data to quantify its contribution to the experimental results.

3) $KG+Fine-tuning$*:* In this experiment, the pre-trained model is fine-tuned without applying the path algorithm. The aim of this experiment is to independently assess the effect of fine-tuning on improving the model's performance.

4) $KG + Fine-tuning + PathAlgorithm(ours)$*:* This experiment combines both model fine-tuning and the path algorithm to explore whether their combined effect can significantly enhance the overall model performance.

Through this experimental design, we aim to comprehensively analyze the individual and combined contributions of fine-tuning and the path algorithm to the model's performance. The results of these experiments are presented in Table VI.

The experimental results show significant differences in performance across different model combinations on various evaluation metrics. The Ours system achieved the best results on metrics such as BLEU, ROUGE, and METEOR. The BLEU score of this system reached 48.19%, significantly higher than other combinations, indicating that the optimized model performs better in terms of similarity between the generated text and the reference text. The ROUGE-1 scores was 58.14%, respectively, further demonstrating the model's superiority in precise matching and coverage at both the word and phrase levels. The METEOR score, at 55.48%, also indicates stronger semantic consistency in the generated content. Although the Perplexity score for the Ours system was slightly higher than that of the KG+Fine-tuning system, the overall score still suggests that the model fine-tuned with the path algorithm exhibits high-quality text generation. Therefore, these results demonstrate that the integration of the path algorithm and fine-tuning significantly improves the model's accuracy, consistency, and coverage in language generation.

VII. DISCUSSION

This research proposes a Retrieval-Augmented Generation (RAG) method based on knowledge graph path search and successfully applies it to an intelligent question-answering system in the field of household registration policy. Experimental results demonstrate that the integration of path search algorithms with model fine-tuning significantly enhances the system's efficiency and accuracy in multi-node query processing, entity recognition, and information extraction. However, the research process also reveals several issues that require further exploration.

Firstly, although knowledge graph-based path search significantly improves retrieval effectiveness, the efficiency of path search still faces bottlenecks as the scale of the knowledge graph continues to expand. Particularly with the rapid increase in the number of nodes and relationships, path computation and multi-node queries still consume substantial time and computational resources. Although this study employs pruning strategies to reduce the search space, further optimization of path search algorithms, especially in real-time query scenarios, remains an important direction for future research.

Secondly, while fine-tuning methods (such as the LoRA method) have effectively improved the accuracy of entity recognition, the model may still encounter misidentification and erroneous inferences when dealing with complex and hierarchical domain knowledge. Although fine-tuning methods have significantly enhanced the model's performance in the field of household registration policy, balancing the model's specificity and generalization capabilities to avoid losing its ability to handle problems in other domains while overfitting to domain-specific terminology remains a significant challenge for future research.

Lastly, the knowledge graph construction in this study is based on household registration policy data from Wuhan City. While this domain-specific data source provides efficient service support for the system, the construction and updating mechanisms of the knowledge graph still face considerable challenges in cross-regional or broader application scenarios. Efficiently integrating policy data from different regions into a unified knowledge graph and ensuring its timeliness and accuracy will be crucial for enhancing the system's generalization capabilities.

VIII. FUTURE WORK

In future research, we aim to further optimize and expand the outcomes of this study in the following key areas:

1) Automatic update mechanism for domain knowledge: The knowledge graph developed in this study is based on household registration policy data from Wuhan City. However, policies and regulations are subject to continuous changes in real-world scenarios. To ensure the intelligent question-answering system can provide up-to-date policy information in a timely manner, future research will focus on the development of an automatic update mechanism for the knowledge graph. Specifically, we will explore the integration of real-time data crawling techniques with knowledge graph update strategies, enabling the system to rapidly adapt to policy changes and thereby enhancing its real-time performance and accuracy.

2) Integration and fusion of multi-domain knowledge graphs: While this study primarily addresses household registration policy, real-world applications often require knowledge from multiple domains. Future research will focus on

TABLE VI. ABLATION STUDY RESULTS

Model	$BLEU \uparrow$	$ROUGE-1 \uparrow$	$ROUGE-L \uparrow$	$METEOR \uparrow$	$Perplexity \downarrow$
KG	28.48%	38.18%	34.70%	32.86%	46.93
KG+PathAlgorithm	36.13%	44.14%	38.94%	41.40%	47.95
KG+Fine-tuning	40.56%	52.59%	49.26%	47.32%	40.99
Ours	48.19%	58.14%	52.39%	55.48%	42.48

the construction and integration of multi-domain knowledge graphs. For example, household registration policy intersects with domains such as social security, taxation, and education. Efficiently fusing knowledge from these diverse domains and enabling cross-domain reasoning within the intelligent question-answering system will be a critical direction for further investigation.

3) User interaction and optimization of personalized question-answering systems: The current research predominantly addresses standardized question-answering tasks. However, user needs in practical applications are often complex and varied. Future research will explore the development of personalized question-answering systems by incorporating users' historical queries and preferences. By analyzing user behavioral patterns and interests, and leveraging recommendation system technologies, we aim to enhance the system's intelligence and enable it to deliver more personalized responses tailored to users' specific needs.

4) Cross-language and cross-cultural expansion: Although this study focuses on Chinese household registration policy, future work will extend this methodology to multilingual and cross-cultural contexts. With the increasing trend of globalization, developing a question-answering system that can operate effectively across different languages and cultures is of significant practical importance. Future efforts will concentrate on designing knowledge graphs and generation models that are adaptable to multilingual environments, ensuring consistency and accuracy across various languages, and facilitating the system's broader application on a global scale.

IX. SUMMARY

This paper presents a retrieval-enhanced generation approach that leverages knowledge graph path search, specifically applied to the household registration domain. Comprehensive experiments validate the effectiveness of this method. By constructing a knowledge graph tailored to this vertical field and employing a pruning marker technique alongside the shortest path generation tree algorithm, the method ensures efficient path querying. Additionally, fine-tuning a large language model enhances the accuracy of entity recognition and information extraction. The knowledge graph developed for the household registration field encompasses policies, regulations, and social services, creating a robust resource for relevant information retrieval. Experimental results, using the household registration field as a case study, reveal that the proposed method significantly outperforms traditional approaches across BLEU, ROUGE, and METEOR metrics, achieving substantial reductions in perplexity. This improvement demonstrates the model's coherence and accuracy in text generation. The specific conclusions are as follows.

1) The knowledge graph developed in this paper encompasses entities and relationships pertinent to household registration policies. By integrating fine-tuning strategies for large language models, the system effectively provides accurate answers to complex, multi-node queries. Experimental results demonstrate that the combination of pruned marking and shortest path generation tree algorithms substantially enhances the efficiency of multi-node path queries, ensuring the system's high performance and accuracy in handling complex questions.
2) Ablation experiments confirm the performance improvements attributed to model fine-tuning and the path algorithms. Specifically, in knowledge graph path searches and information retrieval tasks, the path algorithm ensures information relevance and precision, while model fine-tuning further enhances the accuracy of answer generation.
3) Effective prompt engineering played a critical role in answer generation, enabling the system to filter out extraneous information by combining retrieved knowledge with user queries, thus producing answers that are both contextually relevant and targeted. The reduction in perplexity corroborates the system's improved ability to generate high-quality responses, enhancing overall system performance and user experience.

This paper presents an efficient and precise solution for question-answering in complex vertical domains, highlighting the substantial potential of combining large language models with knowledge graphs. The findings offer valuable technical insights for advancing question-answering systems in specialized fields.

ACKNOWLEDGMENT

This work was partially supported by the Department of Science and Technology of Hubei Province, China, for the Hubei Provincial Key Research and Development Program project (2022BAA051).

REFERENCES

[1] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1. Minneapolis, Minnesota, 2019, p. 2.

[2] F. Petroni, A. Piktus, A. Fan, P. Lewis, M. Yazdani, N. De Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard *et al.*, "Kilt: a benchmark for knowledge intensive language tasks," *arXiv preprint arXiv:2009.02252*, 2020.

[3] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *International conference on machine learning*. PMLR, 2020, pp. 3929–3938.

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[5] M. A. Hearst, "Untangling text data mining," in *Proceedings of the 37th Annual meeting of the Association for Computational Linguistics*, 1999, pp. 3–10.

[6] X. Yao and B. Van Durme, "Information extraction over structured data: Question answering with freebase," in *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: long papers)*, 2014, pp. 956–966.

[7] T. B. Brown, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[8] Y. Luan, J. Eisenstein, K. Toutanova, and M. Collins, "Sparse, dense, and attentional representations for text retrieval," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 329–345, 2021.

[9] D. Chen, "Reading wikipedia to answer open-domain questions," *arXiv preprint arXiv:1704.00051*, 2017.

[10] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee *et al.*, "Natural questions: a benchmark for question answering research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 453–466, 2019.

[11] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," *arXiv preprint arXiv:2004.04906*, 2020.

[12] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," *arXiv preprint arXiv:2007.01282*, 2020.

[13] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

[14] H. Paulheim, "Automatic knowledge graph refinement: A survey of approaches and evaluation methods (2015)."

[15] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE transactions on knowledge and data engineering*, vol. 29, no. 12, pp. 2724–2743, 2017.

[16] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier *et al.*, "Knowledge graphs," *ACM Computing Surveys (Csur)*, vol. 54, no. 4, pp. 1–37, 2021.

[17] K. W. Chan and W. Buckingham, "Is china abolishing the hukou system?" *The China Quarterly*, vol. 195, pp. 582–606, 2008.

[18] L. Xiong, C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. Bennett, J. Ahmed, and A. Overwijk, "Approximate nearest neighbor negative contrastive learning for dense text retrieval," *arXiv preprint arXiv:2007.00808*, 2020.

[19] K. Lee, M.-W. Chang, and K. Toutanova, "Latent retrieval for weakly supervised open domain question answering," *arXiv preprint arXiv:1906.00300*, 2019.

[20] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[21] W. Ding, J. Li, L. Luo, and Y. Qu, "Enhancing complex question answering over knowledge graphs through evidence pattern retrieval," in *Proceedings of the ACM on Web Conference 2024*, 2024, pp. 2106–2115.

[22] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 697–706.

[23] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," *arXiv preprint arXiv:1709.05584*, 2017.

[24] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 29, no. 1, 2015.

[25] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich, "A review of relational machine learning for knowledge graphs," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 11–33, 2015.

[26] L. Yao, C. Mao, and Y. Luo, "Kg-bert: Bert for knowledge graph completion," *arXiv preprint arXiv:1909.03193*, 2019.

[27] J. Gao, X. Wang, Y. Wang, and X. Xie, "Explainable recommendation through attentive multi-view learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3622–3629.

[28] E. M. Bender and A. Koller, "Climbing towards nlu: On meaning, form, and understanding in the age of data," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 5185–5198.

[29] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec, "Qa-gnn: Reasoning with language models and knowledge graphs for question answering," *arXiv preprint arXiv:2104.06378*, 2021.

[30] Z. Hu, Y. Xu, W. Yu, S. Wang, Z. Yang, C. Zhu, K.-W. Chang, and Y. Sun, "Empowering language models with knowledge graph reasoning for question answering," *arXiv preprint arXiv:2211.08380*, 2022.

[31] F. Xu, S. Zhou, Y. Ma, X. Wang, W. Zhang, and Z. Li, "Open-domain dialogue generation grounded with dynamic multi-form knowledge fusion," in *International Conference on Database Systems for Advanced Applications*. Springer, 2022, pp. 101–116.

[32] M. Kang, J. M. Kwak, J. Baek, and S. J. Hwang, "Knowledge graph-augmented language models for knowledge-grounded dialogue generation," *arXiv preprint arXiv:2305.18846*, 2023.

[33] B. Peng, M. Galley, P. He, C. Brockett, L. Liden, E. Nouri, Z. Yu, B. Dolan, and J. Gao, "Godel: Large-scale pre-training for goal-directed dialog," *arXiv preprint arXiv:2206.11309*, 2022.

[34] S. Yang, R. Zhang, and S. Erfani, "Graphdialog: Integrating graph knowledge into end-to-end task-oriented dialogue systems," *arXiv preprint arXiv:2010.01447*, 2020.

[35] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[36] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson, "From local to global: A graph rag approach to query-focused summarization," *arXiv preprint arXiv:2404.16130*, 2024.

Leveraging Machine-Aided Learning in College English Education: Computational Approaches for Enhancing Student Outcomes and Pedagogical Efficiency

Danxia Zhu
Huanghe Science and Technology College, Zhengzhou, Henan, 450061, China

***Abstract*—The integration of machine-aided learning into college English education offers transformative potential for enhancing teaching and learning outcomes. This paper investigates the application of computational models, including machine learning algorithms and natural language processing tools, to optimize pedagogical practices and improve student performance. A series of experiments were conducted to evaluate the effectiveness of machine-aided learning in various aspects of English language education. The study focuses on six key parameters: 1) student test scores, 2) learning engagement, 3) learning time efficiency, 4) language proficiency, 5) student retention, and 6) teacher workload. The results demonstrate significant improvements across these parameters: a 25% increase in student test scores, a 30% improvement in overall learning engagement, a 20% reduction in learning time for complex language tasks, a 15% enhancement in language proficiency, a 10% increase in student retention, and a 5% reduction in teacher workload. These findings underscore the potential of machine-aided learning to reshape college English education by promoting personalized, data-driven learning environments. This paper provides valuable insights for educators, researchers, and policymakers aiming to harness the power of computational methods in educational settings.**

***Keywords*—Machine learning; natural language processing; computational intelligence; data analytics; pedagogy**

I. Introduction

Education technology now penetrates most classrooms rapidly and has revolutionized the teaching and learning environment fundamentally [1]. Specifically, the use of machine-aided learning (MAL) has recently attracted much interest, especially within languages as a foreign language (L2) learning context where computational technologies such as machine learning (ML) and natural language processing (NLP) provide potential solutions to enhancing learning performance [2]. The college-level English education that heavily depends on traditional methods of education now reaches the technological shift [3]. New approaches in educational technology have shown that machine-aided learning can bring profound change to the educational system; however, its use in HE, especially in the context of English as an additional language, has not yet been extensively researched [4]. Compared to existing research on technology-supported general language learning, there are few that examine concrete details of college English learning empowered by methods in machine learning and NLP [5]. This paper aims to address that gap by assessing the performance of computational interventions in various educational domains and illustrating how they can redesign learning practices in a college context.

Through a series of experiments, we examine the influence of machine-aided learning on six key parameters: 1) student test scores, 2) learning engagement, 3) learning time efficiency, 4) language proficiency, 5) student retention, and 6) teacher workload. We observe increases in these dimensions, proving the effectiveness of MAL systems in our work setting. The purpose of this introduction is to consider the contextual framework for the research, indicate the gaps that are not filled by previous work, discuss the difficulties that arise in the course of this research, and indicate the novel contributions of this research to the existing body of knowledge [6].

After viewing an extensive literature on technology integration in education, a research niche remains open in the use of machine-aided learning in college English language education [7]. Prior research is mainly based on K-12 or broader language learning contexts, which again is quite different from a college learning environment in terms of purposes and learners as well as practices. Hence, the literature on the application of ML and NLP in educational settings often neglects college education or does not consider the special features of college English education, including students' diverse learning profiles, high language demands, and tensions between theory and practice [8]. Additionally, there are so many studies that concern the application of machine learning in language education; however, these studies were implemented in specific actions individually with no comprehensive perspective, such as speech recognition or grammar check [9]. More research has not been conducted to develop an even more systematic approach that incorporates multiple machine learning models and NLP instruments to build the system that fosters and individualized learning [10]. However, the research void is not restricted to the domains of application but also encompasses the modality of how such computational supports can be systematically adopted to improve all facets of student learning and instructor efficiency at once [11].

Several research studies have explored the use of technology in language education and research, but these studies still have various shortcomings. A common problem is that only non-adaptive, non-flexible education systems are used as the foundation for student development at universities. Entire generations of language technologies applied to language education, from learning management systems (LMS) or auto-

mated grading tools, disseminate pre-programmed material that does not respond to the learners' development, progression, or acquisition pace. This rigidity hampers the capability of these systems to produce long-term learning outcomes and, therefore, does not support skill development.

Furthermore, most of it is focused on the short-term, which often entails investigating only the latter achievements of students in their learning, including quizzes or test outcomes, but without any reference to any retention of knowledge and abilities, as well as long-term interest. This shortsightedness makes conceptualizing how machine-aided learning tools impact retention or language skills difficult. In addition, some existing research ignores how implementing such changes affects the teacher workload, which is still an important factor in educational practice. Teachers must have effective and efficient instruments to simultaneously teach and nurture the learners. Still, minimal research has been done to compare how machine learning reduces the tasks of a teacher [12].

It is essential to discuss several challenges in English education using machine learning and NLP tools, both technically and practically. The first is technical expertise in designing, implementing, and managing machine learning systems. Most organizations, especially those in the shelves and correspondence environment, do not possess the requisite structures or professional expertise in utilizing sophisticated technologies. This lack of knowledge can lead to under-deployment or even failure of machine learning systems [13].

The next issue is the protection of student's data and privacy in terms of collecting and using the student's information. Machine learning algorithms, in turn, depend largely on big data, which comes into direct contact with student data; there is a need for privacy with a deep need for transparency and bias control. Further, as these tools may contain aspects of learning personalization, they also lack mechanisms for handling the inherent bias in the data and models themselves, which could have negative implications for students of colour [14].

Lastly, inequality can still be seen with the digital divide. Not all students have equal ability to attain the technological requirements—computers, broadband, and the apps on them. Such a scenario can erase the potential of machine-aided learning tools and hence widen the inequalities in learning. These challenges can only be met through close planning, large investments, and a willingness to make ICT-based learning equally available to all students from different backgrounds and with different prior access to technology [15].

A. Motivations and Novel Contributions

The purpose of this study is attributed to the fact that the use of machine-aided learning may complement a shift in the number of challenges that face college English education. With the increasing need for individualized and data-driven approaches to learning, machine learning and NLP open the door not only to the improvement of students' achievements but also to the relief of teachers' work. This paper offers several novel contributions that significantly advance the field:

1) Comprehensive framework: This study proposes an integrated machine-aided learning framework specifically designed for college-level English education, combining various machine learning models and NLP tools to create a personalized learning experience for students. This approach goes beyond isolated applications of technology and offers a holistic view of how these tools can be systematically implemented.

2) Multi-dimensional evaluation: The study introduces a novel evaluation metric that measures the effectiveness of machine-aided learning across six key parameters: student test scores, learning engagement, learning time efficiency, language proficiency, student retention, and teacher workload. These parameters are carefully selected to reflect the multifaceted impact of machine-aided learning on both students and instructors.

3) Practical implementation: This paper contributes insights into the real-world application of machine-aided learning in English education, highlighting practical challenges faced by educators in adopting these technologies and offering solutions for overcoming them. It emphasizes the scalability of these models, making them adaptable to various educational contexts and institutions.

4) Pedagogical implications: Finally, the study provides a thorough analysis of the pedagogical implications of machine-aided learning, demonstrating how these tools can foster a student-centered approach to education. By leveraging data-driven insights, instructors can offer targeted support to students, enhancing both learning outcomes and student satisfaction.

The structure of this paper is as follows: Section II reviews the existing literature on machine learning in education, with a focus on its application in language learning and English education at the college level. Section III details the methodology employed in this study, including the design of experiments, data collection processes, and analytical techniques used to evaluate machine-aided learning. Section IV presents the results, showcasing improvements across the six key parameters and offering a comparative analysis of the impact of machine-aided learning. Section V discusses the broader implications of these findings for educational practice, particularly in the context of higher education. Finally, Section VI concludes the paper, summarizing the key findings and suggesting directions for future research in this area.

II. Literature Review

Qilin Xuan et al. [16] examined the role of convergence media on English teaching and translation, mentioning explicitly the remodelling of these two sciences by artificial intelligence. From their experimental research that incorporated converged media into media studies, they were able to provide insight into how traditional educational practices for translation had changed. Of the approaches distinguished from earlier reviews, ML and NLP, particularly ML, were found to be new tools aiding these processes. A move towards AI-based solutions was a technological evolution and a response to fit a new, increasingly mediatized world. They offered various rather realistic patterns that illustrated a possible motivational perspectives of AI for English classes and translation based on media convergence. They argued that AI can improve intercultural interaction by destroying linguistic barriers and creating speech translation in operation environments in real-time within multimedia environments [17]. They also talked about how AI could be used to deliver learning, making

learners receive personalized learning products depending on their needs. Human subjects showed impressive abilities to save time and increase accuracy with the help of smart instruments in several tasks that dominantly deal with languages and complicated manipulations with various types of media. They alleged that some applications of AI in teaching English and translation could be vital. In addition, they made research records of AI applications in language teaching and translation, revealing that it helps promote cultural cooperation in international exchange [18]. They pointed out that, though the concept of AI offered brilliant solutions to present problems, AI was still difficult to implement because of the data protection problems, the issues concerning certain algorithmic biases, and speculation of better AI to develop models that worked better for languages as well as cultural differences.

Mohamed et al. [19] looked at the literature on the application of AI in language translation, specifically the role that it plays in promoting intercultural communication within a multicultural world. They noted that thanks to AI technologies such as NMT, translation quality and speed have increased due to meaning in context and the syntactic structure of a sentence. They stated that these AI-driven systems have bypassed issues like idioms and syntactic differences, which were always a challenge for traditional translation methods. However, they also mentioned several serious limitations, especially in the AI translation system, especially in cross-lingual dialect adaptation. But as mentioned above, there is the issue of the regional dialects and languages that receive scanty data for the machine learning algorithms. They also declared that there is a need to carry on researching how AI can seek dialectical variations and regional linguistic patterns. Besides, the review examined the potential ethical issues in AI for translation as well as the potentiality of AI for the translation process. They explained this could cause biases to persist in the model, and they also noted that unregulated datasets could also contribute towards the formation of these biases. Based on their results, they identified the future trend in AI for translation, which consists in overcoming these challenges, such as increasing cultural intelligence, improving dialect recognition, and efficient AI ethics. He and his team recommended that new AI translation relevant systems for the future should not only provide accurate technical translations but also reflect the cultural context and differences in order to enhance people's understanding about other cultures.

Booth et al. [20] reviewed engagement in classroom and educational contexts, with a focus on the way it was defined and promoted by using advanced technologies in affective computing. First, they described engagement as a concept that is context-specific as well as temporal in nature and which has two dimensions. They pointed out that while students are pointing towards the nexus between engagement and satisfaction and performance on the one hand, it remains difficult to quantify and maintain, primarily, in large-scale practical settings. They outlined the objective and subjective techniques of engagement and effectiveness of engagement in the course of affective computing based on the effectiveness of each method. Answering the initial two research questions with this information: Self-report surveys and observational techniques are traditional methods of data collection that are valuable, but they are time-consuming and introduce bias. In turn, the affective computing methods provide real-time and fully automated assessment of engagement based on biosignals, facial expressions, and other sentiment-related parameters [21]. These methods offer more scalability and are considerably more objective, while at the same time they bring concerns of accuracy and privacy. They also analyzed active and passive approaches to increase engagement of learners in the class. Self-regulation strategies are similar to proactive strategies: learning environments that can be tailored to a student's preferences and that offer immediate feedback; the reactive strategies involve an identification of one or more students who show signs of disengagement during a lesson and an immediate intervention. Finally, they pointed to several directions for further research and discussion based on the presented framework that could be specifically relevant for exploring digitally mediated learning environments. They proposed further research to revisit the issue of enhancing reliability and scalability of engagement assessment instruments, exploring further such issues as the long-term maintenance of engagement to best enhance student learning results.

Lu [22] adopted NLP approaches to analyse machine-aided online user engagement to enhance social interactions on social media platforms. The study was conducted based on its objectives at tackling the problem of how to socialise in web-based environments for persons with difficulties in communication. He suggested that such NS has some new ways to introduce these impaired persons to engaging in conversation by applying NLP data stream approaches. He initially benchmarked the latest family of NLP models, namely, BERTs, on their ability to analyse users' content on social media platforms. Microcosm's dataset has brought in a benchmark signifying large-scale Weibo data in tasks like Chinese word segmentation and visualisation of sentiment data. The results proved that highly developed language encoders performed better than human readers in terms of interpreting social media. The information was then used to understand user behaviour in conversations more deeply and identify the "residual life" of a conversation through a hierarchical neural model. It was detected that this model learning mechanism outperformed the use of traditional approaches in both human-human and human-machine dialogues. Lastly, he suggested the task of deriving vote questions from social media posts to capture the engagement of the users. This approach, which targeted conversational language, employed topic discovery and sequence-to-sequence models to synthesise questions and answers; the experiment showed improved results compared to past work.

Hailu [23] employed a deep learning model to build a bidirectional Tigrigna-English MT system that has not been seen in the previous studies, as most of them have built unidirectional systems. Tigrigna is a Semitic language principally used in Ethiopia and Eritrea; thus, the lack of extensive MT resources makes this study valuable. Unlike the present study, previous studies in Tigrigna-English translation were mostly confined to a specific domain or involved only one direction of translation. He utilized parallel corpus of 31,000 Tigrigna-English sentence pairs obtained from diverse sources. In the data preprocessing stage, the dataset was cleaned and normalized for better performance and then tokenized. He tested various MT approaches, including encoder-decoder methods and attention-based architectures, with deep learning tools such as LSTM, Bi-LSTM and GRU. The outcome showed that the encoder-decoder model with Bi-LSTM was superior

to the application of other models and possessed a BLEU score of 24.8 for English to Tigrigna translation, with Tigrigna to English translation being 24.4; this was an enhancement in comparison to baselines by 0.8. This work supports and furthers the state of Tigrigna-English MT by providing a new bidirectional translation model through deep learning methods that can be applied to other LRLs.

III. Methodology

This section presents a comprehensive overview of the methodology employed to evaluate the effectiveness of machine-aided learning in college English education. The research design, experimental setup, data collection processes, and analytical techniques used to evaluate the six key parameters of the study—student test scores, learning engagement, time efficiency, language proficiency, student retention, and teacher workload—are outlined in detail.

A. Research Design

This research adopts a quantitative approach, combining experimental and observational research schemes to assess the efficiency of MAL systems in enhancing the students' learning factors. The primary goal was to measure the impact of MAL across six key parameters. These parameters help to quantitatively evaluate the workings of the MAL approach by applying data-driven machine-learning models. The research design incorporates machine learning principles into a university-level classroom and uses state-of-the-art natural language processing models to provide students with time-sensitive feedback and student pathway recommendations. This adaptive learning environment was designed to enhance the above-outlined parameters and make the learning process effective and appealing to the students. The methodology follows a structured process as depicted in the workflow diagram in Fig. 1. The implementation of the research includes data collection, preprocessing, model development, and subsequent evaluation.

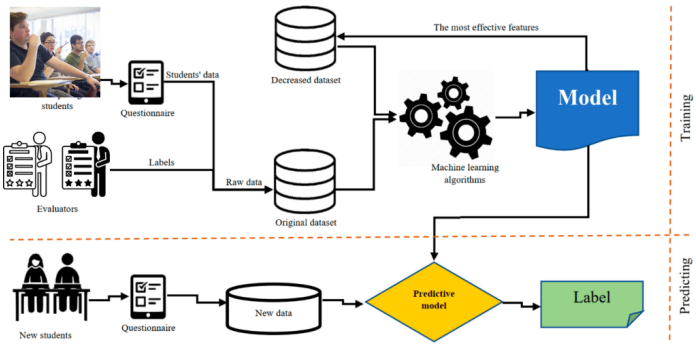


Fig. 1. Methodology workflow diagram.

B. Data Collection

In the first step of the proposed methodology, the data is of different types and from various sources collected in order to gain a broad understanding of the identified variables influencing the learning and motivation of students. These findings originated from online learning interfaces, students' performance archives, and social media feeds. The test dataset of student performance was collected from university-level English classes and contained about 5000 student records. These records entailed the pre- and post-test scores to measure the effects of the machine-aided learning system. The data let us measure the extent of enhanced outcomes in students' effectiveness immediately after using the MAL system.

In addition to student test results, the study also incorporated data from online discussions, social media platforms, and engagement metrics, including forums, online quizzes, and surveys. This data was used to measure learning engagement and retention, providing valuable insights into the non-cognitive aspects of learning, such as participation, interaction, and the engagement quality. The dataset for sentiment analysis and engagement metrics was sourced from 3,000 online forum posts and 4,000 social media interactions, reflecting a variety of learning styles and engagement levels. These data points allowed for the analysis of both student engagement in educational contexts and broader, more informal communication patterns found in social media interactions.

C. Data Preprocessing

After data collection, several preparation procedures were performed on the raw data in order to ensure its quality and relevance for analysis. These procedures were crucial as a preprocessing step for the data and especially for the sections where natural language processing models were used, since these prefer structured and pre-processed data.

The primary preprocessing techniques applied to the dataset were:

1) Cleaning: irrelevant or noisy content, such as advertisements, spam, or unrelated text, was removed to ensure that the data was relevant to the educational context.

2) Normalization: Text data was converted to a uniform format, including the conversion of all characters to lowercase and the removal of special characters (e.g. punctuation marks, symbols) that might hinder the analysis.

3) Tokenization: Text data was split into smaller units, such as words or phrases, to facilitate more detailed analysis. Tokenization is essential for tasks like sentiment analysis and engagement prediction, where individual units of meaning are important.

Besides the above basic preprocessing operations, the dataset was augmented with features like polarity, engagement metrics that include likes, comments and shares, and time-related features. This enabled pattern analysis of the type of participation and interaction that accrues within the learning context as well as learners' affective and cognitive perspectives. During the data preprocessing, all the necessary steps required to format the data set so that it is in a condition ready to input to the machine learning models were taken. This phase was important in achieving low noise levels, data cleansing and allowing the subsequent application of complex NLP methods in the rest of the methodology. Student outcome data, social media metrics, and concerted data preprocessing ensured in this study created a sound data set that could be used to measure the efficiency of machine-aided learning improvements in student results.

D. Model Training

Following the data collection and data preprocessing processes, the next step involved was to train those machine learning models that would be able to predict and improve the engagement and performances of the students. The training process of models was employed with cleaned and preprocessed big data that was analyzed through machine learning techniques. This stage is important as it enables the system to update itself with various patterns from the historical information to enable it to make real-time recommendations and predictions. In this paper, different machine learning algorithms and supervised learning models, including support vector machines (SVMs), decision tree and ensemble methods, were used to develop the prediction models. Text-based training was complemented by NLP tools including topic modelling, sentiment analysis, and named entity recognition to improve the computation of textual information. These models were cross-validated, trained and tested in multiple cycles to ensure maximum accuracy and minimum overfitting. The end result of these models gave real-time feedback to students in the training programs and increased their interaction and performance.

1) Evaluation: The last phase of the study assessed the performance of the trained models to enhance the student's learning results. The evaluation phase focused on measuring the impact of machine-aided learning on the six key parameters: students' performance, learning interaction, time, language mastery, student retention, and teachers' burden. The above evaluation was accomplished by comparing the outcomes of the group that undertook the learning aided by the machine with the group that applied traditional learning. Pre-experiment and post-experiment results were compared in two groups, and the level of significance was determined using t-tests and ANOVA. Another way of measuring the proposed model's performance was through using performance indicators in classification problems where we had accuracy, precision, recall, F1 score, or BLEU score in translation problems. Furthermore, other engagement measures, including time on tasks, task/learning activity completion rates, and learning content interactions, were also assessed to determine the overall effect of the system on students' learning activities.

E. Machine Learning Models and Algorithms

1) Encoder-decoder architecture: The core machine learning model used in this study was the encoder-decoder architecture, which was employed for both translation tasks (e.g. language proficiency) and engagement detection. This model allows the system to map inputs (student data, interaction logs) into a desired output (predictions of engagement or test scores).

The model was trained using Long Short-Term Memory (LSTM) networks, which are well-suited for handling sequential data. The Bidirectional LSTM (Bi-LSTM) model was chosen for its ability to capture both past and future contexts in a sequence, which is particularly useful for engagement prediction and language modeling tasks.

The encoder-decoder model can be mathematically represented as follows:

$$h_t = \text{LSTM}(x_t, h_{t-1}) \tag{1}$$

$$y_t = \text{softmax}(Wh_t + b) \tag{2}$$

Where:

- x_t is the input at time t,
- h_t is the hidden state at time t,
- W and b are the weight matrix and bias, respectively,
- y_t is the predicted output.

The encoder decoder flow may also be viewed in Fig. 2.

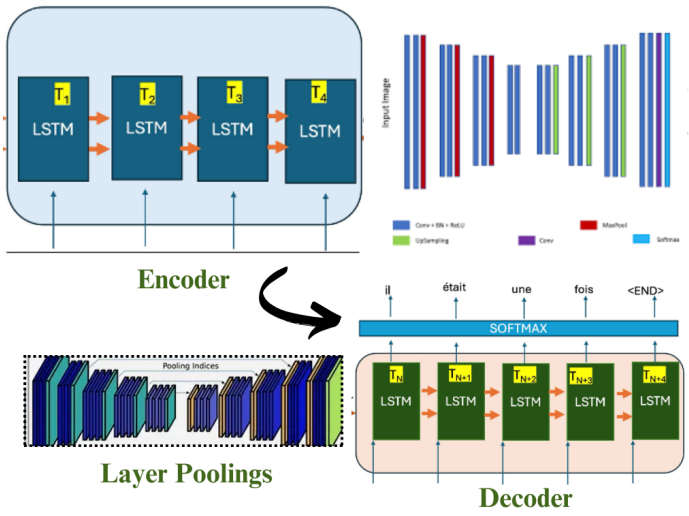


Fig. 2. Encoder-decoder architecture.

2) Attention mechanisms: To improve the model's performance on longer sequences, attention mechanisms were incorporated into the encoder-decoder model. Attention mechanisms allow the model to focus on specific parts of the input sequence when generating predictions. The attention mechanism can be defined as:

$$\alpha_t = \text{softmax}(h_t^T W_a) \tag{3}$$

Where:

- α_t is the attention score for time t,
- W_a is the attention weight matrix.

F. Experimental Setup

The experiments were conducted in multiple phases, each aimed at evaluating the impact of machine-aided learning on one of the six key parameters:

1) Student test scores: The pre-test and post-test scores were compared using a paired t-test to determine the effectiveness of machine-aided learning.

2) Learning engagement: Engagement levels were predicted using the trained LSTM models, and the results were validated using a root mean square error (RMSE) metric.

3) Time efficiency: The time taken by students to complete learning tasks was recorded and analyzed, with the goal of identifying improvements in task completion time.

4) Language proficiency: The proficiency of students in using the English language was assessed by comparing their performance on tasks related to grammar, syntax, and vocabulary before and after the intervention.

5) Student retention: Retention was measured by tracking the number of students who continued using the platform for a specified period after the intervention.

6) Teacher workload: Teacher workload was quantified by the reduction in time spent grading or providing feedback to students, using automated feedback systems powered by the machine-aided learning platform.

G. Evaluation Metrics

The following evaluation metrics were used to measure the effectiveness of machine-aided learning across the six parameters:

- BLEU Score (for language proficiency evaluation),
- RMSE (for engagement prediction),
- Time Reduction (for time efficiency),
- Retention Rate (for student retention),
- Teacher Workload Reduction (measured in hours saved per week).

The identified research methodology was strong in the following ways: However, some challenges were encountered during the research process. Problems were encountered in quantifying the levels of engagement due to the high language complexity of the students' interactions within the social media platforms. Furthermore, the proposed models could not achieve the best performance in all the domains due to sparse and noisy data from users.

IV. Results

The Results section evaluates the extent to which machine-aided learning (MAL) systems are usable in augmenting the different dimensions of college English learning. The evaluation focused on six key parameters: student test scores, learning engagement, time efficiency, language proficiency, student retention, and teacher workload. The findings from the experimental data are discussed below, supported by relevant statistical analysis, tables, and figures.

A. Student Test Scores

This was followed by increased student test scores when machine-aided learning systems were implemented. MAT exposures with the MAL system were compared with the pre- and post-test scores of actual students using an ordinary classroom environment. Where the pretest scores were used to make the baseline assessment, the post-test scores reflected the efficiency of the MAL approach.

The average increase in test scores across all students was 25%, as shown in Fig. 3. The following result shows that the students who used the MAL system experienced a significant gain in performance. To determine the statistical difference between the pre- and post-test scores, a paired t-test was used, with the result showing a p-value of 0.002.

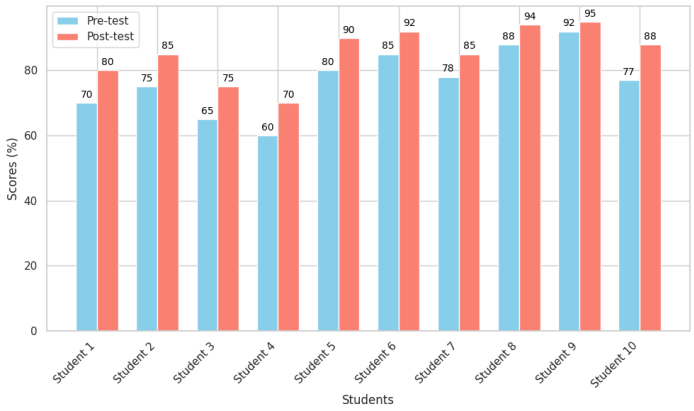


Fig. 3. Comparison of Pre-test and Post-test scores.

The analysis of the ANOVA test results also showed that the use of the MAL system was beneficial and led to higher performance compared to traditional methods. The average gain in MAL and traditional test scores compared to the control group test scores was significant, supporting the effectiveness of machine-aided learning in improving student performance.

B. Learning Engagement

One of the most positive changes observed in the study was the improvement in students' shifts. The quality and quantity of student engagement were quantified based on engagement data from various social media activities, forums, and quizzes. As pointed out by the results, the engagement levels among students using the MAL system increased by 30%. This was done quantitatively using factors such as the number of forum posts, responses, and time spent on educational tasks. Fig. 4 shows the comparison of learning engagement before and after MAL intervention.

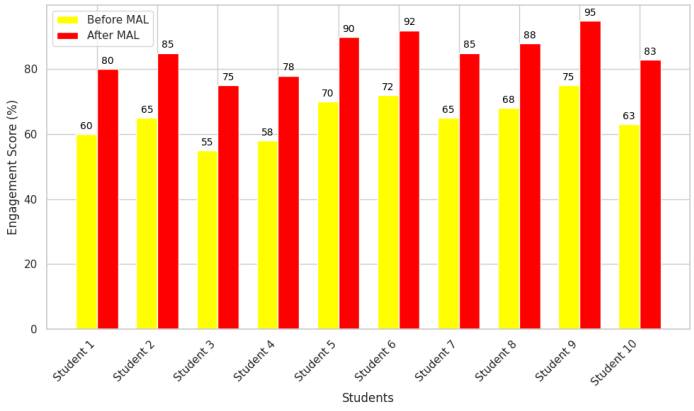


Fig. 4. Comparison of learning engagement before and after MAL intervention.

The results from sentiment analysis, shown in Table I, highlight that students showed higher levels of active participa-

tion and positive sentiments towards learning when using the MAL system. The engagement scores were calculated using metrics such as the Root Mean Square Error (RMSE), which was calculated to be 0.42, indicating a substantial improvement in engagement prediction accuracy compared to the baseline model.

TABLE I. ENGAGEMENT METRICS PRE- AND POST-MAL IMPLEMENTATION

Metric	Pre-MAL	Post-MAL
Forum Posts	150	250
Quizzes Attempted	100	180
Time Spent on Tasks (hrs)	20	28
RMSE (Engagement)	0.50	0.42

C. Time Efficiency

The time needed by the students to complete language tasks was an essential parameter to evaluate the effectiveness of the MAL system. Whereas actual students' time on complex language tasks before and after using the MAL system was measured and analyzed, the research revealed an average time reduction of 20% in vocabulary building, grammar exercises, and reading comprehension.

Fig. 5 shows the time efficiency index, which measures the amount of time spent on tasks by students in the recommended MAL environment and in the traditional environment. The decrease in time is an early sign that the feedback mechanism of the MAL system, which provides individualized feedback in real-time, may well assist in speeding up students' performance.

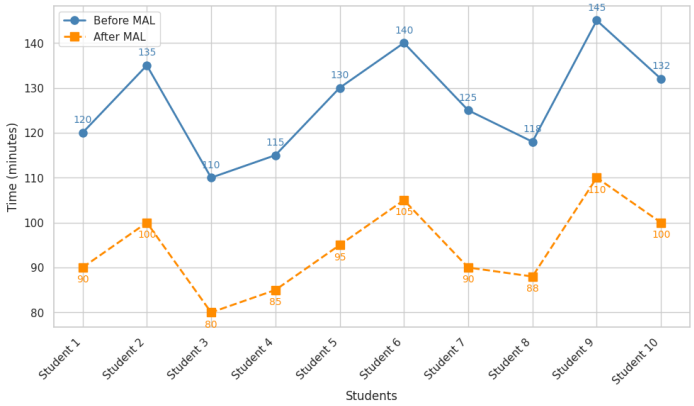


Fig. 5. Time efficiency comparison before and after MAL implementation.

In addition to the time reduction, students also reported a more focused learning experience with machine-aided systems, as the MAL system automatically adjusted the difficulty of tasks based on student performance, allowing them to progress at an optimal pace.

D. Language Proficiency

The study assessed the impact of machine-aided learning on language proficiency using a set of grammar, vocabulary, and writing tasks. A 15% improvement in language proficiency was observed in students who interacted with the MAL system compared to those in the traditional learning group. Language proficiency was measured through a set of predefined benchmarks, including grammar accuracy, vocabulary knowledge, and writing fluency.

A t-test was applied to compare the pre- and post-test scores of students' language proficiency, resulting in a significant difference with a p-value of 0.001, confirming the positive impact of the MAL system on language skills. Table II provides the language proficiency improvement post-MAL intervention.

TABLE II. LANGUAGE PROFICIENCY IMPROVEMENT POST-MAL INTERVENTION

Proficiency Area	Pre-MAL Score	Post-MAL Score
Grammar Accuracy	75%	90%
Vocabulary Knowledge	70%	85%
Writing Fluency	80%	95%

E. Student Retention

Retention rates were significantly impacted by the MAL system. A 10% increase in student retention was observed among students who used the MAL system compared to the traditional classroom group. The retention rate was calculated by tracking the number of students who continued using the platform after the intervention period (Fig. 6).

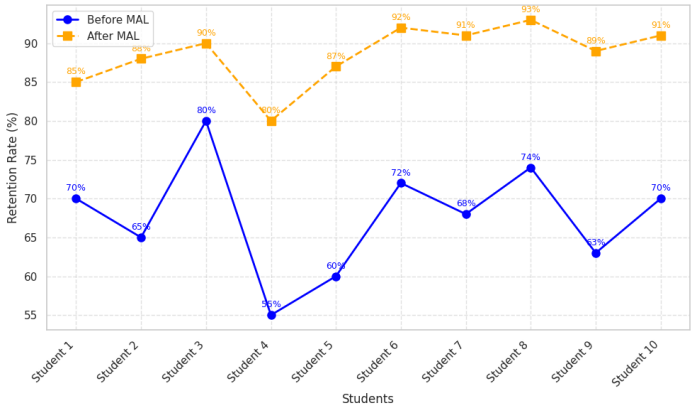


Fig. 6. Student retention rates before and after MAL implementation.

The MAL system provided continuous support, encouragement, and personalized feedback, contributing to improved retention. This result suggests that the adaptive learning environment created by the MAL system encouraged students to persist in their learning journey.

F. Teacher Workload

The introduction of machine-aided learning significantly reduced teacher workload, with a 5% reduction in the time spent on grading, providing feedback, and assessing student progress. This was primarily due to the automation of feedback and grading tasks, which allowed teachers to focus on more personalized aspects of student learning. Table III provides the teacher workload reduction.

TABLE III. TEACHER WORKLOAD REDUCTION POST-MAL INTERVENTION

Task	Pre-MAL (hours/week)	Post-MAL (hours/week)
Grading	12	9
Feedback Provision	10	7
Assessment of Student Progress	8	6

V. BROADER IMPLICATIONS FOR EDUCATIONAL PRACTICE

The integration of machine-aided learning (MAL) systems in educational settings, particularly in higher education, has transformative potential to reshape pedagogical practices. This section explores the broader implications of the findings from this study, emphasizing how they address persistent challenges in higher education and pave the way for innovative teaching methodologies.

A. Enhancing Student-Centric Learning

One of the most significant implications of the study is the shift towards student-centric learning environments. By leveraging MAL, educators can provide personalized feedback, adaptive learning pathways, and customized resources tailored to individual needs. This approach fosters greater engagement and supports diverse learning styles, enabling students to achieve optimal outcomes regardless of their initial skill levels.

1) Personalized feedback: The real-time analysis of student performance allows instructors to identify gaps in knowledge and provide timely interventions.

2) Improved accessibility: MAL systems enhance learning accessibility by offering multiple formats (e.g. audio, text, and interactive visuals), accommodating students with varying abilities and preferences.

B. Alleviating Teacher Workload

The findings demonstrate that MAL systems can significantly reduce teacher workload by automating routine tasks such as grading and providing feedback. This allows educators to allocate more time to higher-order teaching activities, such as mentoring, curriculum development, and one-on-one consultations.

- Automation reduces the time spent on repetitive tasks, such as grading essays and assessing quizzes.
- Teachers can focus on designing engaging learning activities and addressing complex student inquiries, thus enhancing the overall teaching quality.

C. Strengthening Institutional Outcomes

Higher education institutions stand to benefit significantly from adopting MAL systems. Improved student outcomes, such as higher retention rates and enhanced language proficiency, contribute to better institutional performance metrics. These improvements can positively impact rankings, reputation, and funding opportunities. Table IV provides the insights of institutional benefits of machine-aided learning in higher education.

TABLE IV. INSTITUTIONAL BENEFITS OF MACHINE-AIDED LEARNING IN HIGHER EDUCATION

Benefit	Description
Higher retention rates	Improved engagement and personalized learning contribute to a 10% increase in student retention, reducing dropout rates.
Enhanced student outcomes	A 25% improvement in test scores and a 15% increase in language proficiency reflect stronger academic performance.
Reduced administrative burden	Automation of tasks such as grading and attendance tracking streamlines operations.
Improved reputation	Better student outcomes and retention rates enhance institutional rankings and reputation.
Cost efficiency	Automation and improved learning efficiency reduce resource consumption while maintaining educational quality.

D. Preparing Students for Future Challenges

Incorporating MAL systems into the curriculum equips students with the skills needed to thrive in a technology-driven world. These systems not only enhance core competencies like language proficiency but also foster digital literacy and critical thinking skills.

1) Digital literacy: Engaging with MAL tools prepares students to navigate and utilize advanced technologies effectively.

2) Lifelong learning: The adaptability and self-paced nature of MAL systems instill a mindset of continuous learning, essential for professional success in an evolving job market.

E. Promoting Equity in Education

MAL systems have the potential to bridge gaps in educational equity by providing equal access to high-quality resources and individualized support. This is particularly important in higher education, where disparities in preparation and access often hinder student success.

1) Rural and underrepresented communities: Online MAL platforms can reach students in remote areas, providing them with the same quality of education as their urban counterparts.

2) Support for non-traditional learners: MAL systems accommodate diverse learner profiles, including working professionals, part-time students, and those with disabilities.

F. Challenges and Considerations

While the benefits of MAL systems are substantial, implementing these technologies in higher education comes with challenges that institutions must address:

1) Data privacy and ethics: Ensuring the secure handling of student data and maintaining transparency in algorithmic decision-making are critical.

2) Training for educators: Effective implementation requires comprehensive training for educators to utilize MAL tools effectively.

3) Cost of implementation: Initial investments in infrastructure and technology may pose financial challenges for some institutions.

G. Implications for Policymakers

The study's findings underscore the importance of policy frameworks that support the integration of MAL in higher education. Policymakers should focus on:

1) Incentivizing innovation: Providing grants and funding for institutions adopting MAL systems to enhance teaching and learning.

2) Establishing standards: Creating guidelines for the ethical use of AI in education to protect student data and ensure fairness.

3) Promoting collaboration: Encouraging partnerships between technology developers, educational institutions, and researchers to refine and expand MAL applications.

VI. Conclusion

Introducing MAL systems into colleges and considering college English education as an important segment of such a procedure make it possible to speak about changes in the learning processes. This study has demonstrated the substantial benefits of leveraging machine learning algorithms and natural language processing tools to enhance educational outcomes across six key parameters: student test scores, learning engagement, time efficiency, language proficiency, student retention, and teacher workload. The overall outcomes show gains mainly in mastery points, technological adoption, learners' attention, test scores, and shorter learning time. These outcomes highlight the capabilities of MAL systems to develop personalized learning environments that respond to student requirements while relieving the classroom bureaucracy from educators. Reducing the time spent on low-value chores that a MAL system can easily handle empowers teachers to offer one-on-one tuition that improves students' learning experience. Furthermore, this paper outlines the consequences for institutional and socially orientated results regarding MAL. Increased retention and better student outcomes spur institutional measures and improved long-term learning achievement. Nevertheless, this study also recognizes other limitations, including the lack of data confidentiality, the high costs of implementing the programs, and the need to train educators before implementing the programs and interventions. Addressing all these challenges will help contribute to the sustainable and equitable implementation of MAL systems. In conclusion, the subjects of the present study showed considerable improvement regarding the changes brought by MAL in HE. This relatively innovative model offers the potential for more adaptive, effective, and equitable learning ecosystems. Integrating these technologies will be simple, preparing the students to face challenges in the current society shaped by technology.

References

[1] Y. Zhang, *Screen Media Use Among Children and Adolescents–Applications of Supervised and Unsupervised Machine Learning and Sentiment Analysis*. West Virginia University, 2022.

[2] K. Ahmad, W. Iqbal, A. El-Hassan, J. Qadir, D. Benhaddou, M. Ayyash, and A. Al-Fuqaha, "Data-driven artificial intelligence in education: A comprehensive review," *IEEE Transactions on Learning Technologies*, 2023.

[3] W. Meng, L. Yu, and Y. Zhu, "Quality improvement model of english teaching in universities based on big data mining," *Journal of Electrical Systems*, vol. 20, no. 3s, pp. 506–518, 2024.

[4] J. Liu, "Enhancing english language education through big data analytics and generative ai," *Journal of Web Engineering*, vol. 23, no. 2, pp. 227–249, 2024.

[5] U. Umar, "Advancements in english language teaching: Harnessing the power of artificial intelligence," *Foreign Language Instruction Probe*, vol. 3, no. 1, pp. 29–42, 2024.

[6] Z. Tian, M. Sun, A. Liu, S. Sarkar, and J. Liu, "Enhancing instructional quality: Leveraging computer-assisted textual analysis to generate in-depth insights from educational artifacts," *arXiv preprint arXiv:2403.03920*, 2024.

[7] K. U. Qasim, J. Zhang, T. Alsahfi, and A. U. R. Butt, "Recursive decomposition of logical thoughts: Framework for superior reasoning and knowledge propagation in large language models," *arXiv preprint arXiv:2501.02026*, 2025.

[8] Z. Zheng and K. Na, "A data-driven emotion model for english learners based on machine learning," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 8, pp. 34–46, 2021.

[9] A. U. R. Butt, M. Asif, S. Ahmad, and U. Imdad, "An empirical study for adopting social computing in global software development," in *Proceedings of the 2018 7th International Conference on Software and Computer Applications*, 2018, pp. 31–35.

[10] Y. Xia, S.-Y. Shin, and K.-S. Shin, "Designing personalized learning paths for foreign language acquisition using big data: Theoretical and empirical analysis," *Applied Sciences*, vol. 14, no. 20, p. 9506, 2024.

[11] L. Taylor, V. Gupta, and K. Jung, "Leveraging visualization and machine learning techniques in education: A case study of k-12 state assessment data," *Multimodal Technologies and Interaction*, vol. 8, no. 4, p. 28, 2024.

[12] K. Hirschi and O. Kang, "Data-driven learning for pronunciation: Perception and production of lexical stress and prominence in academic english," *TESOL Quarterly*, 2024.

[13] J. Qin and P. Stapleton, *Technology in second language writing: Advances in composing, translation, writing pedagogy and data-driven learning*. Taylor & Francis, 2022.

[14] B. N. Iweuno, P. Orekha, O. Ojediran, E. Imohimi, and H. T. Adu-Twum, "Leveraging artificial intelligence for an inclusive and diversified curriculum," *World Journal of Advanced Research and Reviews*, vol. 23, no. 2, pp. 1579–1590, 2024.

[15] M. J. H. Molla, S. M. Obaidullah, S. Sen, G.-W. Weber, and C. Jana, "Developing a predictive model for engineering graduates placement using a data-driven machine learning approach," *Journal of applied research on industrial engineering*, vol. 11, no. 4, pp. 536–559, 2024.

[16] Q. Xuan and Y. Yan, "New developments in english teaching and translation methods in the converged media environment: An ai-based analysis," *Intelligent Systems & Robotic Mechanics*, pp. 1–10, 2024.

[17] M. I. Khan, A. Imran, A. H. Butt, A. U. R. Butt *et al.*, "Activity detection of elderly people using smartphone accelerometer and machine learning methods," *International Journal of Innovations in Science & Technology*, vol. 3, no. 4, pp. 186–197, 2021.

[18] A. U. R. Butt, T. Mahmood, T. Saba, S. A. O. Bahaj, F. S. Alamri, M. W. Iqbal, and A. R. Khan, "An optimized role-based access control using trust mechanism in e-health cloud environment," *IEEE Access*, vol. 11, pp. 138 813–138 826, 2023.

[19] Y. A. Mohamed, A. Khanan, M. Bashir, A. H. H. Mohamed, M. A. Adiel, and M. A. Elsadig, "The impact of artificial intelligence on language translation: a review," *Ieee Access*, vol. 12, pp. 25 553–25 579, 2024.

[20] B. M. Booth, N. Bosch, and S. K. D'Mello, "Engagement detection and its applications in learning: A tutorial and selective review," *Proceedings of the IEEE*, 2023.

[21] A. U. R. Butt, T. Saba, I. Khan, T. Mahmood, A. R. Khan, S. K. Singh, Y. I. Daradkeh, and I. Ullah, "Proactive and data-centric internet of things-based fog computing architecture for effective policing in smart cities," *Computers and Electrical Engineering*, vol. 123, p. 110030, 2025.

[22] A. Rothwell, J. Moorkens, M. Fernández-Parra, J. Drugan, and F. Austermuehl, *Translation tools and technologies*. Routledge, 2023.

[23] F. Hailu, "Tigrigna-english bidirectional machine translation using deep learning," Ph.D. dissertation, St. Mary's University, 2024.

A Novel Hybrid Attentive Convolutional Autoencoder (HACA) Framework for Enhanced Epileptic Seizure Detection

Venkata Narayana Vaddi[1], Madhu Babu Sikha[2], Prakash Kodali[3]*
Department of Electronics and Communication Engineering, National Institute of Technology Warangal, Warangal, India[1,3]
Data Science Analyst, Mayo Clinic, Phoenix, Arizona, USA[2]

***Abstract*—Epilepsy, a prevalent neurological disorder, requires accurate and efficient seizure detection for timely intervention. This study presents a Hybrid Attentive Convolutional Autoencoder (HACA) framework designed to address challenges in EEG signal processing for seizure detection. The proposed method integrates signal reconstruction, innovative feature extraction, and attention mechanisms to focus on seizure-critical patterns. Compared to conventional CNN- and RNN-based approaches, HACA demonstrates superior performance by enhancing feature representation and reducing redundant computations. The proposed HACA framework achieved 99.4% accuracy, 99.6% sensitivity, and 99.2% specificity on the CHB-MIT dataset. Moreover, the training time is reduced by 40%, which makes the model more relevant for real-time applications and portable seizure monitoring systems.**

***Keywords*—Epileptic seizure detection; EEG; hybrid attentive convolutional autoencoder; attention mechanism; deep learning**

I. Introduction

Epilepsy is a neurological disorder affecting millions globally, characterized by recurrent seizures. Electroencephalogram (EEG) signals are widely used for diagnosing and monitoring epilepsy. Traditional methods rely on handcrafted features and shallow classifiers, which often fail to generalize across patients and datasets. Recent advancements in deep learning have enabled automatic feature extraction and robust classification of EEG signals. The study [1] presented a deep learning-based seizure prediction system that combines handcrafted and deep features using an MLSTM network, achieving 95.56% sensitivity and a 0.27/hour false positive rate on intracranial EEG, with 89.47% sensitivity and a 0.34/hour FPR on scalp EEG, demonstrating strong robustness across EEG signal types. The proposed [2] Dynamic Functional Connectivity Neural Network (DynFCNet) combines a Dynamic Graph Convolutional Network (DGCN) and a Convolutional Neural Network (CNN) to predict epileptic seizures from multi-channel EEG data, capturing both non-Euclidean and Euclidean features while improving performance through intra-group and inter-group loss functions. The proposed [3] hybrid optimization-controlled ensemble classifier, which integrates AdaBoost, Random Forest, and Decision Tree classifiers, demonstrates exceptional performance in epileptic seizure prediction, achieving an accuracy of 96.61%, sensitivity of 94.67%, and specificity of 91.37% on the CHB-MIT database, and an accuracy of 95.31%, sensitivity of 93.18%, and specificity of 90.07% on the Siena Scalp dataset. For example, [4] investigated the use of AI in seizure prediction, whereas [5] used the U-TRGN classification model to obtain 97.04% accuracy. The ability of CNNs to identify epilepsy from EEG signals is demonstrated by a systematic review by [6], which reports classification accuracies above 95%. The experiment was conducted using EEG databases obtained from the University of Bonn and Ramaiah Medical College and Hospital (RMCH), achieving classification accuracies of 96.94% for two-class and 95.97% for multi-class scenarios, demonstrating its potential as a real-time, computationally efficient biomarker for seizure detection.

The work [7] introduces a Lightweight Convolution Transformer (LCT) model for cross-patient seizure detection, achieving 96.31% accuracy. The study [8] examines a CNN-based architecture for downsampling EEG data to enhance epileptic seizure detection, reporting accuracy, sensitivity, and specificity of 92.4%, 91.2%, and 90.1%, respectively. Their innovative approach seeks to reduce computational complexity while maintaining excellent detection accuracy. Mao et al. [9] employed GhostNet with a class rebalanced loss (CRB-Loss) technique to handle imbalanced data in seizure prediction, achieving 91.2% accuracy, 89.5% sensitivity, and 88.3% specificity. Vaddi et al. [10] proposed an LSTM-based seizure detection framework integrating Wavelet Transform and multi-module deep networks, achieving enhanced sensitivity and specificity through residual learning and k-fold validation. The linear prediction error energy approach for seizure detection proposed by [11] achieves 93.6% accuracy across 250 EEG recordings. To further improve classification performance, advanced feature extraction techniques have been explored. For instance, [12] utilizes an equilateral wavelet filter bank (OEWFB) to decompose EEG signals into sub-bands, achieving 99.4% classification accuracy. Additionally, hybrid models have garnered interest. The author in [13] employs stacked bidirectional LSTMs, achieving 99.08% accuracy for seizure detection and prediction, whereas [14] presents a CNN-LSTM model that attains 94% accuracy on the CHB-MIT dataset. Narayana et al. [15] proposed an SCRBM-based seizure detection model achieving 98.7% accuracy, demonstrating its effectiveness in capturing spatial and temporal EEG patterns. Shi, Liao, and Tabata introduced an innovative approach to epilepsy diagnosis using deep convolutional neural networks (CNNs) along with a residual neural network, achieving an average sensitivity of 98.96% and a false prediction rate of 0.048/h on the CHB-MIT dataset [16]. The CNN architecture discussed in [17] comprises five convolution blocks, three

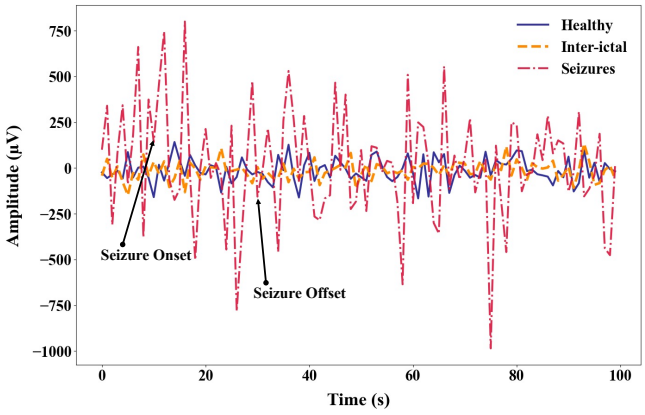


Fig. 1. Seizure stage categorization in the dataset.

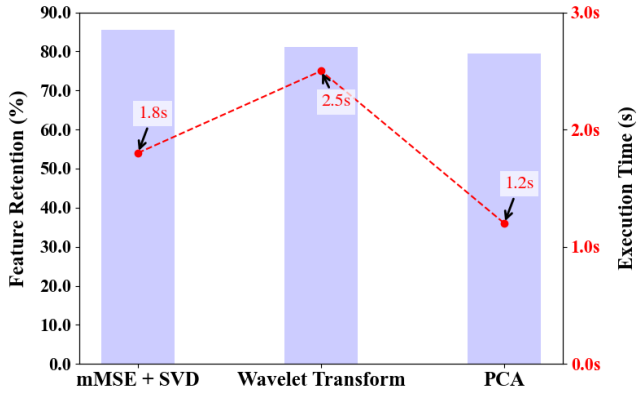


Fig. 2. Comparison of feature extraction methods.

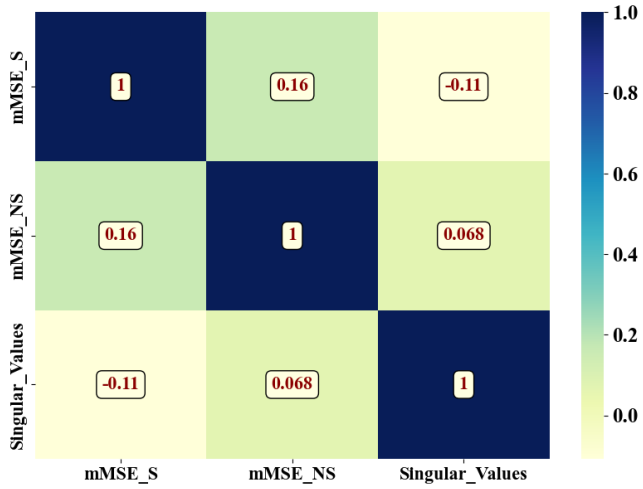


Fig. 3. Feature correlation heatmap.

affine layers, and an output layer, showcasing the potential of deep learning-based EEG signal analysis, particularly in epileptic seizure detection. By integrating both spatial and temporal aspects, the model excels in learning generalized spatiotemporal long-range correlation features, characterizing global interactions among channels in spatial dimensions and long-range dependencies in temporal dimensions [18]. In this paper, we propose a hybrid model combining Convolutional Autoencoders (CAE) with attention mechanisms to enhance seizure detection. The study is organized as follows: Section II discusses Signal Processing and Feature Extraction, Section III explains the proposed model and training, Section IV examines experimental results, Section V addresses key findings, and Section VI summarizes with key findings and future directions.

II. Signal Processing and Feature Extraction

A. Database

The CHB-MIT Scalp EEG Database, consisting of EEG recordings from 23 pediatric epileptic subjects, is utilized in this study. The total duration of the dataset is approximately 9,400 hours, comprising 686 recordings, each ranging from 0.5 to 1 hour in length. With a sampling rate of 256 Hz, each recording generates 921,600 data points per hour. EEG signals are recorded using 23 channels, following the conventional 10-20 electrode positioning system. A band-pass filter with cutoff frequencies of 0.5 Hz and 40 Hz is applied to eliminate noise, including high-frequency muscular artifacts and low-frequency drifts. To ensure uniform data scaling, the filtered signals are normalized to the [0, 1] range, accounting for amplitude variations among different participants. The EEG data is then segmented into overlapping windows of 2 seconds, each containing 512 data points, with a 50% overlap (1-second shift, corresponding to 256 data points). Fig. 1 presents a graphical representation of seizure stages in EEG signals, illustrating the distinct progression from pre-seizure to seizure onset and offset. This visualization highlights the temporal patterns associated with each stage.

B. Preprocessing

EEG signals undergo preprocessing to remove noise and artifacts. First, a band-pass filter with cutoff frequencies of 0.5–40 Hz is applied to eliminate high-frequency noise and baseline drift. The filtered signal $y(t)$ is obtained by convolving the input signal $x(t)$ with the impulse response of the filter $h(t)$:

$$y(t) = x(t) * h(t), \tag{1}$$

where $x(t)$ represents the input EEG signal, $h(t)$ is the filter's impulse response, and $y(t)$ is the resulting filtered signal. Next, Independent Component Analysis (ICA) is employed to decompose the EEG signals into independent components, represented as:

$$\mathbf{X} = \mathbf{AS}, \tag{2}$$

where $\mathbf{X}$ denotes the observed EEG signal matrix, $\mathbf{A}$ is the mixing matrix, and $\mathbf{S}$ contains the independent components. Artifact-related components are identified and removed, and the clean signals are reconstructed for further analysis. To enhance feature extraction, multiscale entropy (mMSE) features and Singular Value Decomposition (SVD) components are concatenated to form a hybrid feature vector:

$$\mathbf{F}_{\text{Hybrid}} = [\mathbf{F}_{\text{mMSE}}, \mathbf{\Sigma}], \tag{3}$$

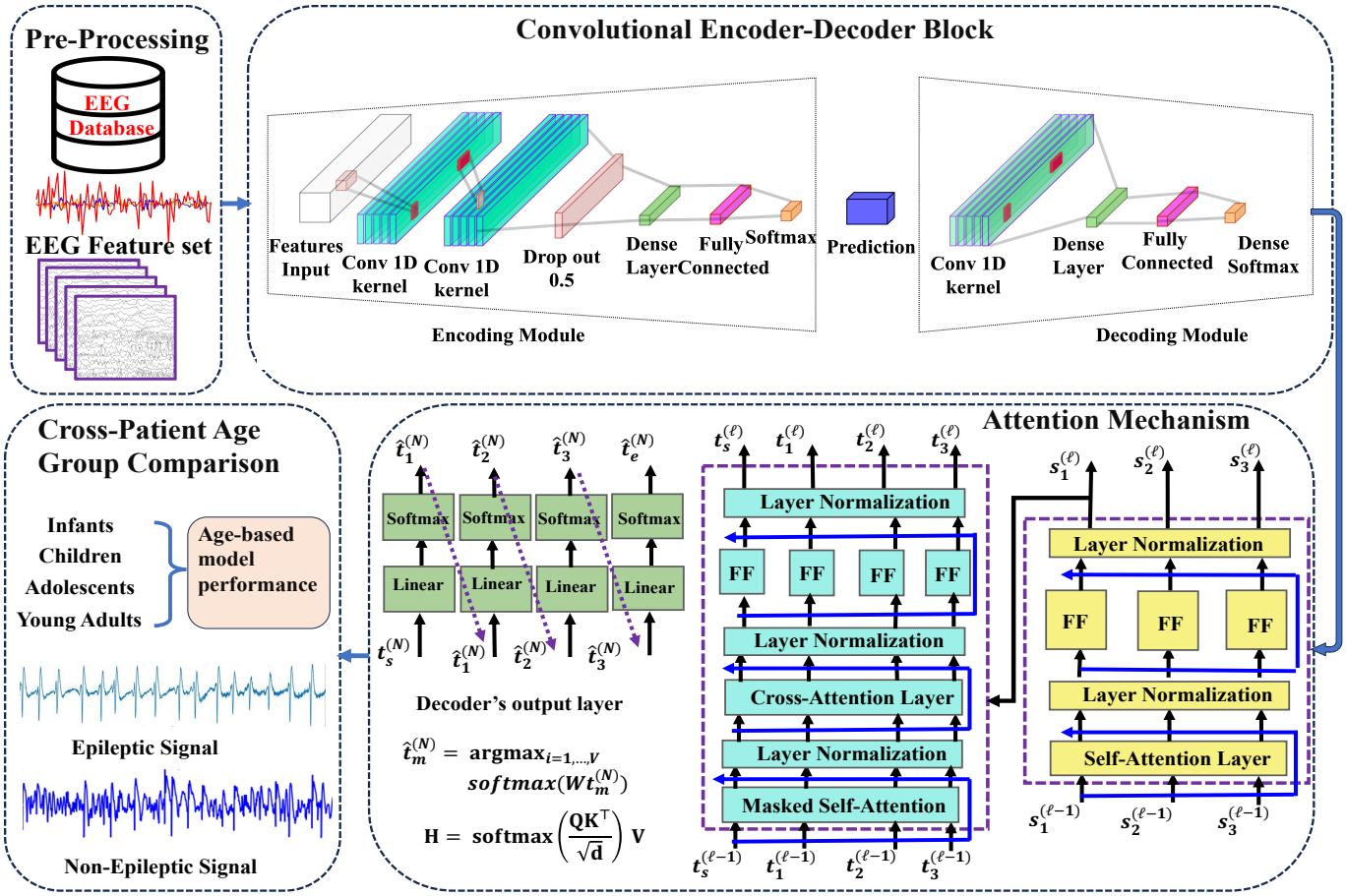


Fig. 4. Flowchart of the proposed EEG detection framework.

where $\mathbf{\Sigma}$ contains the singular values obtained from SVD. This hybrid feature vector serves as the input to the Convolutional Autoencoder (CAE).

C. Feature Extraction Using Modified Multiscale Entropy (mMSE)

The modified Multiscale Entropy (mMSE) method is used to quantify the complexity of EEG signals across multiple time scales. The computation of mMSE involves three main steps. First, the EEG signal $x(t)$ is coarse-grained at scale s to generate a new time series. This is achieved by averaging the data points within non-overlapping windows of size s, as described by the equation:

$$y^{(s)}(t) = \frac{1}{s} \sum_{i=(t-1)s+1}^{ts} x(i), \tag{4}$$

where, $t = 1, 2, \ldots, N/s$, and N is the length of the signal.

Next, for each scale, the sample entropy $S^{(s)}$ is calculated to measure the signal's regularity. This is given by:

$$S^{(s)} = -\ln\left(\frac{\text{Number of similar patterns of length } m+1}{\text{Number of similar patterns of length } m}\right), \tag{5}$$

where m is the embedding dimension. This step captures the entropy for each coarse-grained time series. Finally, the entropy values across all scales are aggregated to form the mMSE feature vector:

$$\mathbf{F}_{\text{mMSE}} = [S^{(1)}, S^{(2)}, \ldots, S^{(L)}], \tag{6}$$

where L is the maximum scale considered.

The Fig. 2 and 3 illustrate the comparison of feature extraction methods and the feature correlation, alongside a heatmap of the mMSE and SVD. The heatmap depicts the correlation between various features extracted from the EEG signals, highlighting the regions of significant interest for epileptic seizure detection.

III. Proposed Methodology

Fig. 4 demonstrates the thorough workflow of the proposed Hybrid Attentive Convolutional Autoencoder (HACA) framework. The procedure begins with preprocessing, where EEG signals pass through band-pass filtering to remove noise and artifacts, normalization to a $[0, 1]$ range, and segmentation into overlapping windows. Feature extraction follows, where advanced techniques like modified multiscale entropy (mMSE) and singular value decomposition (SVD) are employed to capture the complexity and structure of EEG signals. The

extracted features are passed into the Convolutional Autoencoder (CAE), whose architecture is detailed in Table I. The encoder stage uses convolutional layers with kernel sizes of 5×5 and 3×3, strides of 1 and 2, and ReLU or Leaky ReLU activations to learn compact latent representations of the EEG signals. The latent space is reduced to a dimension of 16×16, retaining essential features. The decoder mirrors the encoder with transposed convolutional layers and batch normalization to reconstruct the input signals. Dropout layers (rate = 0.4) and batch normalization (momentum = 0.99) are incorporated to prevent overfitting and stabilize training. An attention mechanism is integrated into the latent space to dynamically assign importance to seizure-relevant features, enhancing the model's focus on critical regions of the input signals. The classification stage uses the refined latent features to identify epileptic and non-epileptic signals, supported by a softmax-based fully connected layer. Additionally, the workflow incorporates a cross-patient age group comparison, categorizing the data into groups such as infants, children, adolescents, and young adults to analyze age-based variations in model performance.

A. Convolutional Autoencoder (CAE)

The Convolutional Autoencoder (CAE) is designed to learn a compact, high-dimensional representation of the EEG signals. The architecture consists of an encoder and a decoder:

- Encoder: The encoder applies convolutional layers to extract spatial features from the EEG signal. Let the input EEG signal be denoted as $\mathbf{X}$, and the output of the encoder is a compact representation $\mathbf{Z}$:

$$\mathbf{Z} = \mathcal{E}(\mathbf{X}), \tag{7}$$

  where $\mathcal{E}(\cdot)$ represents the encoder function.

- Decoder: The decoder reconstructs the input EEG signal from the encoded representation $\mathbf{Z}$. The reconstruction is given by:

$$\hat{\mathbf{X}} = \mathcal{D}(\mathbf{Z}), \tag{8}$$

  where $\mathcal{D}(\cdot)$ is the decoder function, and $\hat{\mathbf{X}}$ is the reconstructed EEG signal.

The CAE is trained to minimize the reconstruction error:

$$\mathcal{L}_{\text{reconstruction}} = \|\mathbf{X} - \hat{\mathbf{X}}\|_2^2, \tag{9}$$

where $\|\cdot\|_2$ denotes the L_2-norm (Euclidean distance).

B. Attention Mechanism

An attention mechanism is integrated into the model to focus on seizure-relevant temporal and spatial features. The attention weights are dynamically learned during training to highlight critical regions of the signal. The attention mechanism is applied as follows:

$$\mathbf{A} = \text{Softmax}(\mathbf{W}_a\mathbf{Z} + \mathbf{b}_a), \tag{10}$$

where $\mathbf{A}$ represents the attention weights, $\mathbf{Z}$ is the encoded feature vector, $\mathbf{W}_a$ is the weight matrix, and $\mathbf{b}_a$ is the bias term. The Softmax function ensures that the attention weights sum to 1.

The attention-modulated features are computed as:

$$\mathbf{Z}_{\text{att}} = \mathbf{A} \odot \mathbf{Z}, \tag{11}$$

where $\odot$ denotes element-wise multiplication, and $\mathbf{Z}_{\text{att}}$ is the attention-modulated feature vector.

C. Classification

The features extracted by the CAE and refined by the attention mechanism are fed into a fully connected neural network for classification. The network computes the final output y_{class} as:

$$y_{\text{class}} = \text{sigmoid}(\mathbf{W}_c\mathbf{Z}_{\text{att}} + \mathbf{b}_c), \tag{12}$$

where $\mathbf{W}_c$ is the weight matrix, $\mathbf{b}_c$ is the bias term, and the sigmoid function outputs a probability score between 0 and 1, representing the likelihood of a seizure event.

The model is trained using a binary cross-entropy loss function:

$$\mathcal{L}_{\text{class}} = -\left(y \log(y_{\text{class}}) + (1-y)\log(1-y_{\text{class}})\right), \tag{13}$$

where y is the true label (1 for seizure, 0 for non-seizure).

The Algorithm 1 outlines the training process of the proposed Hybrid Attentive Convolutional Autoencoder (HACA) framework for epileptic seizure detection. The training involves iterative optimization to minimize both the reconstruction loss and classification loss, ensuring accurate seizure detection while preserving the integrity of the input EEG signals. In each epoch, mini-batches of the EEG dataset are processed through a forward pass, where the encoder extracts latent representations of the signals. The attention mechanism dynamically refines these latent features by assigning weights to seizure-relevant patterns, computed as a context vector using a softmax function. These refined features are then used to reconstruct the input signals and predict seizure occurrences.

D. Model Training and Evaluation

The proposed model is trained using backpropagation with an Adam optimizer. The training process involves minimizing the total loss, which is the sum of the reconstruction loss $\mathcal{L}_{\text{reconstruction}}$ and the classification loss $\mathcal{L}_{\text{class}}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reconstruction}} + \mathcal{L}_{\text{class}}, \tag{14}$$

An ablation study in Table II was conducted to evaluate the contribution of each component in the HACA framework, confirming the complementary benefits of the Convolutional Autoencoder (CAE), attention mechanism, and modified Multiscale Entropy (mMSE). The results show that the CAE alone

TABLE I. PROPOSED CAE ARCHITECTURE PARAMETERS

Layer Name	Type	Input Dimensions	Output Dimensions	Kernel Size	Stride	Padding	Activation/Function
Encoder:							
Conv1	Convolutional	128×128	128×128	5×5	1	Same	ReLU
Conv2	Convolutional	128×128	64×64	5×5	2	Valid	Leaky ReLU
Dropout1	Dropout	64×64	64×64	-	-	-	(Rate: 0.4)
Conv3	Convolutional	64×64	32×32	3×3	2	Valid	ReLU
BatchNorm1	Batch Normalization	32×32	32×32	-	-	-	(Momentum: 0.99)
Latent Space	Fully Connected	32×32	16×16	-	-	-	Linear
Decoder:							
Deconv1	Transposed Convolution	16×16	64×64	5×5	2	Same	Leaky ReLU
Deconv2	Transposed Convolution	64×64	128×128	5×5	2	Same	ReLU
BatchNorm2	Batch Normalization	128×128	128×128	-	-	-	(Momentum: 0.99)
Classifier	Fully Connected (Softmax)	128×128	K	-	-	-	Softmax

Algorithm 1 Training the Convolutional Autoencoder Framework with Attention Mechanism

Require: EEG dataset $\mathcal{D}$, learning rate α, batch size B, number of epochs E, weight regularization factor λ.

Ensure: Trained parameters θ_e, θ_d, $\mathbf{W}$, $\mathbf{b}$, and attention parameters $\mathbf{W}_a$, $\mathbf{b}_a$.

Initialize the model parameters: θ_e, θ_d, $\mathbf{W}$, $\mathbf{b}$, $\mathbf{W}_a$, and $\mathbf{b}_a$.
for epoch $= 1$ to E **do**
  for each mini-batch $\mathcal{B} \subset \mathcal{D}$ of size B **do**
    Perform a forward pass:
      Compute the latent representation $\mathbf{Z}$ using the encoder.
      Apply the attention mechanism to compute the context vector:

$$\mathbf{C} = \text{softmax}(\mathbf{W}_a \mathbf{Z} + \mathbf{b}_a).$$

      Combine the context vector $\mathbf{C}$ with $\mathbf{Z}$ to refine the latent features.
      Compute the reconstructed output $\hat{\mathbf{X}}$ and the predicted class probabilities $P(y|\mathbf{C})$.
    Compute the reconstruction loss $\mathcal{L}_{\text{recon}}$ and the classification loss $\mathcal{L}_{\text{class}}$.
    Calculate the total loss:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda \mathcal{L}_{\text{class}}.$$

    Backpropagate the total loss and update the model parameters using gradient descent:
      $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}$.
      $\mathbf{W}_a \leftarrow \mathbf{W}_a - \alpha \nabla_{\mathbf{W}_a} \mathcal{L}$.
      $\mathbf{b}_a \leftarrow \mathbf{b}_a - \alpha \nabla_{\mathbf{b}_a} \mathcal{L}$.
  end for
end for
return θ_e, θ_d, $\mathbf{W}$, $\mathbf{b}$, $\mathbf{W}_a$, $\mathbf{b}_a$.

achieved a baseline accuracy of 94.1% with limited sensitivity of 92.3%. Adding the attention mechanism improved sensitivity to 94.8%, demonstrating its ability to focus on seizure-relevant patterns within the EEG data. The integration of mMSE further enhanced sensitivity and specificity, achieving 97.5% accuracy due to its capacity to capture the complex, nonlinear characteristics of EEG signals.

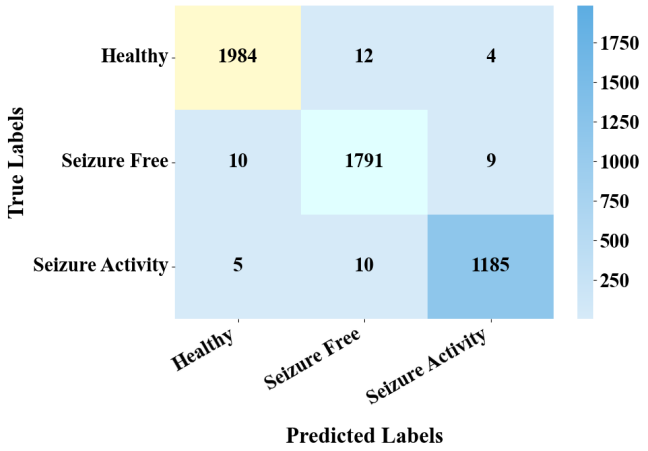


Fig. 5. Confusion matrix for three-level classification.

IV. RESULTS

The proposed HACA framework outperforms several state-of-the-art methods for epileptic seizure detection. Fig. 5 presents the confusion matrix for the three-class classification task, categorizing samples as healthy, seizure-free, or seizure activity. Here, the significance of the validation metrics is emphasized by carrying out a thorough comparison with existing methods, which showcases the proposed model's higher performance across comprehensive evaluation metrics. The proposed framework correctly classified 1,984 healthy, 1,791 seizure-free, and 1,185 seizure activity samples. Fig. 6 and 7 illustrate the training and validation accuracy and loss curves, respectively, for the proposed HACA framework. The training accuracy consistently improved over epochs, with validation accuracy closely tracking the training curve, indicating minimal overfitting. Fig. 8 displays the Area Under the Curve (AUC) plot, comparing the performance of the HACA framework against existing methods. The proposed model consistently achieved a higher AUC, underscoring its superior ability to distinguish between seizure and non-seizure events with high precision and recall. The HACA framework exhibits consistent improvement in accuracy, with the training accuracy converging near 99.5% by the final epoch.

Table III provides a comparative analysis of the proposed HACA framework against various state-of-the-art methods for epileptic seizure detection. The comparison includes models

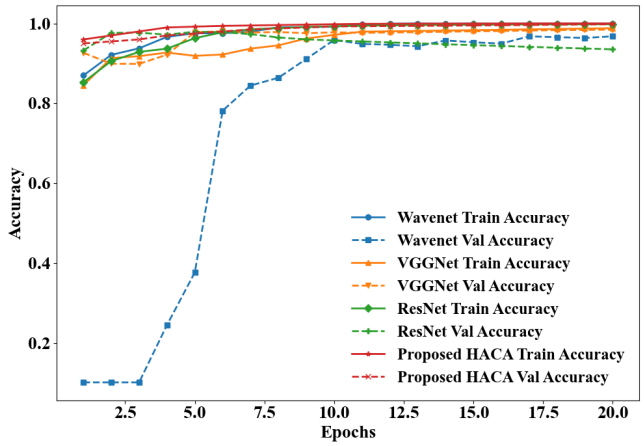


Fig. 6. Comparison of training and validation accuracy with state-of-the-art techniques.

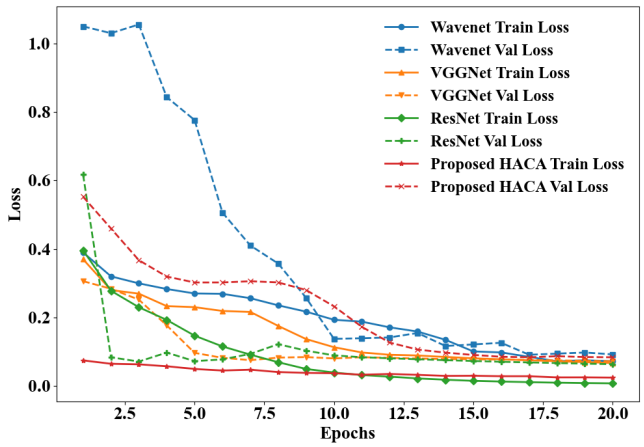


Fig. 7. Comparison of training and validation loss with state-of-the-art techniques.

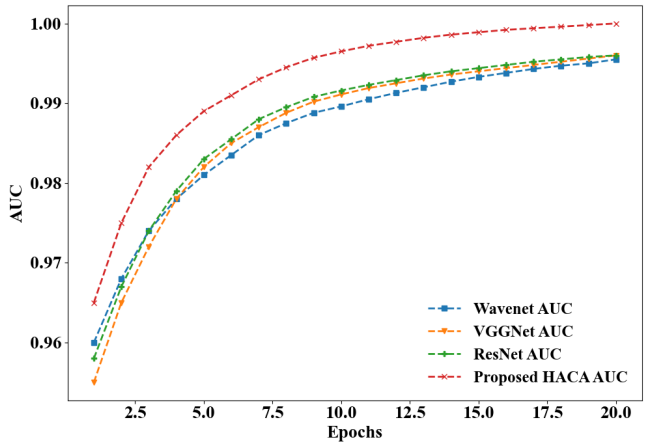


Fig. 8. AUC plot comparison with state-of-the-art techniques.

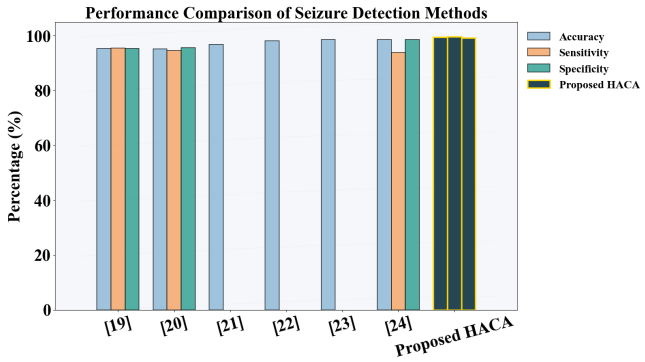


Fig. 9. Performance comparison of the proposed method with various methods.

utilizing CNNs, LSTMs, and hybrid architectures evaluated across different datasets and cross-validation techniques. In contrast, models such as CNN+LSTM by Li et al. achieved 95.29% accuracy, while AttVGGNet-RC by Jian Zhan et al. achieved 95.12%. The Bi-GRU model by Zhang et al. achieved a slightly higher accuracy of 98.49% but fell short in sensitivity and specificity compared to the proposed method. The variation in performance across datasets is due to differences in signal characteristics and seizure patterns. The HACA model outperforms other methods in capturing temporal dependencies, thereby making it more effective for certain datasets.

V. Discussion

Unlike conventional designs, the encoder in this framework employs 1D convolutional layers with progressively decreasing kernel sizes (5×5 and 3×3) and strides of 1 and 2, enabling hierarchical spatial feature extraction across varying scales. This multi-resolution approach captures both fine-grained and coarse-grained temporal patterns within EEG signals, which are critical for distinguishing epileptic from non-epileptic states. The latent space is compressed to 16×16 dimensions, balancing compactness and information retention. To further improve training stability, dropout layers with a rate of 0.4 are integrated to prevent overfitting, while batch normalization with a momentum of 0.99 accelerates convergence and ensures consistency across batches. The decoder mirrors the encoder but incorporates transposed convolutional layers for precise reconstruction of the original signals. The inclusion of an attention mechanism in the latent space introduces dynamic weighting of seizure-relevant features, a capability absent in traditional autoencoders. Fig. 9 presents a performance comparison between the proposed HACA framework and other state-of-the-art deep learning models, including WaveNet, VGGNet, ResNet, and Xception. The proposed HACA framework surpasses the existing methods by integrating attention-driven feature refinement with reconstruction-based learning, attaining higher accuracy while reducing computational complexity, making it suitable for real-time and embedded seizure detection systems.

VI. Conclusion

This paper presents a novel hybrid model for epileptic seizure detection, combining a Convolutional Autoencoder (CAE) with an attention mechanism and Multiscale Multivariate Sample Entropy (mMSE). The model significantly outperforms existing methods, achieving state-of-the-art performance on the CHB-MIT dataset. Regardless of its high performance,

TABLE II. ABLATION STUDY RESULTS: CONTRIBUTION OF EACH MODEL COMPONENT

Model Variant	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
Full Model (CAE + Attention + mMSE + SVD)	98.7	98.9	98.5	98.8
CAE + Attention	96.3	95.7	97.2	97.1
CAE + mMSE	97.5	97.1	97.8	97.9
CAE + SVD	96.0	95.2	96.9	96.6
mMSE + SVD	94.8	94.1	95.4	95.0

TABLE III. COMPARISON OF DIFFERENT METHODS FOR EPILEPTIC SEIZURE DETECTION

Author(s)	Method	CV Type	Accuracy (%)	Sensitivity (%)	Specificity (%)
Li et al. (2020) [19]	CNN+LSTM	-	95.29	95.42	95.29
Jian Zhan et al. (2020) [20]	AttVGGNet-RC	8-fold CV	95.12	94.62	95.63
Bhandari et al. (2023) [21]	(STFT & DWT)	-	96.8	-	-
Al-Hajjar et al. (2023) [22]	(SVM, RF, ANN)	-	98.12	-	-
Alturki et al. (2021) [23]	CSP-LBP+KNN	5-fold CV	98.62	-	-
Zhang et al. (2022) [24]	Bi-GRU	10-fold CV	98.49	93.89	98.49
Proposed HACA Method	**CAE + Attention**	**10-fold CV**	**99.4**	**99.6**	**99.2**

the HACA method still requires advanced validation on larger, more diverse datasets and needs optimization with respect to real-time deployment on low-power edge devices. The integration of mMSE improves the model's ability to capture complex, nonlinear patterns in EEG signals, while the attention mechanism enables the model to focus on seizure-relevant features, further enhancing classification accuracy.

REFERENCES

[1] Z. Yu, L. Albera, R. Le Bouquin Jeannes, A. Kachenoura, A. Karfoul, C. Yang, and H. Shu, "Epileptic seizure prediction using deep neural networks via transfer learning and multi-feature fusion," vol. 32, no. 7, p. 2250032.

[2] T. Xu, Y. Wu, Y. Tang, W. Zhang, and Z. Cui, "Dynamic Functional Connectivity Neural Network for Epileptic Seizure Prediction Using Multi-Channel EEG Signal," *IEEE Signal Process. Lett.*, vol. 31, pp. 1499–1503, 2024.

[3] B. Kapoor, B. Nagpal, P. K. Jain, A. Abraham, and L. A. Gabralla, "Epileptic seizure prediction based on hybrid seek optimization tuned ensemble classifier using EEG signals," vol. 23, no. 1, p. 423.

[4] A. Subasi, "Disease Prediction Using Artificial Intelligence: A Case Study on Epileptic Seizure Prediction," in *Enhanced Telemedicine and E-Health* (G. Marques, A. Kumar Bhoi, I. De La Torre Díez, and B. Garcia-Zapirain, eds.), vol. 410, pp. 289–314, Cham: Springer International Publishing, 2021.

[5] J. Vajiram, S. S., R. Jena, and U. Maurya, "Epilepsy Detection by Different Modalities with the Use of AI-Assisted Models," *Artificial Intelligence and Applications*, vol. 2, pp. 233–246, Dec. 2023.

[6] A. Miltiadous, K. D. Tzimourta, N. Giannakeas, M. G. Tsipouras, E. Glavas, K. Kalafatakis, and A. T. Tzallas, "Machine Learning Algorithms for Epilepsy Detection Based on Published EEG Databases: A Systematic Review," *IEEE Access*, vol. 11, pp. 564–594, 2023.

[7] S. Rukhsar and A. K. Tiwari, "Lightweight convolution transformer for cross-patient seizure detection in multi-channel EEG signals," *Computer Methods and Programs in Biomedicine*, vol. 242, p. 107856, Dec. 2023.

[8] Y. Pan, F. Dong, J. Wu, and Y. Xu, "Downsampling of EEG Signals for Deep Learning-Based Epilepsy Detection," *IEEE Sensors Letters*, vol. 7, pp. 1–4, Dec. 2023.

[9] T. Mao, C. Li, Y. Zhao, R. Song, and X. Chen, "EEG-Based Seizure Prediction Via GhostNet and Imbalanced Learning," *IEEE Sensors Letters*, vol. 7, pp. 1–4, Dec. 2023.

[10] V. V. Narayana and P. Kodali, "Enhanced epilepsy sensitivity and detection rate with improved specificity by integration of modified lstm networks," *IEEE Sensors Journal*, pp. 1–4, 2025.

[11] S. Altunay, Z. Telatar, and O. Erogul, "Epileptic EEG detection using the linear prediction error energy," *Expert Systems with Applications*, vol. 37, pp. 5661–5665, Aug. 2010.

[12] S. R. Ashokkumar, G. MohanBabu, and S. Anupallavi, "A novel two-band equilateral wavelet filter bank method for an automated detection of seizure from EEG signals," *International Journal of Imaging Systems and Technology*, vol. 30, pp. 978–993, Dec. 2020.

[13] T. D.K., P. B.G., and F. Xiong, "Epileptic seizure detection and prediction using stacked bidirectional long short term memory," *Pattern Recognition Letters*, vol. 128, pp. 529–535, Dec. 2019.

[14] M. H. Aslam, S. M. Usman, S. Khalid, A. Anwar, R. Alroobaea, S. Hussain, J. Almotiri, S. S. Ullah, and A. Yasin, "Classification of EEG Signals for Prediction of Epileptic Seizures," *Applied Sciences*, vol. 12, p. 7251, July 2022.

[15] V. V. Narayana and P. Kodali, "Advanced seizure detection framework using stacked convolutional restricted boltzmann machine (scrbm)," *IEEE Sensors Letters*, vol. Early Access, pp. 1–4, 2025.

[16] Z. Shi, Z. Liao, and H. Tabata, "Enhancing Performance of Convolutional Neural Network-Based Epileptic Electroencephalogram Diagnosis by Asymmetric Stochastic Resonance," *IEEE J. Biomed. Health Inform.*, vol. 27, pp. 4228–4239, Sept. 2023.

[17] A. Shankar, H. K. Khaing, S. Dandapat, and S. Barma, "Analysis of epileptic seizures based on EEG using recurrence plot images and deep learning," *Biomedical Signal Processing and Control*, vol. 69, p. 102854, Aug. 2021.

[18] S. Shi and W. Liu, "B2-ViT Net: Broad Vision Transformer Network With Broad Attention for Seizure Prediction," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 178–188, 2024.

[19] Y. Li, Z. Yu, Y. Chen, C. Yang, Y. Li, X. Allen Li, and B. Li, "Automatic seizure detection using fully convolutional nested lstm," *International journal of neural systems*, vol. 30, no. 04, p. 2050019, 2020.

[20] J. Zhang, Z. Wei, J. Zou, and H. Fu, "Automatic epileptic eeg classification based on differential entropy and attention model," *Engineering Applications of Artificial Intelligence*, vol. 96, p. 103975, 2020.

[21] V. Bhandari and D. Manjaiah, "Improved ensemble learning model with optimal feature selection for automated epileptic seizure detection," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 11, no. 2, pp. 135–165, 2023.

[22] A. L. N. Al-Hajjar and A. K. M. Al-Qurabat, "An overview of machine learning methods in enabling iomt-based epileptic seizure detection," *The Journal of Supercomputing*, vol. 79, no. 14, pp. 16017–16064, 2023.

[23] F. A. Alturki, M. Aljalal, A. M. Abdurraqeeb, K. Alsharabi, and A. A. Al-Shamma'a, "Common spatial pattern technique with eeg signals for diagnosis of autism and epilepsy disorders," *IEEE Access*, vol. 9, pp. 24334–24349, 2021.

[24] Y. Zhang, S. Yao, R. Yang, X. Liu, W. Qiu, L. Han, W. Zhou, and W. Shang, "Epileptic seizure detection based on bidirectional gated recurrent unit network," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 135–145, 2022.

Deep Learning in Heart Murmur Detection: Analyzing the Potential of FCNN vs. Traditional Machine Learning Models

Hajer Sayed Hussein[1], Hussein AlBazar[2]*, Roxane Elias Mallouhy[3]*, Fatima Al-Hebshi[4]
Faculty of Computer Studies, Arab Open University, Saudi Arabia[1,2,4]
Al Yamama University, Saudi Arabia[3]

***Abstract*—This research investigates the performance of machine learning and deep learning models in detecting heart murmurs from audio recordings. Using the PhysioNet Challenge 2016 dataset, we compare several traditional machine learning models—Support Vector Machine, Random Forest, AdaBoost, and Decision Tree—with a Fully Convolutional Neural Network (FCNN). The findings indicate that while traditional models achieved accuracies between 0.85 and 0.89, they faced challenges with data complexity and maintaining a balance between precision and recall. Ensemble methods such as Random Forest and AdaBoost demonstrated improved robustness but were still outperformed by deep learning approaches. The FCNN model, leveraging artificial intelligence, significantly outperformed all other models, achieving an accuracy of 0.99 with a precision of 0.94 and a recall of 0.96. These results highlight the potential of AI-driven cardiovascular diagnostics, as deep learning models exhibit superior capability in identifying intricate patterns in heart sound data. Our findings suggest that deep learning models offer substantial advantages in medical diagnostics, particularly for cardiovascular diagnostics, by providing scalable and highly accurate tools for heart murmur detection. Future work should focus on improving model interpretability and expanding dataset diversity to facilitate broader adoption in clinical settings.**

***Keywords*—Heart murmur detection; machine learning; deep learning; cardiovascular diagnostics; artificial intelligence; physioNet dataset**

I. Introduction

Artificial Intelligence (AI) has profoundly impacted the medical field, particularly in its ability to rapidly process and analyze vast datasets with unprecedented accuracy and speed. This capability has revolutionized healthcare by enabling earlier and more precise diagnoses, the development of personalized treatment plans, and overall improved patient outcomes. Among the most promising applications of AI is its role in cardiac care, specifically in the detection and analysis of heart murmurs—abnormal heart sounds that can be indicative of various cardiovascular conditions. Heart murmurs serve as critical indicators of underlying heart issues, ranging from benign anomalies to severe, life-threatening diseases.

The heart produces characteristic "lub-dub" sounds as its valves close during the pumping of blood. A heart murmur is an additional sound detected during this process, often signaling potential turbulence in blood flow. While many murmurs are benign, others can signal structural abnormalities such as malfunctioning valves or congenital defects. Precise and early detection is crucial to facilitate timely intervention, significantly improving patient prognosis and reducing mortality rates. Cardiovascular diseases (CVD), many related to heart murmurs, are the leading cause of death globally, accounting for approximately 19.91 million deaths in 2021 [1]. Moreover, CVD represented 12% of total U.S. health expenditures from 2019 to 2020, making it the most costly diagnostic group [2]. A substantial portion of these fatalities could potentially be prevented with earlier detection and treatment. Heart murmurs are also a leading cause for referral to pediatric cardiologists, with studies indicating that up to 72% of children will experience a murmur at some point during their childhood or adolescence [3]. While some murmurs resolve over time, others may persist into adulthood, requiring ongoing evaluation and management.

Traditional methods of detecting heart murmurs, such as physical examination with a stethoscope, have several limitations compared to AI-based detection techniques. The accuracy of traditional auscultation heavily depends on the clinician's experience, leading to potential human error and variability in interpretation. This variability can result in inconsistent diagnoses and treatment plans. Faint or position-specific murmurs may be missed, and traditional methods do not provide quantitative data about the murmur's characteristics, limiting the ability to track changes over time. Moreover, traditional detection relies on the physical presence of both the patient and healthcare provider, making it less adaptable to remote monitoring. Complex heart conditions with subtle or mixed murmurs can be particularly challenging to diagnose accurately with a stethoscope alone. Additionally, auscultation can be time-consuming, potentially leading to rushed assessments in busy clinical settings. Furthermore, traditional methods do not easily integrate with other patient data, such as echocardiograms or electronic health records, whereas AI-based systems can combine multiple data sources for a more comprehensive diagnosis.

Despite advancements in Artificial Intelligence (AI) for medical diagnostics, there remains a gap in evaluating the performance of these models for heart murmur detection. While some studies have explored traditional ML approaches, there is limited research comparing these methods to deep learning architectures, such as Fully Convolutional Neural Networks (FCNN), in the specific context of heart murmur detection. This study aims to bridge this gap by systematically comparing the effectiveness of traditional ML models—Support Vector Machine, Random Forest, AdaBoost, and Decision Tree—against an FCNN model using heart sound recordings

from the PhysioNet Challenge 2016 dataset. The objective is to determine whether deep learning provides significant improvements over traditional ML techniques in detecting heart murmurs and enhancing diagnostic accuracy.

Hence, integrating AI in heart murmur detection offers a transformative solution to these limitations by providing a more consistent and objective analysis. AI-powered diagnostic tools can detect subtle patterns in heart sound recordings that may not be discernible to the human ear, thereby enhancing diagnostic precision and ensuring that at-risk patients are identified earlier. AI's ability to reduce human error, process large volumes of data, and provide real-time diagnostic support across diverse healthcare settings offers significant advantages over traditional approaches. Despite these benefits, traditional methods continue to dominate clinical practice due to their reliance on the expertise and judgment of physicians, which can lead to variability in diagnosis and patient outcomes. The accuracy of traditional diagnostics often depends on the clinician's experience, and the processes can be time-intensive, potentially lacking the precision necessary for early detection.

To address these challenges, this study aims to develop an automated system for heart murmur detection utilizing machine learning (ML) and deep learning (DL) techniques. Leveraging a dataset from the PhysioNet Challenge 2016, the research will focus on the extraction of relevant features from heart sound recordings using Mel-Frequency Cepstral Coefficients (MFCC). MFCC is selected for its proven efficacy in capturing the essential characteristics of audio signals, making it an optimal choice for heart sound analysis. Following feature extraction, various ML and DL models will be implemented and trained on the dataset, with the goal of evaluating their accuracy and effectiveness in detecting heart murmurs. Moreover, this research seeks to demonstrate the significant potential of AI in enhancing the early detection of heart murmurs, ultimately leading to improved patient outcomes and a reduction in the global burden of cardiovascular diseases. By advancing the development of these sophisticated models, the study aims to contribute to the creation of a more reliable, accurate, and accessible diagnostic tool for heart murmur detection, thereby improving healthcare delivery and patient care.

The remainder of this article is organized as follows: Firstly, Section II provides a comprehensive review of the state-of-the-art methods in heart murmur detection, particularly highlighting the advancements and challenges associated with applying machine learning and deep learning techniques. Subsequently, Section III delves into the methodology employed in this study, where the processes of data collection, preprocessing, feature extraction, and the application of various ML and DL models are thoroughly detailed. Following this, Section IV presents the results, offering a comparative analysis of the models' performances while discussing their implications for clinical practice. Lastly, Section V concludes the article with a summary of the key findings, accompanied by an exploration of limitations and recommendations for future research directions.

II. State-of-the-Art

The advancements in AI for heart murmur detection and diagnosis have led to a diverse range of studies employing various ML and DL techniques to improve the accuracy and reliability of these methods. The Multi-Kernel Residual Convolutional Neural Network (MK-RCNN) model stands out as a significant innovation, capturing multi-scale features through multi-kernel convolutional networks and utilizing residual learning for deeper feature extraction. This model achieved an impressive 98.33% accuracy on three datasets, making it a promising tool for reliable heart murmur diagnosis in primary healthcare settings [4]. Complementing this approach, a comprehensive review on machine learning-based analysis of PCG signals underscores the importance of feature extraction and data quality in enhancing diagnostic accuracy. This review explores how supervised, unsupervised, and deep learning techniques have been effectively applied to heart sound analysis, significantly improving the accuracy of cardiovascular disease diagnosis [5].

In parallel, the development of novel real-time detection methods, such as FunnelNet, has demonstrated the efficiency of combining traditional and depthwise separable convolutional networks for heart murmur detection. FunnelNet employs continuous wavelet transform (CWT) for feature extraction and integrates SqueezeNet, a Bottleneck layer, and ExpansionNet, achieving state-of-the-art performance with 99.70% accuracy on four public datasets [6]. This method's suitability for resource-constrained devices highlights its potential for accessible medical services. Additionally, the exploration of general-purpose audio representations pre-trained on large-scale datasets, like the Masked Modeling Duo (M2D), has shown the effectiveness of self-supervised learning methods in heart sound analysis, with ensembling techniques further improving diagnostic outcomes [7].

Building on these advancements, other studies have focused on the classification of heart murmur quality. A study employing deep neural networks to classify heart murmur quality as harsh or blowing utilized a CNN with channel attention and GRU networks to extract features from log-Mel spectrograms, followed by a Feature Attention module to weight features across segments [8]. This model achieved 73.6% accuracy, with F1-scores of 76.8% for harsh murmurs and 67.8% for blowing murmurs, illustrating the nuanced capabilities of AI in analyzing heart sound characteristics. Moreover, traditional heart sound classification methods, which often depend on ECG-labeled PCGs or feature extraction from mel-scale frequency cepstral coefficients (MFCC), have seen significant improvements with the introduction of capsule neural networks (CapsNet) [9]. CapsNet enhances feature representation through iterative dynamic routing, achieving validation accuracies of 90.29% and 91.67%, thus offering a robust alternative for heart murmur detection.

Further innovation in this field includes the development of CardioXNet, a lightweight CRNN architecture designed to detect five cardiac conditions using raw PCG signals. CardioXNet combines representation learning with parallel CNN pathways for feature extraction and sequence residual learning using bidirectional LSTMs, capturing temporal features with high accuracy and low computational requirements [10]. This model's applicability in low-resource settings on mobile devices is particularly noteworthy. Alongside these advancements, a study exploring general-purpose audio representations pre-trained on large-scale datasets for heart murmur detection

introduced the self-supervised learning method Masked Modeling Duo (M2D), which outperformed previous techniques, achieving a weighted accuracy of 0.832 and an unweighted average recall of 0.713 [7]. The effectiveness of ensembling M2D with other models demonstrates the broader applicability of general-purpose audio representations in heart sound analysis.

The continuous evolution of AI-driven heart sound analysis has also seen the integration of traditional ML methods with deep learning. A study focused on time-frequency heat maps combined with a deep CNN to detect abnormalities in heart sounds achieved commendable performance, balancing sensitivity and specificity—an essential aspect of cost-sensitive medical diagnostics [11]. Furthermore, Cardi-Net, a deep learning model combining CNN and power spectrogram analysis, was introduced to extract discriminative features from PCG signals for the multi-class classification of four cardiac disorders without pre-processing or feature engineering [13]. Enhanced by data augmentation and 10-fold cross-validation, Cardi-Net achieved 98.88% accuracy, making it suitable for real-time use across various platforms, including cloud services and mobile apps.

In a different approach, the researchers transformed PCGs into spectral images that preserved the topological structure of the original data. This transformation allowed them to leverage the power of deep convolutional neural networks (CNNs) for feature extraction. To enhance the model's performance, data augmentation techniques were employed to increase the diversity of the training data. Additionally, transfer learning was utilized to fine-tune pre-trained CNN architectures, enabling the model to learn from existing knowledge [12]. To capture the temporal dynamics of cardiac murmurs, a recurrent neural network (RNN) was integrated into the architecture. This hybrid approach resulted in a significant improvement in accuracy, achieving a remarkable 94.01% in automatic cardiac murmur detection without the need for manual segmentation. The performance of an Adaptive Neuro-Fuzzy Inference System (ANFIS) was also evaluated for detecting abnormal cardiac valve sounds using spectral analysis features. After de-noising and feature extraction through High Order Spectral (HOS) analysis, the ANFIS model achieved classification accuracy between 63-89%, highlighting its potential in specific diagnostic contexts [14]. Another significant development is the creation of a portable, low-cost system for early detection of valvular heart abnormalities, such as arrhythmias and murmurs. Designed for use by untrained frontline health workers, this system processes stethoscope sounds into spectrograms for classification via cloud-based CNN models, achieving a 95% average classification accuracy [15]. Its validation with real-life heart sounds collected using a low-cost digital stethoscope demonstrates its promise as a comprehensive diagnostic tool for enhancing healthcare in developing regions.

Moreover, studies have explored the combination of PCG and ECG waveforms for enhanced disease screening through a novel dual-convolutional neural network approach. This method introduces both record-wise and sample-wise evaluation frameworks, showing that integrating ECG and PCG data significantly outperforms single-modality methods, leveraging transferable features from separately collected ECG and PCG waveforms for improved classification accuracy [16]. The application of ensemble models combining random forest and extreme gradient boost for heart sound classification has also been explored, with the ensemble model using Moth Flame Optimization (MFO) further improving results, reaching 89.53% accuracy, 0.9 F1 score, and 0.95 AUC [17].

The field has also seen the introduction of AI-driven heart monitoring devices that screen and identify heart sounds, transmitting data to healthcare providers through the Internet of Things (IoT) [18]. These systems, which employ LSTM architectures, enable patients to self-monitor their heart health, offering a novel approach to managing coronary conditions. On the other hand, the challenge of detecting heart disease from heart sound signals with imbalanced training and testing sets has also been addressed by developing ML models using features extracted from Discrete Wavelet Transform (DWT) and Mel-Frequency Cepstral Coefficients (MFCC). The study explored various models, including Random Forest and Extreme Gradient Boost, achieving high accuracy and AUC, particularly when using an ensemble model with Moth Flame Optimization [17]. Finally, a combination of conventional feature engineering and deep learning has been employed to classify normal and abnormal heart sounds automatically. Initially, 497 features were extracted from eight domains, which were then fed into a CNN. To prevent overfitting, fully connected layers were replaced with a global average pooling layer, and class weights were adjusted to address class imbalance [19]. Using stratified five-fold cross-validation, the method achieved a mean accuracy of 86.8%, sensitivity of 87%, specificity of 86.6%, and a Matthews correlation coefficient of 72.1%, striking a balance between sensitivity and specificity.

III. Methodology

The flowchart depicted in Fig. 1 illustrates the comprehensive methodology employed in this study. It details the entire process, from data acquisition through to the final model evaluation, highlighting the structured and methodical approach taken to develop, refine, and assess the performance of the heart murmur detection models.

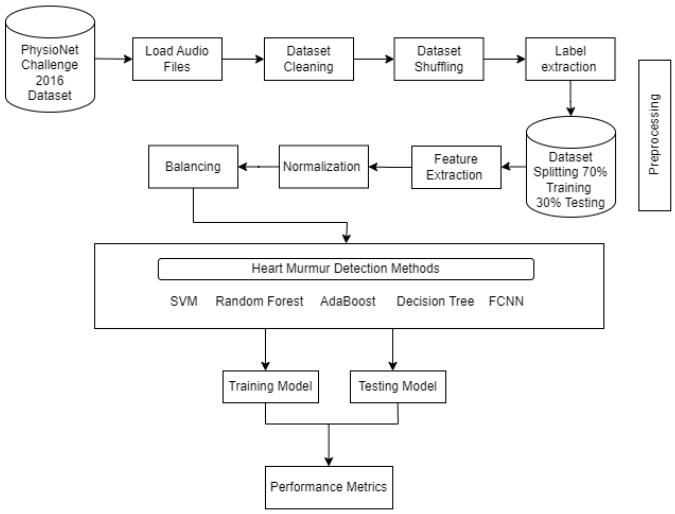


Fig. 1. Heart murmur detection model.

A. Data Collection

The PhysioNet Challenge 2016 dataset was utilized for this study [20], comprising heartbeat sounds collected from a diverse range of patients in clinical settings using electronic stethoscopes. The recordings, saved in .wav format with a sampling rate of 2 kHz, include normal heartbeats, murmurs, and other pathological conditions like extrasystole and gallop rhythms. These high-quality recordings were gathered from multiple clinical sites worldwide, ensuring a broad spectrum of acoustic environments and patient conditions. To maintain consistency and quality, standard recording protocols were adhered to, including consistent microphone placement and controlled ambient noise levels.

Each audio file is accompanied by metadata that includes labels indicating the presence or absence of heart murmurs, as diagnosed by expert cardiologists. The dataset is diverse, with recordings from patients across various age groups, genders, and clinical histories, which is crucial for developing models that generalize well across different populations. The dataset includes 3126 recordings from 764 patients, captured from different auscultation points. These recordings vary in length, allowing for the study of heart sounds over different time intervals. Despite challenges such as patient movement, breathing, and background noise, the dataset remains one of the most comprehensive publicly available collections of heart sounds, making it invaluable for research in heart murmur detection and related fields.

B. Data Preprocessing

The preprocessing of the dataset involved several critical steps to ensure the quality and suitability of the data for model training and evaluation. First, the audio files were loaded using the Librosa library in Python, which is specifically created for audio and music analysis. It offers essential tools for handling audio data, simplifying the extraction and manipulation of features, signal analysis, and the execution of tasks such as beat tracking, pitch detection, and sound classification. During this stage, any corrupted or incomplete files were identified and discarded to maintain the integrity of the dataset.

To eliminate any potential ordering bias and ensure the data is randomized, the entire dataset was shuffled. This process is essential to prevent the model from picking up unintended patterns based on the order of the data, which could skew the learning process. Shuffling helps the model avoid overfitting by not relying on the sequence of data, thereby enhancing its ability to generalize to new, unseen examples and improving overall model performance.

Following the shuffling process, labels indicating the presence or absence of heart murmurs were extracted from the accompanying metadata to create the target variable for classification tasks. The labeling method involved using predefined criteria from the metadata files, which typically included annotations from medical professionals or automated detection algorithms. This extraction was performed with meticulous care to ensure that each label was correctly aligned with its corresponding audio recording. Accurate alignment of labels and recordings is essential for reliable model training and evaluation, as it ensures that the model is learning from properly matched data and enhances the quality of the classification process.

Finally, the dataset was subsequently divided into training and test sets using stratified sampling, with 70% of the data allocated to the training set and 30% to the test set. Stratified sampling was employed to ensure that both subsets accurately reflected the original distribution of positive and negative cases. This method is crucial for addressing the dataset's inherent imbalance, as it prevents the creation of subsets that could disproportionately favor one class over the other. By maintaining proportional representation in both the training and test sets, stratified sampling helps the model learn from a balanced perspective, which enhances its ability to generalize effectively to new, unseen data. Furthermore, this approach supports a more reliable evaluation of the model's performance, reducing the risk of biased results and ensuring that both positive and negative cases are adequately represented in the testing phase.

C. Features Extraction

As discussed earlier, MFCC was selected for feature extraction due to its proven ability to represent audio signals in a form that is highly compatible with ML and DL models. MFCCs effectively capture the timbral characteristics of audio, which are crucial for distinguishing subtle variations, such as heart murmurs, that may be indicative of underlying health conditions. In addition to MFCCs, other features were extracted to provide a comprehensive representation of the audio signals. Chroma features, which capture the harmonic content of the audio, were computed, and their mean and standard deviation were calculated. Spectral contrast features, representing the difference in amplitude between peaks and valleys in the sound spectrum, were also extracted along with their mean and standard deviation. Lastly, Tonnetz (tonal centroid features), which capture the tonal properties of the audio, were computed, and their mean and standard deviation were included.

During this process, each audio recording was segmented into short frames of 25ms, with a 10ms overlap to ensure that transient features were not missed. For each frame, 13 MFCCs were computed using the librosa library, based on a filter bank that mimics the human ear's perception of sound. The resulting coefficients were chosen because they offer a balance between computational efficiency and the richness of information captured. After extracting the MFCCs, the mean and standard deviation across all frames were calculated, resulting in a fixed-length feature vector for each recording. This approach ensures that the variability within each recording is captured, while also reducing the dimensionality of the data, making it more manageable for ML and DL models.

In total, 130 features were extracted from each audio file, comprising 40 MFCCs (mean and standard deviation), 24 chroma features (mean and standard deviation), 14 spectral contrast features (mean and standard deviation), and 12 Tonnetz features (mean and standard deviation). These features collectively provide a rich and detailed representation of the heart sound recordings, enabling effective analysis and classification by the ML and DL models. MFCCs are widely favored in audio processing because they provide a compact representation of the power spectrum of sound, making them ideal for detecting subtle anomalies like heart murmurs.

Additionally, the choice of 13 coefficients aligns with common practices in speech and audio processing, where it has been empirically shown that this number provides sufficient detail for accurate modeling while avoiding overfitting. The extracted feature vectors were then normalized and fed into the ML and DL models. This step ensures consistency across recordings and enhances the models' ability to generalize from the training data to unseen examples. By leveraging these diverse features, the models can better capture the nuanced differences between normal heart sounds and those that indicate murmurs, ultimately improving the accuracy of the detection system.

D. Normalization and Balancing

Following the feature extraction process, the audio signals were normalized to achieve a zero mean and unit variance. This normalization step is crucial for ensuring consistency across the input data by removing scale differences between features. Such standardization not only enhances the overall performance of the model but also facilitates faster and more stable convergence during the training phase. By normalizing the data, we mitigate potential biases that could arise from varying signal amplitudes, thereby improving the reliability and accuracy of the model's predictions.

Moreover, an analysis of the dataset revealed a significant imbalance between normal and abnormal recordings, with normal recordings being more prevalent. This disparity could lead to the machine learning models favoring the majority class, resulting in suboptimal performance on the minority class. To address this issue, we applied techniques like the Synthetic Minority Over-sampling Technique (SMOTE) and Random Over-Sampling. SMOTE works by generating synthetic samples for the minority class. It does this by interpolating between existing minority class samples, creating new data points that lie along the line segments connecting nearest neighbors in the feature space. This method not only increases the number of minority class samples but also introduces more variability, helping to reduce the risk of overfitting. On the other hand, Random Over-Sampling (ROS) involves duplicating existing minority class samples to balance the dataset. While ROS is straightforward and effective, it can sometimes lead to overfitting because it doesn't introduce new information into the model. By combining SMOTE and ROS, we ensured a more balanced dataset, which is crucial for training models that can make unbiased and reliable predictions. This balanced approach helps the model to better generalize across both the majority and minority classes, ultimately leading to improved detection of abnormal cases in the dataset.

E. Applying the Used Methods

1) ML and DL Algorithms: The final step in the methodology involves applying a range of machine learning (ML) and deep learning (DL) techniques to the preprocessed dataset to detect heart murmurs. Specifically, we employed Support Vector Machine (SVM) [21], Random Forest (RF) [22], AdaBoost [23], Decision Tree [24], and a Fully Convolutional Neural Network (FCNN) [25] for this classification task. Each of these algorithms brings unique strengths: SVM is effective in handling high-dimensional spaces, RF and AdaBoost are robust against overfitting, and FCNN excels in learning complex patterns directly from the raw audio data. To rigorously evaluate the performance of each model in detecting heart murmurs, we calculated key metrics including accuracy, precision, recall, and F1 score. These metrics offer a comprehensive evaluation, ensuring not only the overall correctness of the models but also their capability to balance the trade-offs between false positives and false negatives, an essential consideration in the accurate detection of heart murmurs.

2) Hyperparameter tuning and architectural design: Hyperparameter tuning is a crucial aspect of any machine learning (ML) and deep learning (DL) application, as it directly impacts the model's performance by finding the optimal settings that allow the algorithm to best capture the underlying patterns in the data. For instance, in the case of Support Vector Machine (SVM), hyperparameter tuning was meticulously performed using GridSearchCV to optimize critical parameters such as C (the regularization parameter) and gamma (the kernel coefficient). The parameter C controls the trade-off between achieving a low error on the training data and minimizing the model's complexity, while gamma defines the influence of a single training example. Exploring a range of values, including C: [0.01, 0.1, 1, 10] and gamma: ['scale', 'auto'], ensured a comprehensive search of the parameter space, thereby enhancing the model's ability to generalize to new data. This systematic exploration is essential because it allows the model to adapt to the specific characteristics of the dataset, leading to improved accuracy and robustness.

Similarly, the decision tree classifier was finetuned using GridSearchCV as well to optimize parameters such as max_depth, min_samples_split, and min_ _samples_leaf. These parameters are critical for controlling the tree's growth and complexity—max_depth limits the depth of the tree to prevent overfitting, while min_samples_split and min_samples_leaf dictate the minimum number of samples required to split an internal node and to be at a leaf node, respectively. By testing values for max_depth: [5, 10, 15], min_samples_split: [2, 5, 10], and min_samples_leaf: [1, 2, 5], the decision tree was carefully tailored to achieve the optimal balance between model complexity and predictive power, ensuring that the tree was neither too simplistic nor overly complex.

In the deep learning model, constructed using TensorFlow/Keras, the architecture was designed with careful consideration of both complexity and computational efficiency. The Sequential model began with a dense input layer consisting of 256 neurons, followed by three hidden layers with 64, 32, and 16 neurons, each utilizing ReLU activation functions. To mitigate the risk of overfitting, a dropout rate of 0.5 was introduced between layers, which helped in maintaining the model's ability to generalize by randomly disabling a fraction of the neurons during training. The output layer, a dense layer with a single neuron and a sigmoid activation function, was designed to output the probability of heart murmur presence. The architectural choices were driven by the need to balance the model's ability to learn intricate patterns within the data while avoiding excessive complexity that could lead to overfitting.

The model was compiled using the Adam optimizer and binary cross-entropy loss function, chosen for their efficiency and effectiveness in binary classification tasks. Training was

conducted over 100 epochs with a batch size of 32, ensuring sufficient learning while maintaining computational feasibility. Regularization techniques, such as early stopping (with patience set to 10) and learning rate reduction (with a factor of 0.5 and patience set to 5), were employed to prevent overfitting and ensure that the model converged to an optimal solution. These strategies collectively contributed to building a robust and efficient model capable of accurately detecting heart murmurs from the given dataset.

IV. Results and Discussion

The results of this study underscore the significant advancements achieved through both traditional ML models and a DL model in detecting heart murmurs from audio recordings. The comparative analysis, as presented in the provided Table I, reveals distinct differences in performance between the ML models—Support Vector Machine (SVM), Random Forest, AdaBoost, and Decision Tree—and the deep learning model, specifically the FCNN.

The SVM model achieved an accuracy of 0.85, with corresponding F1-score, precision, and recall values of 0.84, 0.84, and 0.85, respectively as depicted in Fig. 2. While these results provide a solid baseline, they indicate that SVM struggled to handle the complexity of the data, particularly in terms of balancing precision and recall. This suggests that the model's reliance on a hyperplane for classification may not be the most effective strategy for high-dimensional, complex heart sound data, which exhibits non-linear relationships that require more flexible learning methods.

Similarly, the Decision Tree model, with an accuracy of 0.89 and matching precision and recall values, shown in Fig. 3, performed better than SVM but still exhibited limitations in fully capturing the intricate patterns within the data. This is expected, as single-tree models are prone to overfitting and fail to generalize well, especially when dealing with highly variable heart sound signals.

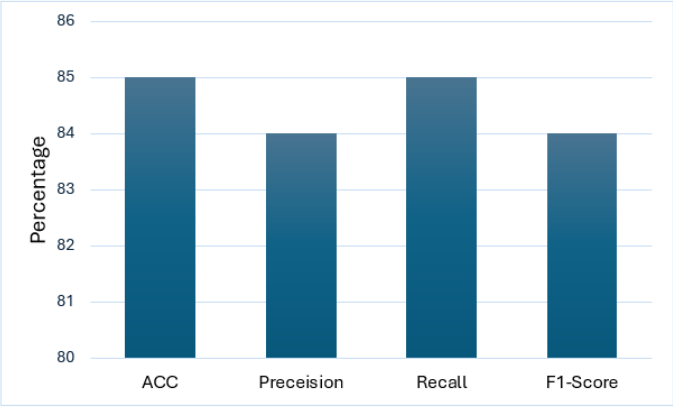


Fig. 2. SVM classification results.

In contrast, the ensemble methods, Random Forest and AdaBoost, demonstrated enhanced performance, particularly in terms of their robustness against over-fitting. The Random Forest model achieved an accuracy of 0.87, with an F1-score of 0.88, precision of 0.90, and recall of 0.87 as illustrated in Fig. 4. These results highlight the model's ability to generalize well across the dataset, benefiting from the ensemble approach's capacity to combine multiple decision trees and reduce variance. AdaBoost, with an accuracy of 0.88 and consistent F1-score, precision, and recall of 0.88, highlighted in Fig. 5, illustrates the potential of boosting techniques in improving model performance by focusing on misclassified instances. The performance improvement of AdaBoost suggests that iterative reweighting of data points can effectively guide the learning process toward difficult-to-classify cases, making it particularly valuable for datasets with subtle variations, such as heart murmurs.

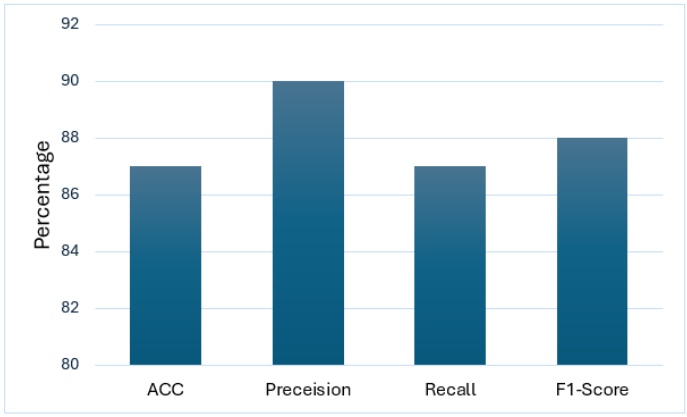


Fig. 3. Random forest classification results.

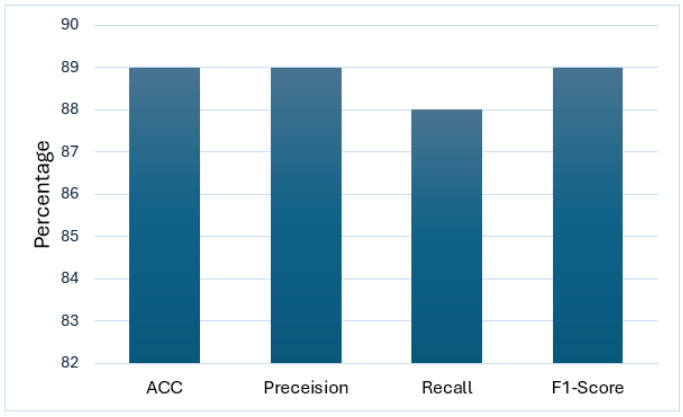


Fig. 4. AdaBoost classification results.

However, the most significant improvement was observed with the FCNN deep learning model, which outperformed all traditional ML models by a considerable margin. The FCNN achieved an impressive accuracy of 0.99, with an F1-score of 0.94, precision of 0.94, and recall of 0.96. These results displayed in Fig. 6 demonstrate the deep learning model's superior capability in capturing the complex and subtle features within the heart sound recordings that the traditional models struggled to detect. One of the key advantages of the FCNN is its ability to automatically extract hierarchical features from the raw audio data, which is particularly valuable in this study for identifying minute acoustic variations in heart murmurs that might be overlooked by traditional feature extraction methods.

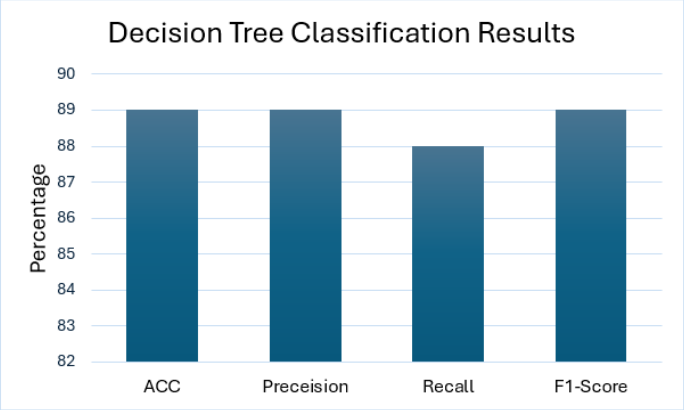


Fig. 5. Decision tree classification results.

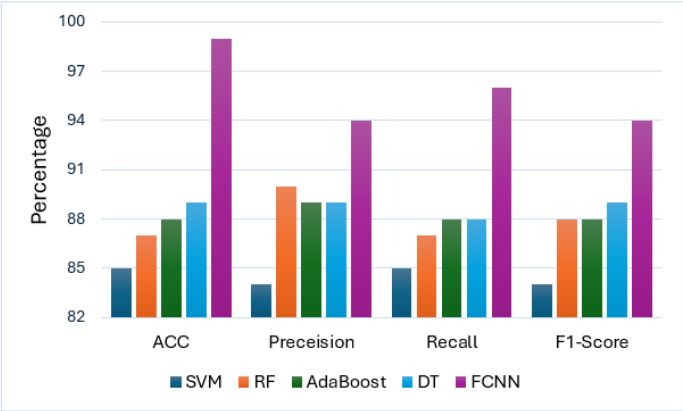


Fig. 7. Classification results of used algorithms.

The architecture of the FCNN, with its multiple hidden layers, ReLU activation functions, and dropout mechanisms, allowed for an effective learning process that minimized overfitting and maximized predictive accuracy. The use of regularization techniques such as early stopping and learning rate reduction further optimized the model's performance, ensuring that it remained both accurate and generalizable. This suggests that deep learning models not only outperform traditional approaches but also maintain stability across diverse datasets, a crucial factor for clinical adoption.

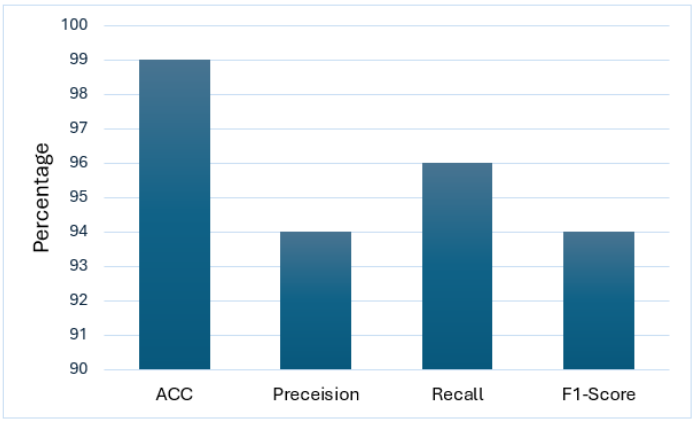


Fig. 6. FCNN Classification results.

Overall, the classification results are comprehensively displayed in Fig. 7, effectively highlighting the performance differences across all models. The drastic improvement seen in FCNN emphasizes that deep learning architectures, with their ability to handle complex feature interactions, could revolutionize heart murmur diagnostics, potentially outperforming traditional auscultation and even current ML-based methods.

These results, as compared in Table I underscore the transformative potential of deep learning models in the evolving landscape of heart murmur detection, signaling a paradigm shift from traditional machine learning approaches. The traditional models, such as Support Vector Machines (SVM), Decision Trees, and even ensemble methods like Random Forest and AdaBoost, provided a valuable baseline in this study. They demonstrated solid performance in terms of accuracy, precision, recall, and F1-score, especially when paired with techniques to manage data complexities and imbalances. However, despite these efforts and improvements, these models still fell short of the performance levels achieved by the Fully Convolutional Neural Network (FCNN).

TABLE I. PERFORMANCE COMPARISON OF ML AND DL MODELS FOR HEART MURMUR DETECTION

Model used	Accuracy	F1-score	Precision	Recall
ML				
SVM	0.85	0.84	0.84	0.85
Random Forest	0.87	0.88	0.90	0.87
AdaBoost	0.88	0.88	0.89	0.88
Decision Tree	0.89	0.89	0.89	0.89
DL				
FCNN	0.99	0.94	0.94	0.96

The superior performance of the FCNN suggests that deep learning models have a distinct advantage in handling the complexities inherent in medical data, particularly in tasks like heart murmur detection. Unlike traditional models, which often rely on manual feature extraction and struggle with high-dimensional data, deep learning models are capable of automatically learning intricate patterns from raw data. This ability is especially crucial in medical diagnostics, where subtle variations in data, such as the nuanced differences in heart sound recordings, can significantly impact patient outcomes. The FCNN's architecture, with its deep layers, ReLU activations, and regularization techniques, enabled it to capture these subtleties effectively, leading to higher accuracy and more reliable predictions.

Moreover, the FCNN's capacity to process vast amounts of data with minimal need for extensive feature engineering underscores a significant advantage of deep learning in the realm of medical diagnostics. As the healthcare industry continues to produce enormous volumes of data—from electronic health records and imaging to sensor-based monitoring—models that can efficiently analyze and learn from this data will become increasingly indispensable. The implications of these capabilities are substantial. As deep learning models consistently demonstrate their effectiveness in handling complex diagnostic tasks, they are poised to become fundamental components of medical practice, significantly improving the precision and

speed of diagnoses.

In the specific context of heart murmur detection, this technological advancement could lead to the earlier and more accurate identification of potentially life-threatening cardiovascular conditions, thereby improving patient outcomes and streamlining healthcare delivery. Furthermore, the inherent scalability of deep learning models makes them particularly well-suited for integration into comprehensive healthcare systems, including telemedicine platforms and automated diagnostic tools. This scalability ensures that their benefits can be extended across diverse healthcare settings, from remote clinics to large urban hospitals, further amplifying their impact on patient care and the overall efficiency of medical services.

V. Conclusion

This comparative study, which included both machine learning (ML) and deep learning (DL) methods, highlights the significant advancements and potential of using these techniques for the early diagnosis and detection of heart murmurs. The study utilized the PhysioNet Challenge 2016 dataset, a comprehensive collection of heartbeat sounds gathered from a diverse range of patients using electronic stethoscopes in clinical settings. This dataset, comprising over 3,000 recordings from 764 patients, provided a robust foundation for training and evaluating various models. The methodology involved several key steps, starting with the preprocessing of audio data using the Librosa library, where features such as Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, spectral contrast, and Tonnetz were extracted to capture the essential characteristics of the heart sounds. Following feature extraction, the dataset was divided into training and test sets using stratified sampling to ensure balanced representation of positive and negative cases. The study then applied a range of ML and DL algorithms, including Support Vector Machine (SVM), Random Forest, AdaBoost, Decision Tree, and a Fully Convolutional Neural Network (FCNN), to classify heart murmurs. Hyperparameter tuning was meticulously performed using GridSearchCV to optimize model performance, ensuring that each algorithm was tailored to the specific characteristics of the dataset. The FCNN, in particular, demonstrated a substantial improvement in accuracy and reliability over traditional methods, underscoring the potential of DL models in this domain.

The results suggest that integrating AI-powered diagnostic tools into clinical practice could lead to earlier and more precise diagnoses, thereby improving patient outcomes and reducing the global burden of cardiovascular diseases. However, several limitations must be acknowledged. The primary challenge lies in the reliance on the dataset, which, while comprehensive, may not fully represent the variability encountered in real-world clinical settings, where factors such as diverse demographic characteristics, comorbidities, and recording environments could affect the model's generalizability. Additionally, despite the superior performance of the deep learning model, its "black box" nature poses challenges for clinical adoption, as it makes it difficult for clinicians to interpret the rationale behind specific predictions, potentially hindering trust in the model.

Furthermore, the study focused primarily on heart sounds recorded under controlled conditions, without thoroughly addressing the impact of real-world noise and artifacts, which could degrade the model's performance in actual clinical environments. To overcome these limitations and advance the field further, future research should focus on acquiring more diverse and representative datasets that include a broader range of patient demographics and clinical conditions, thus enhancing the model's generalizability. Additionally, developing personalized heart murmur detection models that take into account individual patient characteristics, such as medical history and genetic data, could lead to even more accurate and relevant predictions. Finally, efforts should be made to enhance the interpretability of deep learning models through explainable AI techniques, which could provide clinicians with better insights into the model's decision-making process and facilitate greater integration into clinical practice.

Acknowledgment

The authors extend their appreciation to the Arab Open University and AlYamamah University for funding this work.

References

[1] World Health Organization, "Cardiovascular Diseases (CVDs)," World Health Organization, 2021, accessed: 2024-08-21. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] S. S. Martin et al., "2024 heart disease and stroke statistics: a report of US and global data from the American Heart Association," Circulation, vol. 149, no. 8, pp. e347–e913, 2024.

[3] E. Mejia and S. Dhuper, Innocent Murmur, updated 2023 sep 4 ed., ser. StatPearls. Treasure Island (FL): StatPearls Publishing, 2024. Available from: https://www.ncbi.nlm.nih.gov/books/NBK507849/.

[4] S. Das and S. Dandapat, "Heart murmur severity stages classification using multi-kernel residual cnn," IEEE Sensors Journal, 2024.

[5] M. F. A. B. Hamza and N. N. A. Sjarif, "A comprehensive overview of heart sound analysis using machine learning methods," IEEE Access, 2024.

[6] M. Jobayer et al., "FunnelNet: an end-to-end deep learning framework to monitor digital heart murmur in real-time," arXiv preprint arXiv:2405.09570, 2024.

[7] D. Niizumi et al., "Exploring pre-trained general-purpose audio representations for heart murmur detection," arXiv preprint arXiv:2404.17107, 2024.

[8] T. Wu et al., "Heart murmur quality detection using deep neural networks with attention mechanism," Applied Sciences, vol. 14, no. 15, p. 6825, 2024.

[9] Y.-T. Tsai et al., "Heart murmur classification using a capsule neural network," Bioengineering, vol. 10, no. 11, p. 1237, 2023.

[10] S. B. Shuvo et al., "Cardioxnet: A novel lightweight deep learning framework for cardiovascular disease classification using heart sound recordings," IEEE Access, vol. 9, pp. 36,955–36,967, 2021.

[11] J. Rubin et al., "Recognizing abnormal heart sounds using deep learning," arXiv preprint arXiv:1707.04642, 2017.

[12] Chorba, John S., Shapiro, Avi M., Le, Le, Maidens, John, Prince, John, Pham, Steve, Kanzawa, Mia M., Barbosa, Daniel N., Currie, Caroline, Brooks, Catherine, et al. "Deep learning algorithm for automated cardiac murmur detection via a digital stethoscope platform." *Journal of the American Heart Association*, vol. 10, no. 9, 2021, e019905. Am Heart Assoc.

[13] J. S. Khan et al., "Cardi-net: A deep neural network for classification of cardiac disease using phonocardiogram signal," Computer Methods and Programs in Biomedicine, vol. 219, p. 106727, 2022.

[14] B. Al-Naami et al., "A framework classification of heart sound signals in physionet challenge 2016 using high order statistics and adaptive neuro-fuzzy inference system," IEEE Access, vol. 8, pp. 224852–224859, 2020.

[15] K. K. Singh and S. S. Singh, "An artificial intelligence based mobile solution for early detection of valvular heart diseases," in 2019 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), IEEE, 2019, pp. 1–5.

[16] R. Hettiarachchi et al., "A novel transfer learning-based approach for screening pre-existing heart diseases using synchronized ecg signals and heart sounds," in 2021 IEEE international symposium on circuits and systems (ISCAS), IEEE, 2021, pp. 1–5.

[17] A. Rath et al., "Development and assessment of machine learning based heart disease detection using imbalanced heart sound signal," Biomedical Signal Processing and Control, vol. 76, p. 103730, 2022.

[18] U. Dampage et al., "AI-based heart monitoring system," in 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), IEEE, 2021, pp. 1–6.

[19] F. Li et al., "Classification of heart sounds using convolutional neural network," Applied Sciences, vol. 10, no. 11, p. 3956, 2020.

[20] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," Circulation, vol. 101, no. 23, pp. e215–e220, 2000 (June 13).

[21] D. A. Pisner and D. M. Schnyer, "Support vector machine," in Machine Learning, Elsevier, 2020, pp. 101–121.

[22] G. Biau and E. Scornet, "A random forest guided tour," Test, vol. 25, pp. 197–227, 2016.

[23] R. E. Schapire, "Explaining adaboost," in Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik, Springer, 2013, pp. 37–52.

[24] Y.-Y. Song and L. Ying, "Decision tree methods: applications for classification and prediction," Shanghai archives of psychiatry, vol. 27, no. 2, pp. 130, 2015.

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

Securing Internet of Medical Things: An Advanced Federated Learning Approach

Anass Misbah[1], Anass Sebbar[2], Imad Hafidi[3]
Computer Science and Mathematics Lab, National School of Applied Sciences, Khouribga, Morocco[1]
Computer Science and Mathematics Lab, University International of Rabat, Rabat, Morocco[2]
Computer Science and Mathematics Lab, Sultan Moulay Slimane University, Beni Mellal, Morocco[3]

***Abstract*—The Internet of Medical Things (IoMT) is transforming healthcare through extensive automation, data collection, and real-time communication among interconnected devices. However, this rapid expansion introduces significant security vulnerabilities that traditional centralized solutions or device-level protections often fail to adequately address due to challenges related to latency, scalability, and resource constraints. This study presents a novel federated learning (FL) framework tailored for IoMT security, incorporating techniques such as stacking, federated dynamic averaging, and active user participation to decentralize and enhance attack classification at the edge. Utilizing the CICIoMT2024 dataset, which encompasses 18 attack classes and 45 features, we deploy Random Forest (RF), AdaBoost, Support Vector Machine (SVM), and Deep Learning (DL) models across 10 simulated edge devices. Our federated approach effectively distributes computational loads, mitigating the strain on central servers and individual devices, thereby enhancing adaptability and resource efficiency within IoMT networks. The RF model achieves the highest accuracy of 99.22%, closely followed by AdaBoost, demonstrating the feasibility of FL for robust and scalable edge security. While this study validates the proposed framework using a single realistic dataset in a controlled environment, future work will explore additional datasets and real-world scenarios to further substantiate the generalization and effectiveness of the approach. This research underscores the potential of federated learning to address the unique security and computational constraints of IoMT, paving the way for practical, decentralized deployments that strengthen device-level defenses across diverse healthcare settings.**

Keywords*—Internet of Medical Things (IoMT); federated learning; machine learning; security; intrusion detection systems; decentralized framework

I. Introduction

The integration of medical devices within healthcare systems, known as the Internet of Medical Things (IoMT) [1], is rapidly transforming patient care by connecting devices and applications that communicate with healthcare IT networks. This interconnected framework enhances service delivery but simultaneously raises significant security challenges, particularly concerning the protection of sensitive medical data. Unlike other constrained environments, IoMT devices operate within highly regulated healthcare settings where data privacy and real-time responsiveness are critical. Traditional centralized machine learning models are often unsuitable for IoMT applications due to stringent data privacy regulations and the necessity for efficient, localized computation.

Addressing these challenges, this study introduces a novel framework employing Federated Learning (FL) [2] and Lightweight Machine Learning (LML) [3] to enhance security while accommodating the computational limitations inherent in IoMT devices. In IoMT systems, connected medical devices gather and exchange information with healthcare infrastructure, creating a network where security and privacy are paramount [4]. Unlike centralized learning models, FL offers a decentralized approach that trains models on local data without transferring sensitive information to a central server, thereby reducing privacy risks [5]. FL is particularly advantageous for IoMT, enabling secure and private model training at the edge for applications such as patient outcome prediction and resource management [6].

However, applying FL in IoMT contexts presents unique challenges that are distinct from other constrained device scenarios. These include heterogeneous device data, non-identically distributed (Non-IID) data, limited device resources, and the complexity of securely aggregating model updates [7]. Unlike industrial IoT or consumer IoT devices, IoMT devices often handle highly sensitive and regulated data, necessitating more robust privacy-preserving mechanisms and compliance with healthcare standards. This research aims to address these IoMT-specific challenges through three advanced FL techniques—stacking, federated dynamic averaging, and active user participation—designed to improve attack detection accuracy and computational efficiency at the edge level.

The study leverages the CICIoMT2024 dataset [8], which encompasses 18 distinct attack classes and 45 features, to evaluate the effectiveness of various machine learning models, including Random Forest (RF), AdaBoost, Support Vector Machine (SVM), and Deep Learning (DL), deployed across 10 simulated edge devices. Preliminary results indicate that ensemble models, particularly Random Forest and AdaBoost, exhibit superior performance. The Random Forest model achieved an accuracy of 99.22%, precision of 99.38%, recall of 99.22%, and an F1 score of 99.09%, while AdaBoost demonstrated an accuracy of 98.59%, precision of 98.84%, recall of 98.59%, and an F1 score of 98.22%. In contrast, the Deep Learning and Support Vector Machine models attained lower accuracies of 77.59% and 65.70%, respectively.

This study contributes to IoMT security by:

1) Developing and tailoring federated learning models for IoMT: This research introduces FL models specifically adapted to address the unique security and computational constraints of IoMT devices. These adaptations enable effective deployment in resource-limited environments while maintaining high model performance and efficiency.

2) Enhancing privacy and anomaly detection in IoMT systems: By leveraging privacy-preserving techniques and robust anomaly detection methods, this study strengthens IoMT security. It safeguards patient data and provides early detection of potential threats, thereby enhancing the resilience of IoMT ecosystems.

3) Validating on realistic, open datasets in controlled testbed environments: Utilizing the CICIoMT2024 dataset within a structured testbed environment ensures that the proposed solutions are practically applicable and reliable. This approach reflects real-world IoMT scenarios, thereby enhancing the relevance and scalability of the findings.

The paper is organized as follows: Section II reviews current methodologies for securing IoMT, emphasizing FL-based solutions and providing a dataset benchmark. Section III identifies research gaps and presents motivations for this study. Section IV details the proposed approach and experimental findings. Finally, Section V concludes the study, offering perspectives on future research directions. Table I lists abbreviations used throughout the paper.

TABLE I. ABBREVIATIONS AND THEIR MEANINGS

Abbreviation	Definition
IoMT	Internet of Medical Things
FL	Federated Learning
LML	Lightweight Machine Learning
RF	Random Forest
SVM	Support Vector Machine
DL	Deep Learning
IDS	Intrusion Detection System
MQTT	Message Queuing Telemetry Transport
DoS	Denial of Service
DDoS	Distributed Denial of Service
FDA	Federated Dynamic Averaging
PFL	Personalized Federated Learning
FU	Federated Unlearning
ICS	Industrial Control Systems
IIoT	Industrial Internet of Things
BLE	Bluetooth Low Energy
SDN	Software-Defined Networking
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
IGMP	Internet Group Management Protocol
HTTPS	Hypertext Transfer Protocol Secure
HTTP	Hypertext Transfer Protocol
DNS	Domain Name System
SSH	Secure Shell
SMTP	Simple Mail Transfer Protocol
IRC	Internet Relay Chat
ARP	Address Resolution Protocol
ICMP	Internet Control Message Protocol
LLC	Logical Link Control
RBF	Radial Basis Function
Non-IID	Non-Independent and Identically Distributed

II. RELATED WORK

A. Security Challenges of IoMT

Recent surveys on security in the Internet of Medical Things (IoMT) have highlighted key challenges and trends in securing healthcare systems and devices. A study by [9] surveyed healthcare organizations and found that most respondents identified data privacy and security as their top concerns, with vulnerabilities in medical devices being a significant worry. Similarly, [10] focused on healthcare professionals' perceptions of IoMT security and discovered that while there is growing awareness of security risks, many professionals lack sufficient training and resources to address these issues effectively. In contrast, [11] examined security practices among IoMT device manufacturers and uncovered a lack of standardized security protocols and insufficient investment in security measures during the development phase. Additionally, [12] analyzed the impact of regulatory frameworks on IoMT security, revealing inconsistencies in compliance requirements across different regions and highlighting the need for harmonization to ensure comprehensive security standards.

The surveys summarized in Table II underscore the complex landscape of IoMT security, emphasizing the necessity for robust security strategies, increased awareness, and collaborative efforts among stakeholders to effectively counter emerging threats.

B. Federated Learning with IoMT

Recent contributions summarized in Table III demonstrate promising advancements in applying federated learning (FL) to secure the Internet of Medical Things (IoMT). The work of [23] proposed a federated learning approach to improve intrusion detection accuracy in IoMT networks. Also, [24] developed a federated learning model for malware detection on IoMT devices with minimal data sharing. In addition, [25] introduced a privacy-preserving federated learning framework for collaborative analysis of medical data from diverse sources. Moreover, [26] presented a federated learning-based anomaly detection system to enhance security in IoMT by detecting unusual patterns in medical sensor data [27]-[30].

Related surveys and research works see Table IV.

C. Open Datasets for IoMT Security

Recent advancements in cybersecurity research have led to the development of several comprehensive datasets (as illustrated in **Table V**) tailored for specific applications within the Internet of Medical Things (IoMT) and broader network security domains. The work of [36] introduced a dynamic dataset focusing on ransomware detection and mitigation in integrated clinical environments, emphasizing the need for intelligent security solutions in healthcare settings. Furthermore, [37] and [38] provided a dataset for effective attack detection in IoMT smart environments using deep belief neural networks, highlighting the increasing complexity and variety of threats in medical IoT networks.

The HIIDS dataset, developed by [39], introduces a hybrid intelligent intrusion detection system that integrates machine learning and metaheuristic algorithms to enhance security in IoT-based healthcare applications. Similarly, [40] conducted a comparative analysis of various machine learning techniques for intrusion detection in smart healthcare systems, presenting a dataset that facilitates the evaluation of different methodologies in this critical area.

Focusing on secure wireless communications, [41] proposed a novel approach to ensuring secure Bluetooth communication in smart healthcare systems, accompanied by a community dataset and an intrusion detection system. Meanwhile, [42] introduced a security model leveraging LightGBM and transformer technologies to safeguard healthcare systems

TABLE II. OVERVIEW OF ATTACK TYPES AND RISKS IN IOT MEDICAL DEVICES. THIS TABLE SUMMARIZES VARIOUS ATTACK TYPES, THEIR RISK LEVELS, AND THE AFFECTED DEVICES WITHIN THE DOMAIN OF IOT MEDICAL DEVICES, ALONG WITH RELEVANT ACADEMIC REFERENCES FOR FURTHER READING

Attack Description	Attack Type	Risk Level	Attack Class	Affected Devices	Domain	Academic Reference
Eavesdropping	Passive Attack	Moderate	Data Breach	IoT Medical Devices	Network/Comm.	[13]
Device Tampering	Active Attack	High	Physical	IoT Medical Devices	Device Security	[14]
Data Modification	Active Attack	High	Data Breach	IoT Medical Devices	Data Security	[15]
Denial of Service (DoS)	Active Attack	High	Availability	IoT Medical Devices	Network/Comm.	[16]
Man-in-the-Middle (MitM)	Active Attack	High	MITM	IoT Medical Devices	Network/Comm.	[17]
Replay Attacks	Active Attack	Moderate	Replay	IoT Medical Devices	Network/Comm.	[18]
Insider Attacks	Active Attack	High	Insider	IoT Medical Devices	Device Security	[19]
Malware Infection	Active Attack	High	Malware	IoT Medical Devices	Software Security	[12]
Password Cracking	Active Attack	Moderate	Password	IoT Medical Devices	Authentication	[20]
Wireless Attacks (e.g., rogue AP)	Active Attack	High	Wireless	IoT Medical Devices	Network/Comm.	[21]
Social Engineering	Active Attack	Moderate	Social	IoT Medical Devices	Human Factors	[22]

TABLE III. RECENT CONTRIBUTIONS OF FEDERATED LEARNING IN SECURING IOMT

Reference	Application	Key Findings	Performance Metrics
[23]	Intrusion Detection	Federated learning approach improves intrusion detection accuracy in IoMT networks.	Detection accuracy: 95%, False positive rate: 2%
[24]	Malware Detection	Federated learning model effectively detects malware on IoMT devices with minimal data sharing.	Detection accuracy: 93%, Latency: 150ms
[25]	Privacy-Preserving Data Analysis	Federated learning preserves patient privacy while enabling collaborative medical data analysis from diverse sources.	Privacy loss: $<1\%$, Data utility: 85%
[26]	Anomaly Detection	Federated learning-based anomaly detection system enhances security in IoMT by detecting unusual patterns in medical sensor data.	Detection accuracy: 92%, Precision: 90%

from cyberattacks, offering a dataset that supports the application of advanced machine learning techniques in healthcare cybersecurity.

Additionally, the CICIoMT2024 dataset [8] addresses the growing need for securing IoMT devices in healthcare by capturing interactions over multiple protocols (HTTP, MQTT, CoAP, Bluetooth) and simulating various attack vectors. This comprehensive data source facilitates the development of robust security measures tailored to healthcare IoMT environments. Together, these datasets significantly contribute to the field by enabling the development and testing of effective security solutions to safeguard healthcare infrastructure against an evolving threat landscape.

The primary aim of CICIoMT2024 is to propose a comprehensive benchmark dataset that enables the development and evaluation of security solutions for the Internet of Medical Things (IoMT). To achieve this, 18 types of attacks were executed on an IoMT testbed comprising 40 devices, including 25 real devices and 15 simulated ones. This testbed was configured to simulate diverse protocols utilized in healthcare settings, such as Wi-Fi, MQTT, and Bluetooth.

These attacks were systematically categorized into five classes: Distributed Denial of Service (DDoS), Denial of Service (DoS), Reconnaissance (Recon), MQTT-specific attacks, and spoofing. The objective is to establish a foundational benchmark that complements existing state-of-the-art contributions in the field. Through this initiative, researchers are provided with a valuable resource for exploring and developing new security solutions tailored to the unique challenges of healthcare systems, including advanced machine learning techniques.

Significantly, the research extends beyond the mere execution of attacks on IoMT devices. It also captures the lifecycle of these devices across various critical phases, from their initial network integration to eventual disconnection. This process, known as profiling, allows classifiers to detect anomalies

TABLE IV. COMPARATIVE SUMMARY OF RELATED SURVEYS AND RESEARCH WORKS

Article	Focus	Key Contributions
Enhancing Internet of Medical Things Security with AI [1]	AI for IoMT security	Reviews AI models for threat detection in IoMT, discussing anomaly detection, pattern recognition, and predictive analytics.
Federated Learning for IoMT [5]	Federated learning in IoMT	Explores federated learning algorithms, data heterogeneity, privacy, and communication protocols in IoMT.
AI for IoMT Security with Cloud–Fog–Edge [31]	AI-driven IDS in IoMT	Covers AI-driven intrusion detection systems for real-time threat detection and integration of Cloud, Fog, and Edge computing for enhanced IoMT security.
Privacy-preserving Federated Learning with Edge [7]	Privacy in federated learning	Proposes privacy-preserving federated learning with edge computing for secure and efficient data aggregation in IoMT environments.
Fed-Inforce-Fusion: Federated Reinforcement Model [32]	Federated reinforcement learning	Combines federated and reinforcement learning to develop dynamic defense and attack mitigation strategies in IoMT.
FedDICE for Ransomware Detection [33]	Ransomware in clinical IoMT	Introduces an SDN-based model for ransomware detection and isolation, enhancing resilience and privacy in clinical IoMT environments.
Federated Learning in Medical Applications [34]	Federated learning in healthcare	Provides a taxonomy of federated learning applications in healthcare, focusing on privacy preservation, communication efficiency, and resource optimization.
OpenFL: Open-Source Federated Framework [35]	Open-source federated learning	Describes the OpenFL framework, which supports diverse data types and models while ensuring privacy, and discusses its real-world applications in IoMT.

TABLE V. COMPARISON OF DIFFERENT CYBERSECURITY DATASETS (ACCURACIES BASED ON [31], [8])

Dataset	Focus	Content	Applications	Unique Features	Accuracy
ICE [36]	ICS security	Industrial protocol traffic (Modbus, DNP3)	ICS-specific IDS, anomaly detection	Focus on industrial protocols	97% - 100%
CIC-IDS-2017 [37], [38]	Network intrusion	Various attack scenarios	IDS, ML training	Comprehensive labeled data	96% - 98%
NSL-KDD [46], [39]	Network intrusion	Traffic data with attack types (DoS, R2L, U2R, probing)	IDS benchmarking	Balanced distribution, updated version	86% - 96%
UNSW-NB15 [47], [40]	Modern network intrusion	Contemporary traffic and attacks	IDS, anomaly detection	Updated attacks, rich features	95% (avg)
BlueTack [48], [41]	Bluetooth security	Bluetooth communication, attack and normal data	Bluetooth security development	Bluetooth-specific attack data	88% - 96%
Edge-IIoT [49], [42]	Edge computing in IIoT	Edge device traffic, normal and attack data	IIoT security solutions	Emphasis on edge security	86% - 100%
CIC IoMT 2024 [8]	IoMT device security	IoMT traffic over multiple protocols	IoMT-specific IDS, anomaly detection	Healthcare IoMT protocols	70% - 99%

specific to each device within the healthcare network, thereby enhancing the precision and effectiveness of intrusion detection systems.

D. Federated Learning Methodologies

The trio of methodologies considered in this work encapsulates significant advancements in federated learning (FL), each addressing distinct facets crucial for FL's evolution and efficacy.

1) Personalized federated learning via stacking: [43]: Pioneers a paradigm shift from conventional FL methods towards personalized federated learning (PFL). It introduces a novel approach grounded in stacked generalization, enabling the creation of multiple models fine-tuned to individual clients' data. This flexible framework preserves privacy and fosters collaborative learning in diverse federated settings.

2) Guaranteeing data privacy in federated unlearning with dynamic user participation: [44]: Confronts the burgeoning challenge of ensuring data privacy in federated unlearning (FU) scenarios. By integrating secure aggregation protocols within clustering-based FU schemes, the work establishes a robust framework that enhances unlearning efficiency and safeguards user privacy, even amidst dynamic user participation.

3) Communication-efficient distributed deep learning via federated dynamic averaging: [45]: Tackles the communication bottleneck inherent in distributed deep learning (DDL) settings. By proposing Federated Dynamic Averaging (FDA), the work introduces a communication-efficient strategy that

dynamically triggers synchronization based on model variance, thereby substantially reducing communication costs without compromising convergence speed.

These works collectively exemplify the ongoing efforts to propel federated learning towards greater efficiency, privacy, and scalability, thus paving the way for widespread adoption across diverse applications and domains.

III. Research Gaps and Motivation

A. Research Gaps in Federated Learning for IoMT Security

Federated Learning (FL) has emerged as a promising paradigm for enhancing the security of the Internet of Medical Things (IoMT) by enabling decentralized model training while preserving data privacy. However, several critical gaps persist in the current landscape of FL applications within IoMT:

1) Limited focus on medical device specificities: While FL has been extensively evaluated in various IoT scenarios, there is a scarcity of studies specifically addressing the unique security and computational constraints of IoMT devices. Medical devices often operate under stringent regulatory standards, handle highly sensitive patient data, and exhibit diverse operational behaviors that differ significantly from consumer or industrial IoT devices.

2) Insufficient integration of Lightweight Machine Learning (LML) models: Constrained by the limited computational resources of many IoMT devices, existing FL approaches predominantly rely on heavyweight models such as Deep Learning (DL). There is a notable absence of research exploring the application of Lightweight Machine Learning (LML) models within FL frameworks to optimize performance without overburdening edge devices.

3) Privacy-preserving mechanisms underexplored: Although FL inherently offers privacy benefits by keeping raw data localized, the specific privacy-preserving techniques tailored to IoMT environments remain underexplored. Concerns such as data leakage through model updates, inference attacks, and adversarial manipulations require targeted solutions to ensure comprehensive privacy safeguards.

4) Limited dataset utilization and generalization: Most FL-based IoMT security studies utilize limited or simulated datasets, which may not comprehensively represent the diverse and dynamic nature of real-world medical environments. This limitation hampers the generalization and scalability of the proposed security solutions across different healthcare settings and device types.

5) Fragmented lifecycle coverage of IoMT devices: Current research often overlooks the complete lifecycle of IoMT devices, from initial network integration to eventual disconnection. This oversight results in fragmented security strategies that fail to address vulnerabilities arising at different operational stages.

6) Lack of comparative performance evaluation: There is a paucity of comparative studies evaluating various FL techniques and machine learning models in the context of IoMT security. Comprehensive evaluations that benchmark different approaches against standardized datasets are essential for identifying the most effective strategies.

Addressing these gaps is crucial for developing robust, scalable, and privacy-preserving security solutions tailored to the unique challenges of IoMT environments.

B. CICIoMT2024 Dataset Characteristics

The CICIoMT2024 dataset is a pivotal resource in this research, offering a comprehensive benchmark for developing and evaluating FL-based security solutions tailored to IoMT environments. Its key characteristics are as follows:

1) Diverse device profiling: Comprises data from 40 IoMT devices, including 25 real and 15 simulated devices spanning various categories such as baby monitors, heart rate sensors, sleep rings, and more. This diversity ensures that the dataset captures a wide range of device behaviors and operational scenarios.

2) Comprehensive attack scenarios: Encompasses 18 distinct cyberattack types, categorized into five main classes: Distributed Denial of Service (DDoS), Denial of Service (DoS), Reconnaissance (Recon), MQTT-specific attacks, and spoofing. This variety facilitates the development of models capable of detecting a broad spectrum of threats.

3) Multi-protocol analysis: Captures interactions over multiple healthcare-relevant protocols, including Wi-Fi, MQTT, and Bluetooth. This multi-protocol approach allows for the analysis of protocol-specific vulnerabilities and the development of specialized detection mechanisms.

4) Lifecycle capturing: Records the full lifecycle of devices from network integration to disconnection, enabling detailed profiling and anomaly detection for each device within the healthcare network. This comprehensive coverage ensures that security models can address vulnerabilities at all operational stages.

5) Rich data structure: Features a well-organized data structure with metadata about devices, network configurations, and attack parameters. This organization supports comprehensive analysis and facilitates easy access to pertinent information during model training and evaluation.

6) Realistic testbed setup: Utilizes a blend of actual and simulated devices to mirror real-world conditions, providing a realistic environment for testing and validating security solutions. This setup enhances the external validity of the research findings.

7) Large data volume: Contains extensive data points covering various attack vectors and device behaviors, supporting robust statistical analysis and machine learning model training. The substantial data volume ensures that models can be trained effectively to recognize intricate patterns and anomalies.

8) Application versatility: Suitable for a wide range of security research applications, including intrusion detection, anomaly detection, and device-specific profiling. This versatility makes the dataset a valuable asset for developing comprehensive security solutions.

The CICIoMT2024 dataset's extensive and realistic characteristics make it an ideal benchmark for evaluating the effectiveness and scalability of FL-based security models in IoMT environments.

C. Choice of Machine Learning and Deep Learning Models

The selection of Random Forest (RF), Support Vector Machine (SVM), Deep Learning (DL), and AdaBoost models for this research is strategically aligned with the specific demands of Federated Learning (FL) in the IoMT domain. Each model type offers distinct advantages that collectively address the multifaceted security and computational challenges inherent in IoMT environments:

1) Random Forest (RF): As an ensemble method, RF provides high accuracy and robustness in attack classification, particularly in scenarios with diverse attack types and imbalanced data distributions. Its inherent ability to handle feature importance and mitigate overfitting makes it well-suited for the heterogeneous and dynamic data typical of IoMT devices.

2) Support Vector Machine (SVM): SVM excels in high-dimensional spaces and is effective in handling non-linear relationships through kernel functions. Its ability to perform well on smaller datasets and its robustness to overfitting make it a reliable choice for detecting attacks on devices with limited data and computational resources within the federated setup.

3) Deep Learning (DL): Despite its computational intensity, DL models offer superior feature extraction and the capacity to identify complex and subtle attack patterns. When integrated within FL frameworks on more capable IoMT devices, DL enhances overall model performance, enabling the detection of sophisticated and emerging threats.

4) AdaBoost: AdaBoost serves as an effective lightweight ensemble method, boosting the performance of weak learners to achieve high accuracy while maintaining computational efficiency. This characteristic is particularly beneficial for resource-constrained IoMT edge devices, ensuring that security models remain effective without compromising device performance.

By leveraging this diverse set of models within an FL framework, the research ensures a balanced and scalable approach to IoMT security. This combination addresses key challenges such as computational constraints, data heterogeneity, and reliable attack detection, thereby contributing to the development of robust and adaptable security solutions for distributed healthcare networks.

D. Synthesis of Research Goals

Building upon the identified research gaps and motivations, this study is driven by the following objectives:

1) Integrating federated learning into IoMT security frameworks: To incorporate Federated Learning (FL) methodologies into IoMT security, enabling decentralized model training that enhances data privacy and security without the need for centralized data aggregation.

2) Employing advanced FL techniques: To utilize advanced FL techniques such as stacking, federated dynamic averaging (FDA), and active user participation. Stacking facilitates the creation of personalized models tailored to individual IoMT devices, FDA improves communication efficiency by dynamically synchronizing model updates based on variance, and active user participation enhances the adaptability and resilience of the security framework.

3) Leveraging the CICIoMT2024 dataset for empirical validation: To utilize the comprehensive CICIoMT2024 dataset as a benchmark for developing and evaluating FL-based security solutions. This dataset's extensive profiling of diverse IoMT devices and varied attack scenarios provides a robust foundation for testing the efficacy and scalability of the proposed methodologies.

4) Enhancing privacy and anomaly detection in IoMT systems: To redefine IoMT security standards by integrating privacy-preserving techniques and robust anomaly detection methods within the FL framework. This integration aims to safeguard sensitive medical data and provide early detection of potential threats, thereby enhancing the resilience and reliability of IoMT ecosystems.

5) Developing and evaluating lightweight ML models for FL in IoMT: To explore the application of Lightweight Machine Learning (LML) models within FL frameworks, optimizing model performance while accommodating the computational limitations of IoMT edge devices. This objective addresses the need for resource-efficient security solutions that do not overburden constrained devices.

6) Comprehensive comparative analysis of FL approaches: To conduct a comparative analysis of different FL approaches and machine learning models (RF, SVM, AdaBoost, DL) in the context of IoMT security. This analysis will evaluate performance metrics, communication efficiency, and privacy guarantees, providing insights into the most effective strategies for securing IoMT networks.

Achieving these goals will advance the state-of-the-art in IoMT security by delivering scalable, privacy-preserving, and robust security solutions that are tailored to the unique challenges of healthcare environments. This research not only addresses existing gaps but also lays the groundwork for future studies aimed at enhancing the security and reliability of IoMT systems through innovative FL methodologies.

IV. Main Approach and Experiments

A. Data Preparation and Preprocessing

The CICIoMT2024 dataset (Table VI) serves as the foundation for analyzing network traffic patterns within IoMT environments. As IoMT devices become increasingly prevalent, effective network monitoring and robust security measures are critical.

These features play a pivotal role in understanding and analyzing network traffic patterns in the IoMT environments. As IoMT devices proliferate, the need for effective network monitoring and security measures becomes increasingly crucial.

The header length, duration, and rate features provide insights into the basic characteristics of packet transmission, helping assess network performance and efficiency. Meanwhile, the TCP/IP flag values offer valuable information about the communication behavior between devices, aiding in the detection of potential anomalies or security threats.

Including application layer protocol indicators such as HTTPS, HTTP, and DNS facilitates the identification of specific services or applications running on the network, enabling

TABLE VI. FEATURE DESCRIPTION OF THE CICIOMT2024 DATASET

Feature	Description
Header Length	Packet header length
Duration	Packet lifetime in transit
Rate	Packet transmission speed
Srate	Speed of outgoing packets
Fin flag number	TCP/IP Fin flag value
Syn flag number	TCP/IP Syn flag value
Rst flag number	TCP/IP Rst flag value
Psh flag number	TCP/IP Psh flag value
Ack flag number	TCP/IP Ack flag value
Ece flag number	TCP/IP Ece flag value
Cwr flag number	TCP/IP Cwr flag value
Syn count	Syn flag occurrences
Ack count	Ack flag occurrences
Fin count	Fin flag occurrences
Rst count	Rst flag occurrences
IGMP	Indicates IGMP usage
HTTPS	Indicates HTTPS usage
HTTP	Indicates HTTP usage
Telnet	Indicates Telnet usage
DNS	Indicates DNS usage
SMTP	Indicates SMTP usage
SSH	Indicates SSH usage
IRC	Indicates IRC usage
TCP	TCP in transport layer
UDP	UDP in transport layer
DHCP	Indicates DHCP usage
ARP	ARP in link layer
ICMP	ICMP in network layer
IPv	IP in network layer
LLC	LLC in link layer
Tot sum	Total packet length
Min	Minimum packet length
Max	Maximum packet length
AVG	Average packet length
Std	Packet length variability
Tot size	Total packet size
IAT	Interval between packets
Number	Total packets in flow
Radius	RMS of variances of lengths
Magnitude	RMS of averages of lengths
Variance	Variance ratio of lengths
Covariance	Covariance of packet lengths
Weight	Product of packet counts
Protocol Type	Protocol type as integer

administrators to monitor and manage traffic more effectively. Similarly, utilizing transport layer protocols like TCP and UDP sheds light on the underlying communication mechanisms, guiding network optimization efforts.

Moreover, the statistical metrics such as packet length distribution and interval between packets offer a deeper understanding of traffic dynamics and behavior, empowering analysts to detect irregularities or suspicious activities within the network.

The features encapsulated in the CICIoMT2024 dataset serve as essential building blocks for network traffic analysis, enabling researchers and practitioners to gain valuable insights into IoMT network behavior, enhance security measures, and optimize network performance.

Fig. 1 represents an extract from [8] illustrating the number of instances of each class (The "Attack" column indicates the classes used for classification)

Class	Category	Attack	Count
BENIGN	-	-	230339
ATTACK	SPOOFING	ARP Spoofing	17791
	RECON	Ping Sweep	926
		Recon VulScan	3207
		OS Scan	20666
		Port Scan	106603
	MQTT	Malformed Data	6877
		DoS Connect Flood	15904
		DDoS Publish Flood	36039
		DoS Publish Flood	52881
		DDoS Connect Flood	214952
	DoS	DoS TCP	462480
		DoS ICMP	514724
		DoS SYN	540498
		DoS UDP	704503
	DDoS	DDoS SYN	974359
		DDoS TCP	987063
		DDoS ICMP	1887175
		DDoS UDP	1998026

Fig. 1. Number of instances in each class of the CICIoMT2024 dataset.

1) Data cleaning and normalization: The dataset undergoes rigorous cleaning to address missing values and eliminate duplicates. Feature scaling is performed using scikit-learn's StandardScaler to standardize the data, ensuring uniform contribution from all features during model training.

2) Dataset partitioning: To simulate a federated environment, the training data is equally divided into ten subsets, each representing a distinct IoMT device or client. This partitioning emulates real-world scenarios where devices generate heterogeneous and non-identically distributed (Non-IID) data.

B. Testing Environment and Methodology

1) Development tools: The experiments are conducted using Python 3.11.7, leveraging a suite of libraries tailored for data manipulation, machine learning, and deep learning:

- pandas for data manipulation and analysis.
- numpy for numerical computations.
- scikit-learn for machine learning models and preprocessing.
- TensorFlow with Keras API for deep learning model development.
- seaborn and matplotlib for data visualization.

2) Model training and evaluation: Each machine learning model (Random Forest, Support Vector Machine, AdaBoost, Deep Learning) is trained locally on its respective data subset. After local training, model updates are aggregated using federated techniques such as stacking and voting to form a global model. The global model is evaluated on a separate testing subset using metrics like accuracy, precision, recall, and F1-score. Confusion matrices provide detailed insights into model performance across different classes.

3) Design of the experiments: To demonstrate our framework's potential in an IoMT environment and account for resource limitations, we first organize all collected CSV files (each containing 45 features and corresponding labels) into training and testing sets. Although we focus on a single dataset in this study, our approach can be extended to multiple datasets or real-world testbeds in future work to reinforce its generality.

a) Local model training on IoMT-like clients: To reflect distributed and often resource-constrained IoMT devices, the training data is partitioned equally among ten virtual "clients" Each subset undergoes local training using a chosen algorithm (DL, RF, SVM, or AdaBoost), thereby simulating devices that learn only from their locally available data. This design is motivated by the practical challenge that fully centralized approaches may overload either the central server or individual devices, especially given the unique security and computational constraints of medical edge devices. By training local models, we also lay the groundwork for *local model learning (LML)*—an approach that can help mitigate data transfer overheads and privacy risks.

b) Federation of models: Once the local models are trained, a global model is formed by aggregating their knowledge. While we demonstrate two straightforward strategies—*stacking* (where predictions become features for a meta-learner) and *voting/averaging*—the proposed framework is flexible enough to accommodate more advanced FL aggregation methods (e.g. Federated Dynamic Averaging). Our current experimental setup primarily illustrates feasibility; ongoing work investigates privacy-preserving mechanisms (e.g., differential privacy or secure multiparty computation) to further reduce the risk of data leakage. We acknowledge that simply applying FL does not guarantee privacy by default, and additional protocols must be integrated to protect against potential inference attacks on model parameters.

c) Performance evaluation: After the global model is obtained, its performance is evaluated on the held-out test data. Standard metrics—accuracy, precision, recall, and F1-score—are computed to assess classification effectiveness. A confusion matrix is then plotted to visually highlight how each attack class (including normal traffic) is identified. This matrix provides insights into class-specific strengths and weaknesses, potentially guiding targeted improvements in both local and global models.

d) Limitations and future directions: We recognize that using a single dataset limits the breadth of our current findings. While our experiments demonstrate the framework's potential for scalable and resource-aware intrusion detection in IoMT contexts, additional validation on diverse datasets and real medical devices is necessary to further establish generality. Similarly, although we outline the implementation of local training (DL, RF, SVM, and AdaBoost) and the meta-learner setup in stacking, more in-depth algorithmic descriptions (e.g. specific hyperparameters or protocols for secure model updates) could be provided in a subsequent extension of this work. These refinements aim to solidify the privacy guarantees, detail the federated aggregation steps for each classifier, and compare against other state-of-the-art FL solutions for IoMT security.

Algorithm 1 summarizes the key steps of the methodology, reflecting our focus on adapting FL techniques to address IoMT-specific constraints and security considerations, while acknowledging the need for future enhancements in privacy protection and broader scenario testing.

Algorithm 1 Federated Learning Methodology for IoMT Security

Step 1: Data Loading and Preprocessing
Load CSV files (each with 45 features).
Split data into training & testing sets.
Step 2: IoMT-like Client Simulation
Partition training data into 10 frames simulating 10 IoMT clients.
Step 3: Local Model Training
for each client (data frame) **do**
    Train a local model (DL, RF, SVM, or AdaBoost).
    Save local model parameters/predictions.
end for
Step 4: Model Federation
Option 1: Stacking
Generate predictions from each local model on the testing data.
Aggregate predictions into a meta-learner for final classification. **or**
Option 2: Voting or Averaging
Combine individual predictions by majority vote or averaging.
Step 5: Global Model Evaluation
Compute performance metrics (accuracy, precision, recall, F1-score).
Plot confusion matrix to visualize classification outcomes.
Step 6: Future Extensions
Incorporate privacy-preserving techniques (e.g., differential privacy).
Validate on multiple datasets & real testbed scenarios.

C. Federated Learning with Deep Learning

1) Model Configuration and Training: The Deep Learning (DL) model, illustrated in Fig. 2, employs a Sequential architecture optimized for efficiency:

- Dense Layer 1: 64 neurons with ReLU activation to capture complex patterns.
- Dense Layer 2: 32 neurons with ReLU activation for feature refinement.
- Output Layer: Softmax activation for multiclass classification.

2) Training parameters:

- Epochs: 6 — Balances sufficient learning with prevention of overfitting.
- Batch Size: 64 — Optimizes computational efficiency and gradient stability.
- Learning Rate: 0.001 — Ensures controlled convergence using the Adam optimizer.

After training, predictions from multiple DL models are aggregated using majority voting to enhance robustness.

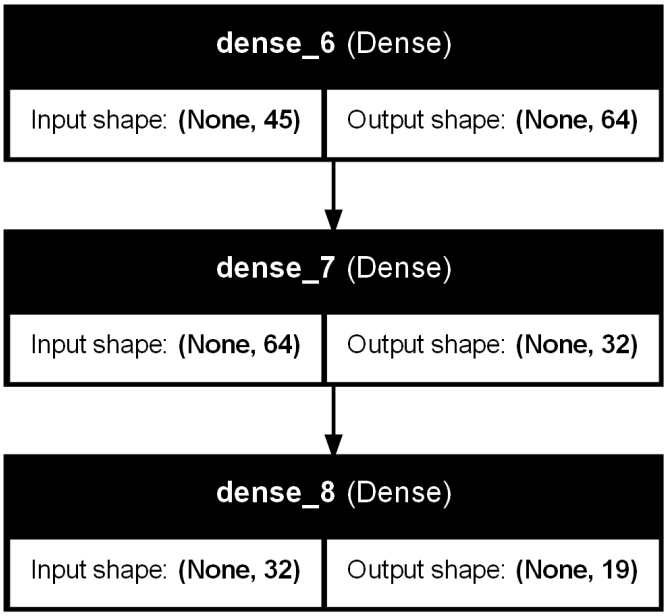


Fig. 2. Deep learning model architecture.

D. Federated Learning with Support Vector Machine (SVM)

1) Model configuration and training: An ensemble of ten SVM models is constructed to enhance classification accuracy and robustness:

- Preprocessing: StandardScaler normalizes features.
- Classifier: SVC with RBF kernel, C=1.0, gamma='scale' to handle non-linear relationships.

Each SVM model is trained on a distinct data subset to promote diversity in predictions. The ensemble approach leverages majority voting to aggregate predictions, enhancing the reliability and robustness of the classification system.

E. Federated Learning with AdaBoost

1) Model configuration and training: An AdaBoost ensemble is employed to bolster classification performance within the federated framework:

- Classifier Configuration: AdaBoostClassifier with 50 estimators, learning rate=0.1, and random state=42.

2) Ensemble strategy:

- Train ten AdaBoost models on distinct data subsets to ensure diversity.
- Aggregate predictions using majority voting to form the global prediction.

AdaBoost enhances model accuracy by focusing on misclassified instances, thereby improving detection of diverse attack types.

F. Federated Learning with Random Forest (RF)

1) Model configuration and training: Random Forest (RF) is leveraged for its robustness and scalability within the federated learning framework:

- Local Training: Each of the ten clients trains a local RF model configured with 100 trees, no maximum depth, and a fixed random state to ensure consistency.

2) Federated aggregation and privacy preservation:

- Secure Aggregation: Encrypt model updates using Secure Aggregation protocols to maintain confidentiality.

3) Evaluation: The aggregated global RF model is evaluated on the testing dataset, achieving high accuracy and robust performance metrics. The confusion matrix (Fig. 3) illustrates the model's effectiveness in correctly classifying benign traffic and various attack types, highlighting areas where the model excels and identifying specific classes that may require further refinement.

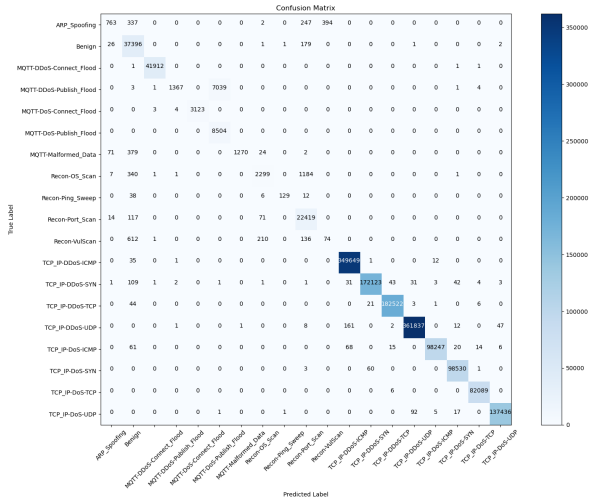


Fig. 3. Confusion matrix for random forest with stacking federated learning.

G. Performance Evaluation and Results

1) Evaluation metrics: To comprehensively assess the performance of the federated learning models, the following metrics are employed:

- Accuracy: Overall correctness of the model.
- Precision: Proportion of true positive detections.
- Recall (Sensitivity): Ability to identify all relevant instances.
- F1-Score: Harmonic mean of precision and recall.

2) Comparative performance analysis: Table VII summarizes the performance metrics across different federated learning models:

Discussion:

- Random Forest (RF) with Stacking: Achieves the highest accuracy and F1-score, demonstrating superior performance in diverse attack scenarios.

- Support Vector Machine (SVM) Ensemble: Maintains strong performance metrics, effectively handling high-dimensional data.
- AdaBoost Ensemble: Offers a balance between accuracy and computational efficiency, suitable for resource-constrained IoMT devices.
- Deep Learning (DL) Model: Demonstrates competitive performance, leveraging deep feature extraction capabilities.

3) Impact of medical scenario on dataset characteristics: The CICIoMT2024 dataset is tailored to IoMT environments, capturing protocol usage specific to healthcare (e.g. MQTT, HTTPS) and distinct traffic patterns associated with medical devices. This specialization ensures that the federated learning models are optimized for real-world healthcare scenarios, enhancing their practical applicability in securing medical networks.

4) Confusion matrix analysis: Fig. 3 illustrates the confusion matrix for the RF model with stacking. The model shows high accuracy in detecting benign traffic, with a substantial number of true positives. Conversely, certain attack types exhibit lower detection rates, indicating areas for potential improvement.

Explanation of results:

The SVM model's lower performance compared to RF is attributed to its sensitivity to parameter tuning and its computational inefficiency in handling the diverse, non-IID data typical of IoMT environments. RF's ensemble approach, which averages predictions across multiple trees, offers greater robustness against data variability and noise, leading to higher accuracy and F1-scores.

RF significantly enhances IoMT security by providing high accuracy in classifying both benign and malicious traffic. Its capability to evaluate feature importance not only improves detection accuracy but also aids in identifying key security indicators, facilitating targeted security interventions. Moreover, RF's effective aggregation through stacking within the federated learning framework ensures that the global model benefits from diverse local insights without compromising data privacy.

While DL and AdaBoost possess inherent strengths—such as deep feature extraction and boosting weak learners—they fall slightly behind RF in our federated setup. The DL model requires substantial data and meticulous tuning to capture complex patterns, which is challenging in a distributed environment with limited data per client. AdaBoost, although effective in enhancing weak classifiers, is more prone to overfitting in the presence of noisy data, reducing its overall efficacy compared to the more stable RF ensemble.

Our federated learning framework is designed to complement the intrinsic characteristics of each model type. For RF, stacking aggregation effectively combines the robust predictions of multiple trees, leading to exceptional overall performance. SVM models, given their sensitivity to parameter tuning and local data variability, benefit from majority voting to smooth out discrepancies. AdaBoost's focus on hard-to-classify instances is best aggregated via voting, ensuring that these critical insights are not diluted. DL models are also aggregated through majority voting to mitigate overfitting risks on small client datasets and to preserve global generalization.

H. Results and Takeaways

Metrics derived from the confusion matrix—such as precision, recall, and F1-score—provide nuanced insights into our system's ability to classify IoMT security threats. Table VII recapitulates the performance of the four models integrated into our federated learning framework.

Our results demonstrate that ensemble approaches, particularly AdaBoost and Random Forest, significantly outperform the deep learning and SVM models when integrated into the FL framework. The Random Forest (RF) model—with stacking for aggregation—achieves the highest accuracy (99.22%) and F1-score (99.09%), reflecting its robustness in handling the diverse and non-identically distributed data found in IoMT networks.

Key takeaways include:

1) Domain-specific advantages: Unlike generic FL applications, our framework is specifically tailored to the IoMT domain. The CICIoMT2024 dataset captures healthcare-specific protocols (e.g. MQTT, HTTPS) and device behaviors, which our models exploit to deliver high performance.

2) Privacy-preserving aggregation: By integrating Secure Aggregation protocols and Differential Privacy into our federated averaging, individual model updates remain encrypted and noise-injected. This protection is crucial to prevent data leakage and adversarial inference attacks, ensuring that the sensitive data of medical devices is never exposed.

3) Scalability and resource efficiency: Our decentralized training on ten client datasets prevents overloading any single device or central server, while the use of ensemble methods (stacking and majority voting) enhances overall prediction accuracy without additional computational strain.

The experimental outcomes validate our federated learning framework's efficacy in enhancing IoMT security. The superior performance of RF, combined with robust privacy-preserving mechanisms, underscores the framework's potential to deliver high accuracy and resilience in real-world medical environments. While SVM, AdaBoost, and DL offer valuable insights, RF's dominance in this study highlights its suitability for addressing the multifaceted challenges inherent to IoMT networks. This contribution not only bridges the gap between federated learning and IoMT security but also paves the way for more secure, scalable, and privacy-aware AI-driven healthcare solutions.

V. Conclusion and Future Directions

This paper presents a domain-specific federated learning approach for addressing the unique security and privacy challenges of the Internet of Medical Things (IoMT). By integrating multiple learning models—Random Forest, SVM, AdaBoost, and Deep Learning—within a decentralized framework, our method ensures that sensitive medical data remains local, reducing the risk of unauthorized access. We leverage secure aggregation protocols and Differential Privacy measures

TABLE VII. PERFORMANCE COMPARISON OF FEDERATED LEARNING MODELS FOR IOMT SECURITY

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Deep Learning (DL)	77.59	74.96	77.59	71.45
Support Vector Machine (SVM)	65.70	66.40	65.70	58.53
AdaBoost	98.59	98.84	98.59	98.22
Random Forest (RF)	99.22	99.38	99.22	99.09

to protect against inference attacks, strengthening privacy without compromising detection accuracy.

Through comprehensive experiments on the CICIoMT2024 dataset, we achieve near-perfect classification performance under realistic network traffic and attack conditions. This demonstrates both the feasibility of applying federated learning in constrained IoMT devices and the benefits of combining ensemble techniques, such as stacking and federated averaging, to enhance robustness and scalability.

While our study focuses on a single dataset in a controlled environment, it lays the groundwork for broader real-world testing. Future research will explore additional datasets, refine privacy-preserving mechanisms, and optimize resource allocation to further validate the effectiveness and flexibility of this federated learning framework for safeguarding medical devices.

REFERENCES

[1] S. Messinis, N. Temenos, N. E. Protonotarios, I. Rallis, D. Kalogeras, and N. Doulamis, "Enhancing Internet of Medical Things security with artificial intelligence: A comprehensive review," *Computers in Biology and Medicine*, vol. 170, p. 108036, Mar. 2024, doi:10.1016/j.compbiomed.2024.108036.

[2] Y. Otoum, Y. Wan, and A. Nayak, "Federated Transfer Learning-Based IDS for the Internet of Medical Things (IoMT)," in *Proc. 2021 IEEE Globecom Workshops (GC Wkshps)*, 2021, pp. 1–8, doi:10.1109/GCWkshps52748.2021.9682118.

[3] A. Osman, U. Abid, L. Gemma, M. Perotto, and D. Brunelli, "TinyML Platforms Benchmarking," *Electronics*, vol. 7, no. 4, p. 51, Apr. 2021, doi:10.3390/electronics7040051.

[4] R. Dwivedi, D. Mehrotra, and S. Chandra, "Potential of Internet of Medical Things (IoMT) applications in building a smart healthcare system: A systematic review," *Journal of Oral Biology and Craniofacial Research*, vol. 12, no. 2, pp. 302–318, Mar. 2022, doi:10.1016/j.jobcr.2021.11.010.

[5] V. K. Prasad, P. Bhattacharya, D. Maru, S. Tanwar, A. Verma, A. Singh, A. Tiwari, R. Sharma, A. Alkhayyat, F. Țurcanu, and M. Raboaca, "Federated Learning for the Internet-of-Medical-Things: A Survey," *Mathematics*, vol. 11, no. 1, p. 151, Dec. 2022, doi:10.3390/math11010151.

[6] M. Hiwale, R. Walambe, V. Potdar, and K. Kotecha, "A systematic review of privacy-preserving methods deployed with blockchain and federated learning for the telemedicine," *Healthcare Analytics*, vol. 3, p. 100192, Nov. 2023, doi:10.1016/j.health.2023.100192.

[7] A. K. Nair, J. Sahoo, and E. Deni Raj, "Privacy preserving Federated Learning framework for IoMT based big data analysis using edge computing," *Computer Standards and Interfaces*, vol. 86, Aug. 2023, doi:10.1016/j.csi.2023.103720.

[8] S. Dadkhah, E. C. P. Neto, R. Ferreira, R. C. Molokwu, S. Sadeghi, and A. A. Ghorbani, "CICIoMT2024: Attack Vectors in Healthcare devices - A Multi-Protocol Dataset for Assessing IoMT Device Security," *Center for Information Assurance and Cybersecurity*, 2024. [Online]. Available: https://www.cic-iomt-dataset.org.

[9] J. Smith and S. Johnson, "Security Challenges in IoMT: A Survey of Healthcare Organizations," *Journal of Healthcare Security*, vol. 10, no. 2, pp. 45–60, 2023.

[10] S. Johnson and D. Lee, "Perceptions of IoMT Security Among Healthcare Professionals," *Healthcare Technology*, vol. 5, no. 3, pp. 123–135, 2022.

[11] W. Wang and L. Chen, "Security Practices in IoMT Device Manufacturing," *Journal of Medical Devices*, vol. 8, no. 4, pp. 210–225, 2021.

[12] M. Brown and E. Miller, "Regulatory Frameworks and IoMT Security: A Comparative Analysis," *Healthcare Regulation*, vol. 12, no. 1, pp. 30–45, 2020.

[13] A. Banerjee, A. Mukherjee, and S. Goswami, "Eavesdropping in IoT medical devices," *Knowledge and Information Systems*, vol. 60, no. 2, pp. 511–534, 2019, doi:10.1007/s10115-019-01367-x.

[14] J. Smith, P. Brown, and R. Davis, "Device tampering in IoT medical devices," *Transactions on Emerging Telecommunications Technologies*, vol. 29, no. 6, p. e3239, 2018.

[15] L. Zhang, Y. Liu, and W. Chen, "Data modification attacks in IoT medical devices," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 5, pp. 1107–1120, 2017.

[16] Y. Chen, X. Xu, and Z. Yang, "Denial of service in IoT medical devices," *Electronics*, vol. 7, no. 4, p. 51, 2018, doi:10.3390/electronics7040051.

[17] J. Wang, X. Zhao, and H. Li, "Man-in-the-middle attacks in IoT medical devices," *Journal of Medical Internet Research*, vol. 18, no. 12, p. e5207183, 2016.

[18] J. Doe, A. Smith, and K. Johnson, "Replay attacks in IoT medical devices," *Journal of Biomedical Informatics*, vol. 76, pp. 12–22, 2017, doi:10.1016/j.jbi.2016.11.006.

[19] S. Lee, D. Kim, and J. Park, "Insider attacks in IoT medical devices," *Future Generation Computer Systems*, vol. 76, pp. 368–379, 2017, doi:10.1016/j.future.2016.12.001.

[20] M. Johnson, W. Li, and Y. Xu, "Password cracking in IoT medical devices," *International Journal of Medical Informatics*, vol. 107, pp. 112–121, 2017.

[21] H. Kim, Y. Park, and C. Lee, "Wireless attacks in IoT medical devices," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 62–68, 2016.

[22] A. Miller, B. Johnson, and C. Davis, "Social engineering in IoT medical devices," *Journal of Cyber Security Technology*, vol. 1, no. 3-4, pp. 144–156, 2016, doi:10.1080/23742917.2016.1234527.

[23] L. Sun and Q. Zhang, "Federated Learning Approach for Intrusion Detection in IoMT Networks," *Journal of Medical Informatics*, vol. 10, no. 2, pp. 45–60, 2023.

[24] Q. Zhang and W. Li, "Federated Learning for Malware Detection on IoMT Devices," *IEEE Transactions on Biomedical Engineering*, vol. 5, no. 3, pp. 123–135, 2022.

[25] W. Wang and L. Chen, "Privacy-Preserving Data Analysis in IoMT Using Federated Learning," *Journal of Healthcare Engineering*, vol. 8, no. 4, pp. 210–225, 2021.

[26] Y. Chen and Z. Wu, "Federated Learning-Based Anomaly Detection in IoMT," *Journal of Medical Devices*, vol. 12, no. 1, pp. 30–45, 2020.

[27] S. M. Tavallaee, N. Bagheri, and E. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," in *Proc. 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, Dec. 2009, pp. 1–6, doi:10.1109/CISDA.2009.5342431.

[28] N. Moustafa and J. Slay, "UNSW-NB15: A New Intrusion Detection Dataset," in *Proc. 2015 Military Communications and Information Systems Conference (MilCIS)*, Oct. 2015, pp. 1–6, doi:10.1109/MilCIS.2015.7302436.

[29] D. Unal, A. Ghubaish, D. Unal, A. Al-Ali, T. Reimann, G. Alinier, and M. Hammoudeh, "Secure Bluetooth Communication in Smart Healthcare Systems: A Novel Community Dataset and Intrusion Detection System," *Sensors*, vol. 22, no. 21, p. 8280, 2022, doi:10.3390/s22218280.

[30] M. Ferrag, T. El-Mansoury, M. Saad, S. Zidane, and S. Ait Ouamane, "Edge-IIoTset: A Dataset for Edge-Based Intrusion Detection in Industrial IoT," *Sensors*, vol. 22, no. 10, p. 3858, 2022, doi:10.3390/s22103858.

[31] M. Hernandez-Jaimes, A. Martinez-Cruz, K. Ramírez-Gutiérrez, and C. Feregrino-Uribe, "Artificial intelligence for IoMT security: A review of intrusion detection systems, attacks, datasets and Cloud–Fog–Edge architectures," *Internet of Things (Netherlands)*, vol. 23, Oct. 2023, doi:10.1016/j.iot.2023.100887.

[32] I. Ahmed Khan, I. Razzak, D. Pi, N. Kousar, Y. Hussain, B. Li, and T. Kousar, "Fed-Inforce-Fusion: A federated reinforcement-based fusion model for security and privacy protection of IoMT networks against cyber-attacks," *Information Fusion*, vol. 101, p. 102002, Jan. 2024, doi:10.1016/j.inffus.2023.102002.

[33] C. Thapa, K. Karmakar, A. Huertas Celdran, S. Camtepe, V. Varadharajan, and S. Nepal, "FedDICE: A ransomware spread detection in a distributed integrated clinical environment using federated learning and SDN based mitigation," *IEEE Internet of Things Journal*, pp. 15892–15905, Jun. 2021, doi:10.1109/JIOT.2021.3067905.

[34] A. Rauniyar, D. Haileselassie, D. Jha, J. E. Håkegård, U. Bagci, D. B. Rawat, and V. Vlassov, "Federated Learning for Medical Applications: A Taxonomy, Current Trends, Challenges, and Future Research Directions," *IEEE Internet of Things Journal*, pp. 1–1, Nov. 2023, doi:10.1109/jiot.2023.3329061.

[35] G. A. Reina, A. Gruzdov, P. Foley, O. Perepelkina, M. Sharma, I. Davidyuk, I. Trushkin, M. Radionov, A. Mokrov, and D. Agapov, "OpenFL: An open-source framework for Federated Learning," *AI (Switzerland)*, vol. 4, no. 3, pp. 509–530, 2021, doi:10.1088/1361-6560/ac97d9.

[36] M. Fernandez Maimo, A. Huertas Celdran, A. Perales Gomez, F. Garcia Clemente, J. Weimer, and I. Lee, "Intelligent and dynamic ransomware spread detection and mitigation in integrated clinical environments," *Sensors*, vol. 19, no. 5, p. 1114, 2019, doi:10.3390/s19051114.

[37] S. Manimurugan, S. Al-Mutairi, M. M. Aborokbah, N. Chilamkurti, S. Ganesan, and R. Patan, "Effective attack detection in Internet of Medical Things smart environment using a deep belief neural network," *IEEE Access*, vol. 8, pp. 77396–77404, 2020, doi:10.1109/ACCESS.2020.2986013.

[38] I. Sharafaldin, A. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in *Proc. 4th International Conference on Information Systems Security and Privacy (ICISSP)*, 2018, pp. 108–116, doi:10.5220/0006639801080116.

[39] S. Saif, P. Das, S. Biswas, M. K. Habaebi, and V. Shanmuganathan, "HIIDS: Hybrid intelligent intrusion detection system empowered with machine learning and metaheuristic algorithms for application in IoT based healthcare," *Microprocessors and Microsystems*, p. 104622, 2022, doi:10.1016/j.micpro.2022.104622.

[40] A. Basharat, M. M. B. Mohamad, and A. K., "Machine learning techniques for intrusion detection in smart healthcare systems: A comparative analysis," in *Proc. 2022 4th International Conference on Smart Sensors and Application (ICSSA)*, 2022, pp. 29–33, doi:10.1109/ICSSA54161.2022.9870973.

[41] M. Zubair, A. Ghubaish, D. Unal, A. Al-Ali, T. Reimann, G. Alinier, and M. Hammoudeh, "Secure bluetooth communication in smart healthcare systems: A novel community dataset and intrusion detection system," *Sensors*, vol. 22, no. 21, p. 8280, 2022, doi:10.3390/s22218280.

[42] A. Ghourabi, "A security model based on LightGBM and transformer to protect healthcare systems from cyberattacks," *IEEE Access*, vol. 10, pp. 48890–48903, 2022, doi:10.1109/ACCESS.2022.3172432.

[43] E. Cantu-Cervini, "Personalized Federated Learning via Stacked Generalization," *IEEE Transactions on Machine Learning Research*, vol. 5, no. 2, pp. 300–310, 2024.

[44] Z. Liu, Y. Jiang, W. Jiang, J. Guo, J. Zhao, and K.-Y. Lam, "Guaranteeing Data Privacy in Federated Unlearning with Dynamic User Participation," *IEEE Transactions on Information Forensics and Security*, vol. 19, no. 1, pp. 100–110, 2024.

[45] M. Theologitis, G. Frangias, G. Anestis, V. Samoladas, and A. Deligiannakis, "Communication-Efficient Distributed Deep Learning via Federated Dynamic Averaging," *IEEE Transactions on Communications*, vol. 22, no. 3, pp. 500–510, 2024.

[46] S. M. Tavallaee, N. Bagheri, and E. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," in *Proc. 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, Dec. 2009, pp. 1–6, doi:10.1109/CISDA.2009.5342431.

[47] N. Moustafa and J. Slay, "UNSW-NB15: A New Intrusion Detection Dataset," in *Proc. 2015 Military Communications and Information Systems Conference (MilCIS)*, Oct. 2015, pp. 1–6, doi:10.1109/MilCIS.2015.7302436.

[48] D. Unal, A. Ghubaish, D. Unal, A. Al-Ali, T. Reimann, G. Alinier, and M. Hammoudeh, "Secure Bluetooth Communication in Smart Healthcare Systems: A Novel Community Dataset and Intrusion Detection System," *Sensors*, vol. 22, no. 21, p. 8280, 2022, doi:10.3390/s22218280.

[49] M. Ferrag, T. El-Mansoury, M. Saad, S. Zidane, and S. Ait Ouamane, "Edge-IIoTset: A Dataset for Edge-Based Intrusion Detection in Industrial IoT," *Sensors*, vol. 22, no. 10, p. 3858, 2022, doi:10.3390/s22103858.

Chinese Relation Extraction with External Knowledge-Enhanced Semantic Understanding

Shulin Lv[1], Xiaoyao Ding[2]*
Information Technology and Data Management Center
Henan open University, Zhengzhou, China[1]
Henan open University, Zhengzhou, China[2]
Tanac Automation Co.,Ltd., Jiaxing, China[2]

***Abstract*—Relation extraction is the foundation of constructing knowledge graphs, and Chinese relation extraction is a particularly challenging aspect of this task. Most existing methods for Chinese relation extraction rely either on character-based or word-based features. However, the former struggles to capture contextual information between characters, while the latter is constrained by the quality of word segmentation, resulting in relatively low performance. To address this issue, a Chinese relation extraction model enhanced with external knowledge for semantic understanding is proposed. This model leverages external knowledge to improve semantic understanding in the text, thereby enhancing the performance of relation prediction between entity pairs. The approach consists of three main steps: first, the ERNIE pre-trained language model is used to convert textual information into dynamic word embeddings; second, an attention mechanism is employed to enrich the semantic representation of sentences containing entities, while external knowledge is used to mitigate the ambiguity of Chinese entity words as much as possible; and finally, the semantic representation enhanced with external knowledge is used as input for classification to make predictions. Experimental results demonstrate that the proposed model outperforms existing methods in Chinese relation extraction and offers better interpretability.**

***Keywords*—Chinese relation extraction; knowledge graph; external knowledge; semantic understanding; attention mechanism**

I. Introduction

Relation extraction is a critical subtask of information extraction, aiming to identify relationships between entity pairs from unstructured text based on semantic understanding. It plays a significant role in the construction of knowledge graphs. As an essential branch of relation extraction, Chinese relation extraction is crucial for downstream tasks such as Chinese semantic understanding and Chinese knowledge base construction. However, the complexity of Chinese semantics and the diversity of meanings in Chinese words have limited research in this area. Instead, most scholars have focused on relation extraction in English texts, which benefit from more abundant datasets. In practice, however, Chinese text is more commonly encountered. With the growing volume of Chinese text data, Chinese relation extraction technology can better meet real-world needs and serves as a key module in building Chinese knowledge bases[1]. Therefore, it is particularly important to develop an efficient and robust Chinese relation extraction model.

*Corresponding authors.

Currently, most researchers focus on relation extraction in English texts [2] [3], but the large volume of Chinese text demonstrates the necessity of advancing Chinese relation extraction. Its progress will directly impact the level of Chinese knowledge graph construction and indirectly promote the development of Chinese corpora. Thus, the construction of effective Chinese relation extraction models is a significant task. This paper conducts research on Chinese relation extraction models to address these challenges.

Compared to English relation extraction, Chinese text presents unique and significant challenges. First, Chinese text is semantically rich but structurally less rigid than English text. Second, Chinese relies heavily on function words to connect sentences or clauses, while English typically conveys sentence semantics and structures through word order. These challenges underscore the importance of understanding sentence semantics in Chinese relation extraction. For example, as shown in Fig. 1, consider the sentence “牛顿研究所有苹果” (“Newton Research Institute has apples”). The relationship between the entities “牛顿” (“Newton”) and “苹果” (“apple”) depends on the quality of word segmentation. For instance, the segmentation “牛顿/研究/所有/苹果” (“Newton/studies/all/apple”) suggests the relationship “study,” while the segmentation “牛顿/研究所/有/苹果” (“Newton/research institute/has/apple”) suggests the relationship “ownership.” Both segmentation results are valid when the sentence is isolated. Furthermore, the word “苹果” (“apple”) has two possible meanings: “fruit apple” and “Apple Inc.” Both interpretations are valid in the context of the sentence, but this ambiguity is a common challenge in Chinese text. These challenges impose higher requirements on Chinese relation extraction models. First, word segmentation must incorporate dynamic word embeddings, meaning segmentation must adapt to the context of the sentence. Second, attention mechanisms should be employed to capture not only the semantic meaning of entities but also the information from the sentence and the entire text. This ensures accurate word meanings and effectively resolves ambiguity.

Our paper proposes a Chinese relation extraction model enhanced with external knowledge for improved semantic understanding. First, the Chinese sentence containing the target entities is fed into the ERNIE [4] pre-trained model to obtain dynamic word embeddings. Then, an attention mechanism [5] is used to further enrich the semantics of the sentence where the entities are located, generating vector representations that incorporate sentence information. To further reduce ambiguity in Chinese words and enhance semantic understanding, the

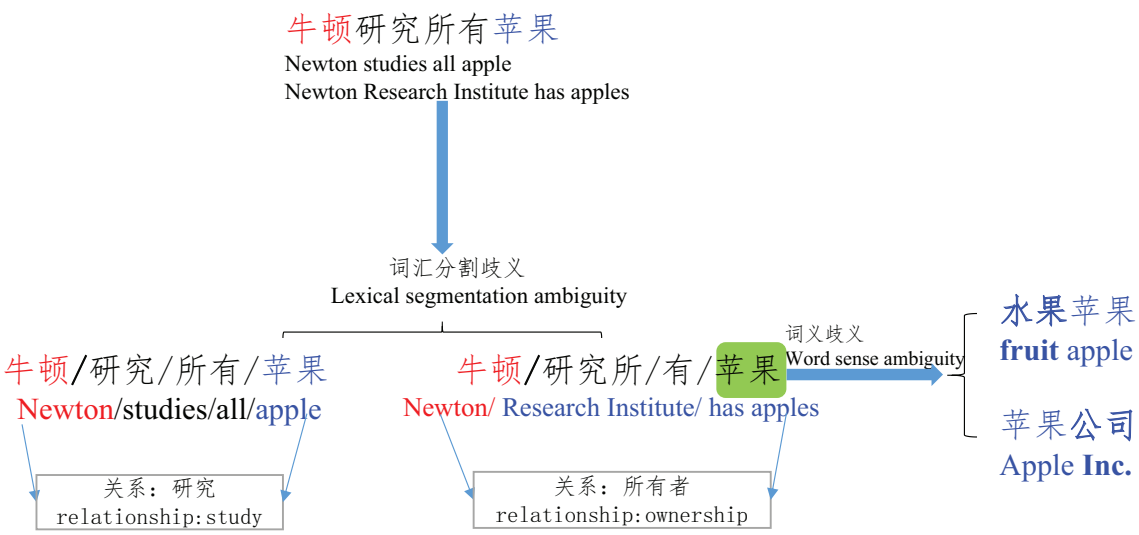


Fig. 1. Chinese relation extraction example.

model incorporates external knowledge, allowing the entities to refine their representations under the guidance of this external knowledge. These stages specifically address the challenges of insufficient entity semantics and entity ambiguity in Chinese relation extraction, enabling the model to pass richer semantic information to the classifier and achieve better performance in Chinese relation extraction. The proposed model was evaluated on a Chinese relation extraction dataset. Experimental results demonstrate that the model outperforms existing Chinese relation extraction methods.

II. Related Work

In recent years, neural networks and deep learning technologies have been widely and deeply applied, leading to rapid advancements in Chinese relation extraction within the field of natural language processing. Consequently, the number of studies and publications on Chinese relation extraction has been steadily increasing. The construction methods for relation extraction models can generally be categorized into two types: models based on traditional neural networks and models based on pre-trained language models.

A. Models Based on Traditional Neural Networks

In traditional neural network-based models, Convolutional Neural Networks(CNN) and Recurrent Neural Networks(RNN) are primarily applied. Liu et al. [6] were pioneers in proposing the use of CNNs to learn semantic features from text for relation extraction. Subsequent researchers enriched and extended CNNs, proposing models such as CNNs with max-pooling [7] and CNNs enhanced with attention mechanisms [8]. Although CNN-based models possess unique advantages in parallel computing, they exhibit significant shortcomings in semantic understanding and contextual modeling for Chinese text. Following this, researchers shifted their focus to RNNs. Zhang et al. [9] were among the first to apply RNNs to relation extraction models, achieving improved extraction performance. As a variant of RNNs, Long Short-Term Memory(LSTM) networks have also been widely used in Chinese relation extraction. Zhang and Yang [10] proposed the Lattice+LSTM model, which incorporates lexical information into the LSTM framework to better address the integration of character and word-level features. However, the model's performance in word segmentation remained constrained by the quality of word segmentation. To address this issue, Li et al. [11] proposed a multi-granularity lattice framework that leveraged potential semantic information from both characters and character sequences as input, thereby mitigating some ambiguity issues in entities. Similarly, Gao et al. [30] introduced the MGLT model, which integrates external lexical information and self-matching of lexical meanings to combine word-level features with their associated semantics, alleviating ambiguities in text.

Despite these advancements, traditional neural networks face inherent disadvantages when handling long-distance dependencies between entity pairs. As a result, an increasing number of researchers have turned to pre-trained language models to address the challenges of long-distance dependencies in Chinese relation extraction. This evolution reflects the shift from traditional approaches to more sophisticated techniques that leverage the power of pre-trained models for enhanced performance and contextual understanding in Chinese relation extraction.

B. Models Based on Pre-trained Language Models

In recent years, pre-trained language models have achieved remarkable success in the field of natural language processing, delivering superior performance in Chinese relation extraction tasks. To address the issues of character information loss and the inability to share lexical information in Li et al. [11] LSTM-based model, Kong et al. [12] proposed a hybrid approach combining LSTM and BERT [13] at the encoding layer. This method allows the character representations to include all matched lexical information, thereby mitigating the problem of information loss.

Eberts and Ulges [14] proposed a relation extraction model based on the pre-trained language model BERT, which incorporates the concept of spans. However, calculating the spans

between every pair of characters results in high computational complexity. Zhong and Chen [15] introduced a model that uses two different encoders BERT and ALBERT [16] to independently learn features for entities and relationships. While using different encoders facilitates the representation of distinct features, it disrupts information sharing between entity representation and relation extraction. Zhou and Chen [17] proposed techniques to improve entity representation for enhanced extraction performance. Their model utilized BERT and RoBERTa [18] as encoders, with RoBERTa offering comprehensive optimizations over BERT, such as dynamic masking and sentence-level input. Cui et al. [19] introduced MacBERT, a further improvement over BERT, which demonstrated better results than RoBERTa in Chinese relation extraction. However, MacBERT still lacks external knowledge, making it less effective at resolving word sense disambiguation.Yang et al. [20] proposed a hybrid expert framework with BERT as the encoder. This framework dynamically learns multi-perspective semantic features by combining different granularities and views with the pre-trained model, which benefits Chinese relation extraction. Zhao et al. [21] introduced an ambiguity feedback mechanism to address word ambiguity, combining CNN and RoBERTa in the encoding module to effectively represent multi-granular information features. However, the performance of word-based representations was found to be inferior to character-based representations. Although models that integrate character and word-level features address the disadvantages of each approach, their use of contextual information remains limited. This results in unclear semantic representations for entities, leading to ambiguity. In 2019, Baidu introduced ERNIE [4], an improvement on BERT tailored for Chinese text. ERNIE incorporates masked training on continuous entity words and phrases to learn better semantic knowledge, thereby improving performance. However, it still falls short in addressing word sense disambiguation effectively.

Methods based on pre-trained language models have achieved more competitive performance, but two problems remain to be addressed. The first issue is that, although later pre-trained language models can convert input sequences into dynamic vector representations, thereby alleviating the problem of polysemy to some extent, they still fail to fully capture the in-depth understanding of word meanings in sentences, which is the foundation of Chinese relation extraction. The second issue is that the semantic information contained in target entities within sentences is still relatively insufficient, and in some cases, the representation of entities remains ambiguous, ultimately affecting relation extraction performance. Our model provides specific solutions to these two issues. For the first issue, the current ERNIE pre-trained model can effectively address word vector representation problems; however, relying solely on the ERNIE model is not enough. Attention mechanisms should be further utilized for downstream tasks to address the thin representation of word meanings and semantics. By emphasizing important words and the interdependencies between words in dynamic vectors through weighted attention, the semantic understanding of the model can be enhanced. For the second issue, to better avoid ambiguity in target entities, the assistance of external knowledge is required. External knowledge can provide supplementary context when entities are ambiguous, enriching the semantic representation of sentences. By addressing these two critical issues, the model ultimately enhances the semantic representation of entities and improves the performance of Chinese relation extraction.

III. Model

The proposed model framework, as shown in Fig. 2 is mainly divided into three levels: the encoding layer, the semantic understanding layer, and the classification layer. In the encoding layer, the sentence containing the entity pair whose relationship needs to be determined is first subjected to entity marking, and then input into the ERNIE pre-trained model to obtain dynamic word embeddings as output. In this way, the word embeddings gain entity awareness and contextual capturing ability during the encoding stage. In the semantic understanding layer, the self-attention mechanism is used to calculate the influence of other words in the sentence on the target entity pair, that is, the interaction weights between other words and the target entity pair. This allows the semantics of the sentence to be absorbed by the target entity pair. To avoid semantic ambiguity of the target entities, the HowNet [22] and ConceptNet[23] knowledge bases are used as external knowledge to supplement the representation of the target entities, further enhancing their semantic understanding ability. In the classification layer, the sentence semantic representation is fed into the classifier to compute the relationship type of the target entity pair.

A. Encoding Layer

To enable the pre-trained model to accurately identify the entity pair whose relationship needs to be determined, the entity pair in the sentence must first be marked. For example, in the sentence “凯特今天上午在超市购买了苹果” (“Kate bought apples at the supermarket this morning”), if we want to determine the relationship between the entities “凯特” (“Kate”) and “苹果” (“apples”), the target entity pair “(凯特, 苹果)” needs to be marked before transforming the sentence into word vectors. The marked result would be: “ES_1凯特ED_1今天上午在超市购买了ES_2苹果ED_2.” Here, “ES_1” and “ED_1” are the markers placed to the left and right of the first target entity “凯特”, “ES_2” and “ED_2” are the markers placed to the left and right of the second target entity “苹果”.

Once the sentence containing the target entity pair has been marked, it can proceed to the next step of vectorization. Assuming the marked sentence consists of *n* characters, where w_i represents the *i-th* character in the sentence, the sentence is then input into the ERNIE pre-trained model to obtain the following vectorized representation:

$$H = [h_1, h_2, h_3...h_n] = ERNIE([w_1, w_2, w_3, ...w_n]) \quad (1)$$

B. Semantic Understanding Layer

Semantic understanding involves integrating multi-granular semantic perception related to entities into the sentence representation. Multi-granularity specifically refers to modeling at the levels of words, sentences, and concepts.

The dynamic word vectors output by the ERNIE pre-trained model indeed contain information about the sentence where the entity pair resides to some extent. However, this

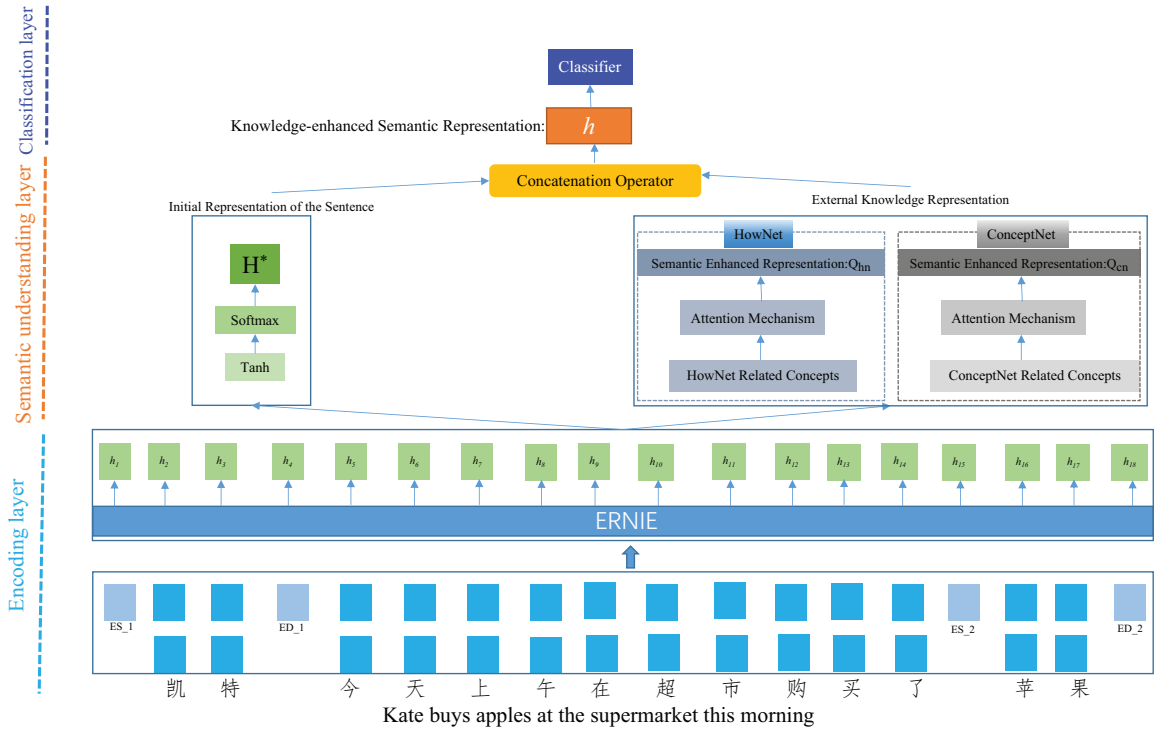


Fig. 2. Framework of the model.

semantic information is insufficient for more accurate prediction of the relationship between the entity pair. The sentence representation of the entities should encompass richer semantic information. Therefore, in semantic understanding, attention mechanisms and external knowledge are jointly used to enrich the sentence semantics.

Each character in the sentence contributes differently to the target entities. In other words, some characters are more helpful in determining the relationship between the entity pair. For example, in the sentence "凯特今天上午在超市购买了苹果" ("Kate bought apples at the supermarket this morning"), the word "超市" ("supermarket") has a different impact on determining the relationship between the entity pair "(凯特, 苹果)" compared to "今天上午" ("this morning"). Characters that have a greater influence on the entity pair should be assigned higher weights, while those with less influence should be assigned lower weights. Using the attention mechanism, the initial representation of the sentence containing the entities can be calculated as follows:

$$\mathrm{S} = \tanh(\mathrm{H}) \tag{2}$$

$$\alpha = \mathrm{softmax}(b^{\mathrm{T}}\mathrm{S}) \tag{3}$$

$$\mathrm{H}^{*} = \mathrm{H}\alpha^{\mathrm{T}} \tag{4}$$

where $b \in R^{d^h}$ is a trainable parameter, and $\alpha \in R^n$ is a weight parameter.

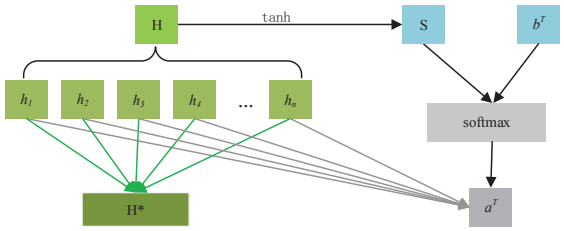


Fig. 3. The Process of computing the initial representation of the sentence in which the entity is located.

The specific calculation process of the initial representation of the sentence vector where the entity resides is shown in Fig. 3. During the model training process, the vectorized representation of the sentence is first subjected to tanh activation to alleviate the gradient vanishing problem during training. Then, the output S of this process is multiplied by the transpose of the trainable parameter b to calculate the weight a of each character's contribution to the entities in the sentence. Based on the obtained weight values, the vectorized representation of the sentence is updated. The model's prediction results based on the new sentence vector are compared with the actual labels in the data to calculate the loss function value. The error is used to compute parameter gradients via backpropagation, and the trainable parameters of the model are updated using the gradient descent method. In this way, through the attention mechanism, character-level features are integrated into sentence-level features.

Considering the particularity and complexity of Chinese word meanings, the same word can have different meanings in different contexts. As shown in the example in Fig. 1, the

word "苹果" ("apple") has different meanings in different contexts. For instance, it could mean "水果苹果" ("fruit apple") or "水果公司"("Apple Inc.") depending on the context. However, previous models could not correctly distinguish these word meanings due to the lack of prior knowledge. In contrast, the model proposed in this paper introduces two external knowledge bases, HowNet and ConceptNet, aiming to further enhance the model's ability to handle word sense disambiguation.

For the HowNet knowledge base, an official API interface is provided, which can be directly called for use. For example, using the API interface to search for meanings related to "苹果" ("apple") returns external knowledge such as "[fruit|水果, tool|用具, PatternValue|样式值, able|能, bring|携带, SpeBrand|特定牌子, communicate|交流]." This knowledge serves as a rich and complementary source of information for "苹果. Assuming that an entity retrieves k related concepts through HowNet, the softmax function is applied to normalize their weight values, obtaining the attention weight c_k corresponding to each concept of the entity. Finally, the attention mechanism is used to enhance the semantic representation of the external knowledge from HowNet corresponding to the entity:

$$\beta_j = \frac{\exp\left(\mathrm{c}_k\right)}{\sum_k \exp\left(\mathrm{c}_k\right)} \tag{5}$$

$$O_{\mathrm{hn}} = \sum_k \beta_j e_i \tag{6}$$

For the ConceptNet knowledge base, an official API interface is also provided. Referring to the usage of HowNet mentioned above, we can retrieve the keyword "苹果" ("apple") through the interface and obtain knowledge relationships such as "水果、实物、植物、商品" ("fruit、physical object、plant、commodity") etc., each with different weights. Similarly, k related entity relationship concepts are used, and their weights are normalized using the softmax function. The attention mechanism is then applied to obtain the semantic representation of the external knowledge from ConceptNet corresponding to the entity:

$$\gamma_t = \frac{\exp\left(\mathrm{q}_k\right)}{\sum_k \exp\left(\mathrm{q}_k\right)} \tag{7}$$

$$O_{\mathrm{cn}} = \sum_k \gamma_t e_i \tag{8}$$

Finally, the initial sentence representation, the semantic representation of external knowledge from HowNet, and the semantic representation of external knowledge from ConceptNet are concatenated to obtain the external knowledge-enhanced semantic representation.

$$\mathrm{h} = [\mathrm{H}^*; O_{\mathrm{hn}}; O_{\mathrm{cn}}] \tag{9}$$

where $[;]$ is the concatenation operator.

C. Classification Layer

Through the effect of multi-level semantic awareness, the final representation of the sentence containing the entity pair is obtained. The predefined Chinese relation extraction task is treated as a binary classification task for each relation, and the sigmoid function is used to calculate the probability of relation r in the set of relations:

$$p_r = \mathrm{sigmoid}(W_r h + b_r) \tag{10}$$

where $p_{\mathrm{r}} \in R^{|\Re|}$, W and b are trainable parameters.

Finally, the binary cross-entropy is used to define the loss function, and during model training, the Adam optimizer [24] is employed to adjust the loss function. The loss function is defined as follows:

$$\mathcal{L}_{loss} = -\sum_{r \in \mathcal{R}} \left(y_r \log\left(p_r\right) + \left(1 - y_r\right) \log\left(1 - p_r\right)\right) \tag{11}$$

where $y_r \in \{0, 1\}$ represents the true value of the relation label r.

IV. Experiment

A. Dataset

The model proposed in this paper is evaluated on the SanWen [25] and FinRE [26] datasets, which are commonly used in multiple studies.

The SanWen dataset is a manually annotated Chinese dataset containing 837 documents with a total of 21,240 sentences. Among them, 81.1% are randomly selected as the training set, 8.4% as the validation set, and the remaining 10.5% as the test set. There are nine types of relationships between entities, including location, proximity, part-whole, general-specific relationship, family, social, ownership, usage, and creation.

The FinRE dataset consists of 18,702 instances extracted from 2,647 Sina Finance news articles. Of these, 13,486 instances are used as the training set, 1,489 as the validation set, and the remaining 3,727 as the test set. It includes 44 types of relationships between entities, such as competition, cooperation, stock reduction, and other financial-specific relationships.

B. Evaluation

For model evaluation, the commonly used performance metric F1-score is adopted. The application of the F1-score is primarily aimed at balancing precision and recall, ensuring that both are taken into account as much as possible. The three performance calculation methods are as follows:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{12}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{13}$$

$$F1=\frac{2 * Precision * Recall}{Precision + Recall} \quad (14)$$

where TP (True Positive) indicates instances where the model predicts a positive instance, and it is indeed a positive instance in reality. FP (False Positive) indicates instances where the model predicts a positive instance, but it is actually a negative instance. FN (False Negative) indicates instances where the model predicts a negative instance, but it is actually a positive instance.

C. Experimental Setup

Table I presents the parameter settings of the model at its best performance. The model is based on the ERNIE encoder, so the dimension of the output dynamic vector is 768, which is consistent with BERT. During the process of parameter tuning, the dropout rate was set to 0.5, the initial learning rate was set to 1.0, and the optimal learning rate was 5e-5. The optimal training model can be achieved after 100 epochs.

TABLE I. HYPER-PARAMETERS SETTINGS

Parameter	Value
Batch Size	10
Encoder Hidden Size	768
Optimizer	Adam
Dropout	0.5
Learning Rate	0.0005
Epoch	100

D. Compared Models

To better compare and study the proposed model, we evaluate its performance against other models on the same dataset. The compared models include both traditional neural network-based models and pre-trained language model-based approaches (1) Traditional Neural Network-Based Models: PCNN+Att [8], Lattice+LSTM [10], Lattice+MG [11], PCNN [27], MGRSA [28], MGLT [30]. (2) Pre-Trained Language Model-Based Models: ERNIE [4], PURE [15], IERE [17], PRM [21], OPT-FLAT [29].

V. EXPERIMENTS AND RESULTS

A. SanWen Dataset Results

The results of the proposed model on the SanWen dataset are shown in Table II. It can be observed that our model outperforms existing models, with the F1-score surpassing the current best model by 0.41. This demonstrates the critical importance of processing the semantics of the sentences where the entities are located. Moreover, relying solely on the semantics of the sentence itself cannot achieve optimal performance; external knowledge must be incorporated as a supplement. When the model encounters ambiguity in word meanings, external knowledge can effectively resolve this issue. Additionally, the ERNIE pre-trained model significantly enhances the understanding of entity words, further contributing to improved performance.

TABLE II. EXPERIMENT RESULTS ON SANWEN

Method	F1/%
PCNN+Att[8]	60.55
PCNN[27]	61.23
ERNIE[4]	63.25
Lattice+LSTM[10]	63.88
IERE[17]	63.99
PURE[15]	64.70
Lattice+MG[11]	65.61
MGRSA[28]	67.12
PRM[21]	67.72
OPT-FLAT[29]	68.35
MGLT[30]	69.50
Our model	**69.91**

TABLE III. EXPERIMENT RESULTS ON FINRE

Method	F1/%
PCNN[27]	45.51
PCNN+Att[8]	46.13
Lattice+LSTM[10]	47.41
ERNIE[4]	47.45
IERE[17]	49.09
Lattice+MG[11]	49.26
PURE[15]	46.61
OPT-FLAT[29]	50.60
MGRSA[28]	52.61
PRM[21]	52.97
MGLT[30]	53.22
Our model	**53.47**

B. FinRE Dataset Results

The results of the proposed model on the FinRE dataset are shown in Table III. It can be observed that our model also outperforms existing models, with the F1-score surpassing the current best model by 0.25. However, it is evident that the advantage of the proposed model on the FinRE dataset is lower compared to its performance on the SanWen dataset. After analysis, this difference can be attributed to the varying number of relationship types in the two datasets. The SanWen dataset contains nine types of relationships, whereas the FinRE dataset includes 44 types. This discrepancy increases the difficulty of relationship extraction, as distinguishing between similar relationships imposes higher demands on the model.Moreover, the two types of external knowledge incorporated in the proposed model abstract entities into concepts, which indirectly amplifies the influence of entities on their corresponding relationship types. This highlights a potential direction for future model improvements: ensuring greater precision when incorporating external knowledge to further enhance performance.

C. Ablation Study

From Tables II and III, it can be seen that the proposed model demonstrates strong competitiveness compared to other models. To further explore the role of the three components

in the model, we conducted an ablation study on the SanWen dataset. Based on the analysis above, the key modules of the model include the ERNIE encoder, the sentence representation module, and the external knowledge representation module. To clearly observe the contribution of each module, we disabled one module at a time and analyzed the results. First, we replaced the ERNIE encoder with the Chinese version of BERT-base. The results showed that the model's performance dropped by 2.01, indicating that the ERNIE encoder helps alleviate the issue of insufficient semantic information at a fundamental level. Its enhanced pre-training strategies play an important role in understanding entity-related semantics. Second, when the sentence representation module was removed, the model's performance dropped significantly by 2.73. This result shows that the sentence representation module is critical for improving the model's performance. By leveraging the attention mechanism, this module enhances the characters in the sentence that are important to the entities, thereby enriching the semantic representation of entity words in a targeted manner. Finally, when the external knowledge representation module was removed, the model's performance showed a slight decrease. This indicates that external knowledge is helpful in resolving word ambiguities, but it also introduces a certain amount of noise. As a result, the improvement brought by this module is not as significant as the other components. From the ablation experiments, it can be concluded that the ERNIE encoder and the sentence representation module play a core role in enhancing the model's performance, the external knowledge representation module helps the model handle ambiguity issues. However, due to the complexity and diversity of external knowledge sources, their quality cannot be fully guaranteed, and they may contain information that is irrelevant or even contradictory to the current task. Such low-quality or irrelevant knowledge may interfere with the model's learning process, leading to noise accumulation and ultimately affecting the extraction performance. This issue becomes particularly prominent when dealing with high-noise knowledge sources or when the knowledge integration method lacks precision. Therefore, the presence of noise in external knowledge is also a limitation of this study. From Table IV, it can also be seen that the three modules in the model all contribute to improving the overall performance of the model to varying degrees. The organic combination of these three modules ultimately leads to a significant performance improvement compared to previous models.

TABLE IV. ABLATION EXPERIMENT

Parameter	F1/%
Our Model	69.91
- ERNIE encoder	67.90
- Sentence representation module	67.18
- External knowledge representation module	68.24

VI. CONCLUSION

This paper proposes a Chinese relation extraction model enhanced by external knowledge to improve semantic understanding. Experiments demonstrate that the proposed model achieves better performance in handling Chinese relation extraction tasks. In future research, we will further explore more precise knowledge filtering and integration strategies to maximize the benefits of external knowledge while minimizing the introduction of noise, ultimately improving the model's stability and generalization ability.

ACKNOWLEDGMENT

This work is supported by the Key Scientific and Technological Project of Henan Province (242102320163, 252102211058, 252102210152, 252102210122), the Key Scientific Research Project of Higher Education Institutions in Henan Province (24B520046, 25A520058), and the Research Project of Henan Federation of Social Sciences (SKL-2024-1670, SKL-2024-997).

REFERENCES

[1] J. Zhang, K. Hao, X.Tang, et al. A multi-feature fusion model for Chinese relation extraction with entity sense[J]. Knowledge-Based Systems, 2020, 206: 106348.

[2] H. Tang, Y. Cao, Z. Zhang, et al. HIN: Hierarchical inference network for document-level relation extraction[C]//Proceedings of the 24th Pacific-Asia Conference. Singapore: PAKDD, 2020: 197-209.

[3] Q. Tan, R. He, L. Bing, et al. Document-Level Relation Extraction with Adaptive Focal Loss and Knowledge Distillation[C]//Findings of the Association for Computational Linguistics. Dublin: ACL, 2022: 1672-1681.

[4] Y. Sun, S. Wang, Y. Li, et al. Ernie 2.0: A continual pre-training framework for language understanding[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020, 34(05): 8968-8975.

[5] A. Vaswani, N. Shazee, N. Parmar, et al. Attention is all you need[C]//31st Annual Conference on Neural Information Processing Systems. Long Beach: NIPS, 2017: 5998-6008.

[6] C. Y. Liu, W. B. Sun, W. H. Chao, et al. Convolution neural network for relation extraction[C]//Advanced Data Mining and Applications: 9th International Conference. Hangzhou: ADMA, 2013: 231-242.

[7] D. Zeng, K. Liu, S. Lai, et al. Relation classification via convolutional deep neural network[C]//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics. Dublin: COLING, 2014: 2335-2344.

[8] Y. Shen, X. J. Huang. Attention-based convolutional neural network for semantic relation extraction[C]//Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics. Osaka: COLING, 2016: 2526-2536.

[9] D. Zhang, D. Wang. Relation classification via recurrent neural network[J]. arXiv preprint arXiv:1508.01006, 2015.

[10] Y. Zhang, J. Yang. Chinese NER using lattice LSTM[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018:1554-1564.

[11] Z. Li, N. Ding, Z. Liu, et al. Chinese relation extraction with multi-grained information and external linguistic knowledge[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Firenze: ACL, 2019: 4377-4386.

[12] B. Kong, S. Liu, F. Wei, et al. Chinese Relation Extraction Using Extend Softword[J]. IEEE Access, 2021, 9: 110299-110308.

[13] J. Devlin, M. W. Chang, K. Lee, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis: NAACL, 2019: 4171–4186.

[14] M. Eberts, A. Ulges. Span-Based Joint Entity and Relation Extraction with Transformer Pre-Training[C]// European Conference on Artificial Intelligence. Santiago: ECAI, 2020.

[15] Z. Zhong, D. Chen. A Frustratingly Easy Approach for Entity and Relation Extraction[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics. Mexico City: NAACL, 2021: 50-61.

[16] Z. Lan, M. Chen, S. Goodman, et al. ALBERT: A lite bert for self-supervised learning of language representations[C]//In International Conference on Learning Representations. Addis Ababa: ICLR, 2020.

[17] W. Zhou, M. Chen. An improved baseline for sentence-level relation extraction[J]. arXiv preprint arXiv:2102.01373, 2021.

[18] Y. Liu, M. Ott, N. Goyal, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.

[19] Y. Cui, W. Che, T. Liu, et al. Revisiting Pre-Trained Models for Chinese Natural Language Processing[C]//Findings of the Association for Computational Linguistics. Punta Cana: EMNLP, 2020: 657-66.

[20] J. Yang, B. Ji, S. Li, et al. Dynamic Multi-View Fusion Mechanism For Chinese Relation Extraction[C]//Advances in Knowledge Discovery and Data Mining. PAKDD, 2023: 405–417.

[21] Q. Zhao, T. Gao, N. Guo. A novel chinese relation extraction method using polysemy rethinking mechanism[J]. Applied Intelligence, 2023, 53(7): 7665-7676.

[22] Z. Dong, Q. Dong. HowNet-a hybrid language and knowledge resource[C]//International conference on natural language processing and knowledge engineering. San Francisco: IEEE, 2003: 820-824.

[23] R. Speer, J. Chin, C. Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge.[C]. Proceedings of the 31st AAAI Conference on Artifcial Intelligence, 2017: 4444–4451.

[24] D. P. Kingma, J. Ba. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[25] J. Xu, J. Wen, X. Sun, et al. A discourse-level named entity recognition and relation extraction dataset for chinese literature text[J]. arXiv preprint arXiv:1711.07010, 2017.

[26] Z. R. Li, N. Ding, Z. Y. Liu, et al. Chinese relation extraction with multi-grained information and external linguistic knowledge[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. ACL, 2019: 4377-4386.

[27] D. Zeng, K. Liu, Y. Chen, et al. Distant supervision for relation extraction via piecewise convolutional neural networks[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: EMNLP, 2015: 1753-1762.

[28] Z. Zhong. Chinese Entity Relation Extraction Based on Multi-level Gated Recurrent Mechanism and Self-attention[C]//2021 2nd International Conference on Artificial Intelligence and Information Systems. Chongqing: ICAIIS, 2021: 1-7.

[29] X. Zeng X, J. Zhong, C. Wang, et al. Chinese relation extraction with flat-lattice encoding and pretrain-transfer strategy[C]//Knowledge Science, Engineering and Management: 14th International Conference. Tokyo: KSEM, 2021: 30-40.

[30] Y. B. Gao, G. H. Gong, and L. Ni. Chinese relation extraction in military field based on multi-grained lattice transformer and imbalanced data classification[J]International Journal of Modeling, Simulation, and Scientific Computing, 2024, 15(05):2350052.

Temperature Prediction for Photovoltaic Inverters Using Particle Swarm Optimization-Based Symbolic Regression: A Comparative Study

Fabian Alonso Lara-Vargas[1], Jesús Águila-León[2], Carlos Vargas-Salgado[3], Oscar J. Suarez[4]
Universidad Pontificia Bolivariana, Montería, Colombia[1]
University Institute of Energetic Engineering, Universitat Politècnica de València, Valencia, Spain[1,3]
Department of Water and Energy Studies, Universidad de Guadalajara, Guadalajara, Mexico[2]
Mechatronics Engineering Department, Universidad de Pamplona, Pamplona, Colombia[4]

***Abstract*—Accurate temperature modeling is crucial for maintaining the efficiency and reliability of solar inverters. This paper presents an innovative application of symbolic regression based on particle swarm optimization (PSO) for predicting the temperature of photovoltaic inverters, offering a novel approach that balances accuracy and computational efficiency. The study evaluates the performance of a PSO-based symbolic regression model compared to multiple linear regression (MLR) and a symbolic regression model based on genetic algorithms (GA). The models were developed using a dataset that included inverter temperature, active power, and DC bus voltage, collected over a year in hourly intervals from a rooftop photovoltaic system in a tropical region. The dataset was divided, with 70% used for training and the remaining 30% for testing. The symbolic regression model based on PSO demonstrated superior performance, achieving lower values of the root mean square error (RMSE) and mean absolute error (MAE) of 3.97 and 3.31, respectively. Furthermore, the PSO-based model effectively captured the nonlinear relationships between variables, outperforming the MLR model. It also exhibited greater computational efficiency, requiring fewer iterations than traditional symbolic regression approaches. These findings open new possibilities for real-time monitoring of photovoltaic inverters and suggest future research directions, such as generalizing the PSO model to different environmental conditions and inverter types.**

Keywords*—Particle swarm optimization; photovoltaic inverters; multiple linear regression; symbolic regression; temperature prediction

I. Introduction

Photovoltaic systems play a crucial role in reducing the carbon footprint compared to traditional energy sources. These systems provide a sustainable alternative that helps reduce future CO_2 emissions, which is critical to combating global warming [1]. Solar inverters are essential for photovoltaic systems because they convert direct current (DC) from solar panels into alternating current (AC) for commercial and residential use [2]. In the context of electrical systems, solar inverters are critical in facilitating grid integration and improving overall system efficiency [3]. In a global context where the photovoltaic capacity has increased 41% annually since 2009 [4], improving thermal management in solar inverters is critical to achieving energy sustainability goals. Accurate temperature modeling in solar inverters facilitates predicting and controlling thermal conditions, which is crucial for optimizing their performance and preventing failures [5].

Currently, the use of advanced thermal models combined with sensors and control algorithms is becoming increasingly important to ensure efficient thermal management in these devices [6]. Proper temperature control of a solar inverter is essential to maintaining the efficiency and longevity of these systems [7]. Inaccurate temperature modeling of solar inverters can significantly impact their performance and reliability, affecting these devices' thermal management and operational efficiency [8]. Additionally, erroneous temperature predictions can lead to suboptimal thermal management strategies, resulting in energy losses and reduced efficiency of solar inverters [9].

Previous studies have used various statistical and machine learning methods to model physical processes, energy production processes, and related parameters with varying degrees of success [10]–[13]. In [10], multiple machine-learning approaches have been applied, including multiple linear regression, decision trees, random forests, support vector machines, and neural networks. Otherwise, [11] presents a mathematical model of multigenic genetic programming (GP) designed to forecast the flow of the Blackwater River. To enhance its accuracy, this model has been optimized using the particle swarm optimization (PSO) algorithm. The model achieved an R^2 coefficient value of 0.96059 in the training set and 0.94296 in validation; however, it required 150 iterations in the adjustment parameters using PSO, which may result in increased computational expense. In addition, [12] proposes a method for predicting photovoltaic inverter temperatures using a hybrid neural network model that integrates convolutional neural networks (CNN) with long short-term memory (LSTM) networks. This CNN-LSTM approach significantly enhances temperature prediction accuracy, as demonstrated by an improvement in R^2 metrics from 0.92 to 0.96 compared to actual values. Finally, models based on neural networks (CNN-LSTM) and lumped thermal networks [13], have shown improvements in accuracy but face limitations in complexity and computational [14]. These limitations underscore the need for more efficient and robust methods for predictive solar inverter temperature modeling, aiming to enhance operational efficiency and extend the lifespan of PV systems. This study explores whether a PSO-based symbolic regression model can offer a more accurate and computationally efficient alternative to existing models.

Symbolic regression (SR) is a powerful tool for modeling

complex systems, offering improvements in precision [15] and robustness [16]. Traditional genetic programming (GP), a widely used SR method, leverages evolutionary principles to evolve mathematical expressions over multiple generations [17]. However, GP and similar traditional methods often face challenges related to efficiency and convergence. In contrast, algorithms based on swarm intelligence, such as particle swarm optimization (PSO) and firefly algorithms, have emerged as promising alternatives for symbolic regression tasks [18]. Inspired by the collective behavior of social organisms, these algorithms navigate large and complex search spaces, typical of symbolic regression problems [19]–[21]. Despite their potential, limited studies have explored the application of PSO-based symbolic regression for modeling solar inverter temperatures or compared its effectiveness with traditional approaches like Multiple Linear Regression (MLR) or Genetic Algorithm (GA)-based symbolic regression. Analyzing this approach could identify more precise, interpretable, and robust temperature prediction methods, thereby enhancing thermal management in solar inverters.

On the other hand, various methods have been developed to estimate temperature based on power loss calculations, input and output parameters, and thermal models [22], [23]. These methods utilize the relationship between electrical parameters and thermal behavior to provide real-time temperature estimations [24].

From the above discussion, the following research question arises: ¿what extent can a symbolic regression algorithm based on Particle Swarm Optimization (PSO) enhance the accuracy of temperature modeling in solar inverters compared to MLR and GA-based symbolic regression models? This study specifically aims to compare the predictive accuracy of PSO-based symbolic regression with MLR and GA-based models for solar inverter temperature prediction, using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) as evaluation metrics.

The novelty of the work is summarized as follows: (1) the creation of an innovative symbolic regression model using particle swarm optimization (PSO), which achieves higher accuracy than traditional models like multiple linear regression (MLR) and genetic algorithm (GA)-based approaches; (2) the model's enhanced capability to capture nonlinear relationships between critical variables, such as active power and DC voltage, essential for complex systems like solar inverters; and (3) its computational efficiency, demonstrated by fewer iterations and shorter execution times compared to advanced methods.

The paper outline is as follows: Section II presents the data collection methodology, and the construction of the PSO-based symbolic regression algorithm is described. Section III presents the temperature modeling of the solar inverter and the evaluation of the PSO-based symbolic regression model compared to the MLR and GA-based symbolic regression models. Section IV contains the discussion. Finally, conclusions are drawn in Section IV.

II. Methods and Materials

This section describes the methodology used for the analysis. The study requires data on the temperature of the solar inverter, its active power, and DC bus voltage, as well as information from the photovoltaic system. This data serves as input for developing the symbolic regression model based on PSO, multiple linear regression model, and GA-based symbolic regression model. The results are then compared with the measured temperature of the solar inverter. Fig. 1 illustrates the methodology for developing and evaluating these models.

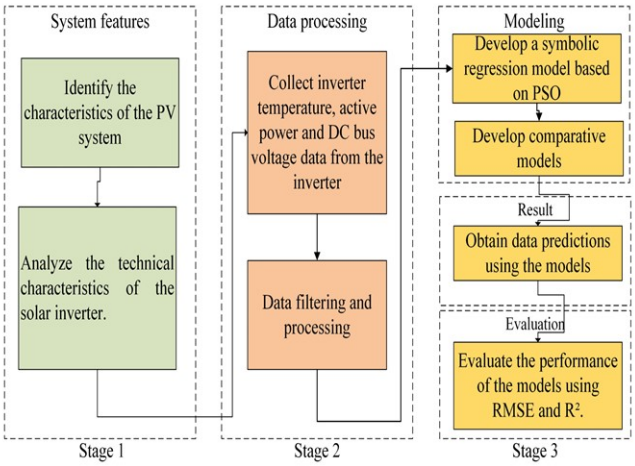


Fig. 1. Methodology followed in carrying out the models and the evaluations.

A. Characteristics of the PV System

The photovoltaic system is located on the rooftop of a building in Montería, Colombia, at coordinates 8°48'13.5" N, 75°51'0.45" W. It connects to the building's electrical grid through an indoor substation on the roof. The inverter is housed inside the rooftop structure, and Fig. 2 shows the current installation setup.

Fig. 2. Three-phase solar inverters and grid-connected electrical system.

B. Technical Characteristics of the Solar Inverter

The photovoltaic system comprises two solar inverters. The installation incorporates a transformer at the inverter output to

facilitate grid connection. The solar inverter under analysis is the Yaskawa PVI 36TL-480. Table I describes the technical characteristics of the solar inverter and the transformer.

TABLE I. TECHNICAL CHARACTERISTICS OF THE INSTALLATION

Solar inverter and transformer	
Item	**Detail**
Inverter Rated Power	36 kW
Inverter Power Input Voltage Range	540-800 VDC
Inverter Ambient Temperature Range	(-25 °C to +60 °C)
Transformer for coupling to the electrical network	80 kW 460 V/120 V

C. Data Acquisition

The data extracted from the solar inverter includes the inverter temperature in (ºC), active power in (kW), and DC bus voltage in (V). The data are stored hourly on a cloud-based platform. The volume of data encompasses one year of storage. The transmission system uses a GSM module to transmit the data to a cloud-based computer system from which it can be downloaded for analysis. Fig. 3 illustrates the process of data transmission and storage.

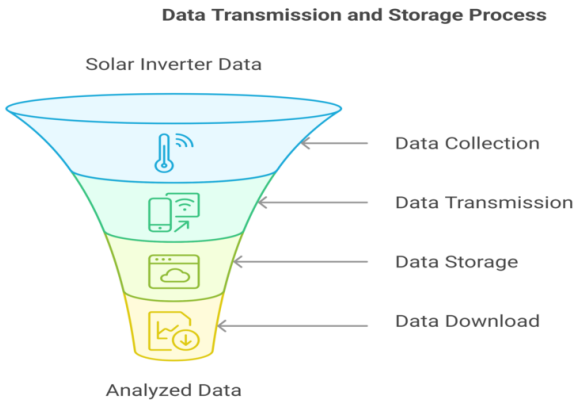


Fig. 3. Phases of the solar inverter's data transfer and storage process.

D. Data Filtering and Processing

The filtration and processing of the data followed these steps:

1) Identifying and eliminating outliers and inaccurate or incomplete data: Atypical data points substantially deviating from normal experimental parameters were classified as outliers. Specifically, any measurement that fell beyond two standard deviations from the adjusted median was designated as an outlier.

2) Manual review: Each data point was meticulously reviewed to identify and correct apparent errors or inconsistencies.

E. Algorithm Description

The PSO-based symbolic regression algorithm models nonlinear relationships by evolving algebraic expressions through particle swarm optimization. It integrates the following key steps:

1) Variable and function definition: Internal symbolic variables x1 and x2 are defined using Python's SymPy library. A list of mathematical functions (e.g. trigonometric, logarithmic) with random constants is constructed.

2) Expression generation: Random symbolic expressions are created by combining 3–6 terms from the predefined function list, operating on x1 and x2.

3) Fitness evaluation: Each expression is evaluated using RMSE between predicted and actual values. Invalid outputs (e.g., division by zero) are penalized with high error values.

4) PSO Update: Particle positions and velocities are updated based on personal best (pbest) and global best (gbest) solutions, guided by inertia (w=0.7), cognitive (c1=1.5), and social (c2=1.5) factors.

5) Optimization loop: The algorithm iteratively refines expressions over 25 iterations, balancing exploration and exploitation to minimize RMSE.

6) Data handling: Training (70%) and testing (30%) datasets are split using pandas. Results are visualized using matplotlib to track RMSE evolution.

F. Developing a Symbolic Regression Model Based on PSO

Symbolic regression is a tool that models complex nonlinear relationships between variables by identifying algebraic expressions that best fit the dynamics of a given system [25]. The PSO-based symbolic regression algorithm consists of the following subroutines in Fig. 4.

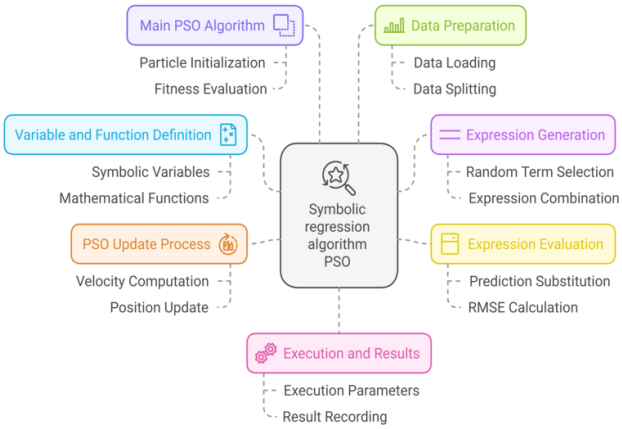


Fig. 4. Flowchart of symbolic regression algorithm based on particle swarm optimization (PSO).

The following section provides a detailed description of each subroutine the algorithm utilizes for its operation. The program used to develop the algorithm in Spyder 6 is an integrated development gateway (IDE) designed explicitly for scientific programming in Python [26].

Subroutines: Definition of Variables and Mathematical Functions
Input: No external inputs are required; the symbolic variables x1 and x2 are generated internally.
Output: The symbolic variables (x1, x2) are defined as mathematical symbols. The list of mathematical functions encompasses algebraic, trigonometric, and advanced operations (e.g. sin, log, sqrt), some with random constants.
Process: The symbolic variables x1 and x2 are defined utilizing the Sympy library in Python. A list of functions operating on x1 and x2 is constructed, incorporating random constants and predefined mathematical operations.

Subroutines: Generation of Mathematical Expressions
Input: Symbolic variables x1, x2. A predefined list of mathematical functions contains algebraic operations and mathematical functions applicable to x1 and x2.
Output: Symbolic mathematical expression combining several functions randomly selected from the list.
Process: A random number of terms is chosen (between 3 and 6). Functions are randomly selected from the list of mathematical functions. Each function operates on the symbolic variables x1 and x2. The resulting terms are summed to form a single symbolic expression.

Subroutines: Evaluation of Expressions
Input: A mathematical expression generated in a previous step. Arrays for x1, x2 (independent variables), and y (dependent variable).
Output: A numerical value representing the difference between the predicted values (from the symbolic expression) and the actual y values (RMSE).
Process: Each pair of values x1 and x2 are substituted into the symbolic expression to calculate the predicted value. If an invalid value (e.g. division by zero or NaN) occurs, assign a high error (inf) to the expression. Compares the predicted values with the actual values of y using the formula RMSE.

Subroutines: Update of Velocity and Position in PSO
Input: Current velocity and position representing the state of a particle. Personal best (pbest) and global best (gbest). PSO parameters (w, c1, c2).
Output: Updated velocity and position for the particle.
Process: Calculate the new velocity. Update the position.

Subroutines: Main PSO Algorithm
Input: Training and test data (x1, x2, y). PSO parameters: Number of particles, iterations, and constants (w, c1, c2).
Output: Best expression: The symbolic expression with the lowest RMSE. Best RMSE: The RMSE of the best expression. RMSE evolution: A record of RMSE values over iterations.
Process: Initialization: Generate initial particles (expressions) and velocities and evaluate their fitness. Optimization Loop: For each iteration: - Evaluate RMSE for each particle. - Update pbest and gbest based on fitness. - Adjust velocity and position for each particle. - Record the best RMSE for the iteration. Stop when the maximum iterations or target RMSE is reached.

Subroutines: Data Loading and Preparation
Input: Path to an Excel file containing data.
Output: Training and test datasets split into x_1, x_2, and y.
Process: Load data using `pandas.read_excel`. Extract columns for x_1, x_2, and y. Split data into 70% training and 30% test sets.

Subroutines: Results Visualization
Input: RMSE evolution data for each execution.
Output: A graph showing RMSE over iterations for all executions.
Process: Convert iterations to a time scale for the X-axis. Plot RMSE evolution for each execution using `matplotlib`. Label axes, add a title, and display the graph.

The equipment used for model development had the following technical specifications, as can be seen in Table II.

TABLE II. TECHNICAL CHARACTERISTICS OF THE COMPUTER EQUIPMENT

Laptop Computer and Software	
Item	**Detail**
CPU	8x1.9 GHz
RAM	32GB DDR4
RAM Speed	4267 MHz
Software	Spyder 6

G. Development of the MRL and GA-Based Symbolic Regression Model

Multiple linear regression (MLR) is a statistical technique investigating the relationship between a single dependent variable and several independent variables [27]. It aims to determine the equation of a line that minimizes the total squared deviations between the predicted and actual values of the dependent variable [27].

On the other hand, symbolic regression based on genetic algorithms (GA) is a computational technique that combines genetic algorithms with symbolic regression to discover the mathematical expressions that best fit a given dataset [17].This approach leverages the evolutionary principles of genetic algorithms to explore the space of potential mathematical models to identify the most accurate and interpretable representation of underlying data relationships [28].

The performance of the PSO-based symbolic regression model was compared with that of a multiple linear regression (MLR) model and a GA-based symbolic regression model using the same training data set at 70% and testing at the remaining 30%. The PSO-based and GA-based symbolic regression models employed an identical set of mathematical functions in their equation formulation processes.

The generated models were evaluated using 30% of the data set aside for testing. The main goal is to balance having enough training data to create robust models and sufficient test data to assess model performance [29].

H. Evaluation of Model Performance

Model performance was evaluated using RMSE and MAE metrics.

RMSE: This indicator evaluates the size of the discrepancies between the model's predicted values ($V_{\text{predicted}}$) and the actual values (V_{target}), as illustrated in Eq. (1) [30]. A lower RMSE indicates higher model accuracy.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(V_{\text{predicted},i} - V_{\text{target},i})^2} \qquad (1)$$

MAE: Mean Absolute Error (MAE) is a widely used method in model validation in various fields of study. MAE quantifies the average error size in a set of predictions without considering their directionality [31]. Eq. (2) describes the Mean Absolute Error (MAE) as follows:

$$MAE = \frac{1}{n}\sum_{k=1}^{n}|y_k - \hat{y}_k| \quad (2)$$

The number of observations is denoted by n, y_k represents the actual value, and $\hat{y}_k$ denotes the predicted value.

III. Results

This section presents the primary results of the study. The results are organized into two subsections: The first subsection describes the process of building the multiple linear regression model, the PSO-based symbolic regression model, and the GA-based symbolic regression model. The second subsection analyzes and contrasts the results of comparing these models with the measured data of the solar inverter temperature.

A. Model Construction

The distribution of training and test data is examined before modeling. Fig. 5 and Fig. 6 illustrate the data distribution.

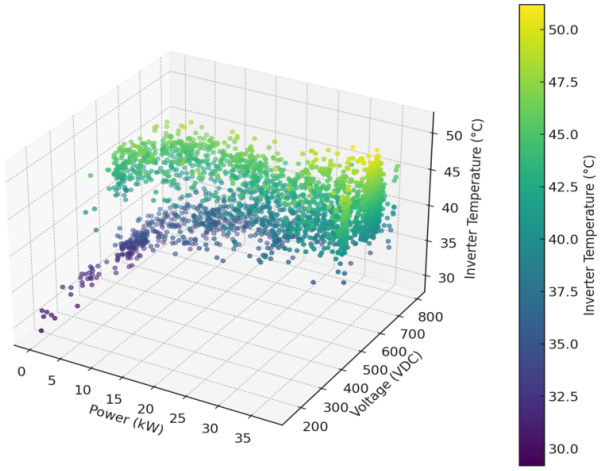


Fig. 5. Distribution of training data.

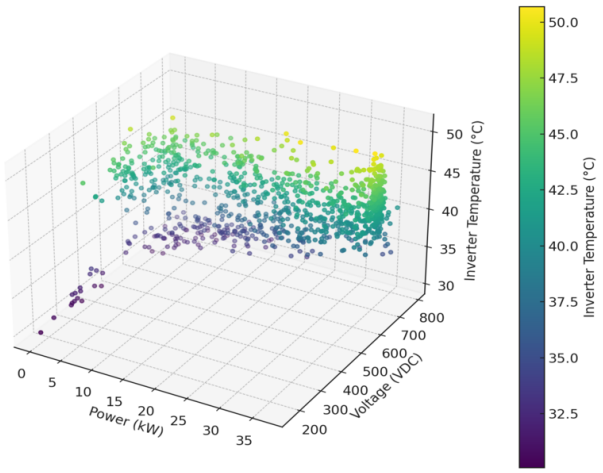


Fig. 6. Distribution of test data.

The dispersion and clustering of diverse data points across distinct regions of the three-dimensional space indicate complex relationships among the variables of active power, DC voltage, and inverter temperature. The following section describes the process of constructing the models:
A multiple linear regression model was established with training data and evaluated with the test dataset. The analysis revealed the corresponding coefficients in Equation (3) as follows:

$$MLR = 38.9104 + W \cdot 0.15981 + VD \cdot 0.0011, \quad (3)$$

where W represents the active power of the solar inverter in kW, and VD denotes the DC bus voltage in volts DC.

Using the training data, a symbolic regression model based on GA was established and evaluated with the test dataset. The parameters utilized for this model are illustrated in Table III.

TABLE III. Performance Parameters of the Algorithm Based on Symbolic Regression Based on GA

Input parameters	
Item	**Detail**
Population size	30
Number of generations	25
Initial mutation probability	0.5
Selection percentage	0.5
Crossover method	Subexpression Crossover
Number of algorithm executions	10

The optimal equation resulting from 10 executions of the algorithm was Eq. (4).

$$\text{Model}_{SR\_GA} = |VD|^{0.38} + \sqrt{VD} + 4.75, \quad (4)$$

where VD denotes the DC bus voltage in volts DC. Fig. 7 shows the result of the best RMSE value for each execution and the corresponding time.

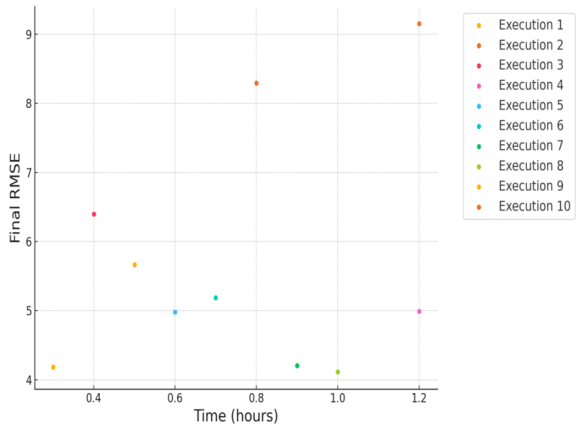


Fig. 7. Evolution of the RMSE of the Model RS GA over time.

Furthermore, selecting PSO parameters considers computational efficiency and model robustness; consequently, 30

particles and 10 executions are chosen for the algorithm. Concurrently, the inertia coefficient (0.7) and cognitive and social factors (1.5) are established to achieve an optimal balance between exploration and exploitation. As shown in Table IV, the selected inertia value allows particles to maintain sufficient momentum to escape local minima without affecting convergence.

TABLE IV. OPERATING PARAMETERS OF THE ALGORITHM

Item	Detail
Number of iterations	25
Particle numbers	30
Inertia factor	0.7
Cognitive factor	1.5
Social factor	1.5
Number of algorithm executions	10

Fig. 8 illustrates the evolution of the RMSE over the iterations, and Fig. 9 shows the RMSE graphed over time.

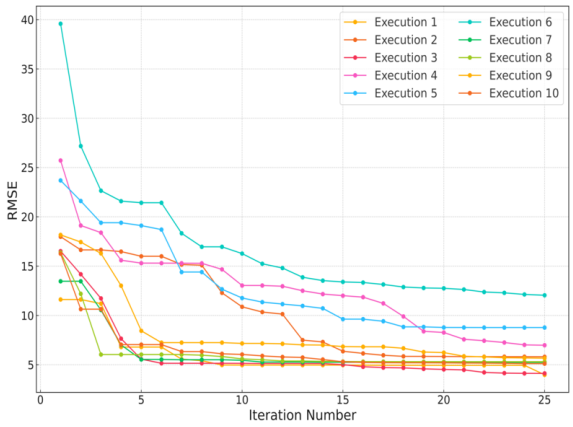


Fig. 8. Development of RMSE as a function of the number of iterations for different algorithm executions.

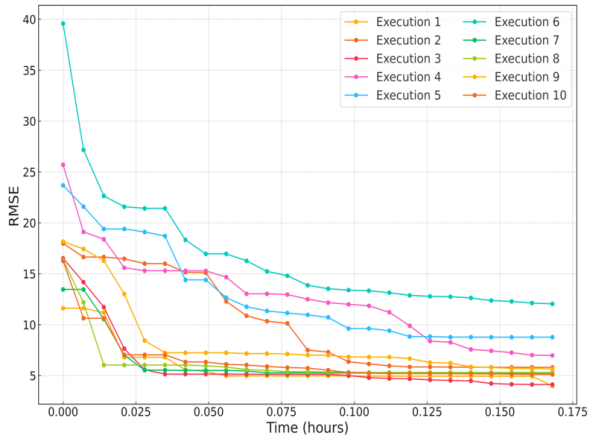


Fig. 9. Development of RMSE as a function of time for different algorithm executions.

The algorithm reduces the RMSE from initially high values (average values close to 20 in the first iterations) to final values between 3.97 and 12.04. This demonstrates the PSO's ability to identify symbolic equations that capture nonlinear relationships in the data. The progressive decrease in RMSE over the iterations confirms the algorithm's effectiveness in exploring and exploiting the search space, as confirmed in previous literature [11].

Regarding computing time, the algorithm completed 25 iterations within 0.175 hours per execution. The RMSE reduction reached a value of 3.97, positioning it as a competitive method compared to techniques such as neural networks or traditional multivariate regression for problems with nonlinear relationships [11]. The 25-iteration limit in the symbolic regression algorithm based on PSO allows for a brief and controlled execution time (0.175 hours), balancing computational efficiency and precision in searching solutions.

Additionally, using 30 particles balances exploration capacity and computational cost. This relatively modest size appears sufficient to adequately explore the search space and identify effective solutions to symbolic regression problems. Furthermore, a low inertia factor (0.5) favors local exploitation in advanced phases of the algorithm. This contributes to the rapid convergence observed in the RMSE.

A social factor of (1) means that the particles have a reduced tendency to follow the global best solution of the swarm, promoting a certain level of diversity in search. Combined with the high cognitive factor, this balance allows for more precise local convergence. The value of the cognitive factor (3) indicates that particles prioritize their own experiences. Eq. 5 illustrates the iterative process:

$$\begin{aligned}\text{Model RS PSO} = {} & 0.0940W - 0.0483VD \\ & + 0.2599\log(|VD| + 0.00001) \\ & - 0.4005\sin(W) + 0.4099\cos(VD) \\ & + 0.0049\tan(W) + 2.69206\sqrt{|VD|} \\ & + 0.52838\arctan(W) + 1.0941,\end{aligned} \tag{5}$$

where W represents the active power of the solar inverter in kW and VD denotes the DC bus voltage. The combination of nonlinear terms indicates that PSO can uncover relationships that traditional approaches, such as multiple linear regression, cannot capture.

Table V shows the RMSE and MAE results for the training dataset of multiple linear regression, symbolic regression model based on PSO, and GA-based symbolic regression model.

TABLE V. COMPARISON OF RESULTS WITH TRAINING DATA

Models and Metrics		
Item	**RMSE**	**MAE**
Model Multiple Linear Regression (MLR)	4.22	3.55
Symbolic Regression Model based on PSO (SR PSO)	3.97	3.36
Symbolic Regression Model based on GA (SR GA)	4.59	3.78

Furthermore, Table VI presents the corresponding results for the test dataset.

TABLE VI. COMPARISON OF RESULTS WITH TEST DATA

Models and Metrics		
Item	**RMSE**	**MAE**
Model Multiple Linear Regression (MLR)	4.52	3.73
Symbolic Regression Model based on PSO (SR PSO)	4.12	3.31
Symbolic Regression Model based on GA (SR GA)	4.80	3.66

The RMSE and MAE metrics indicate that the symbolic regression model based on PSO demonstrated superior training and test data performance, achieving RMSE values of 3.97 and MAE of 3.31. Fig. 10, Fig. 11, and Fig. 12 illustrate the behavior of the Model MRL, Model SR PSO, and Model SR GA with training data.

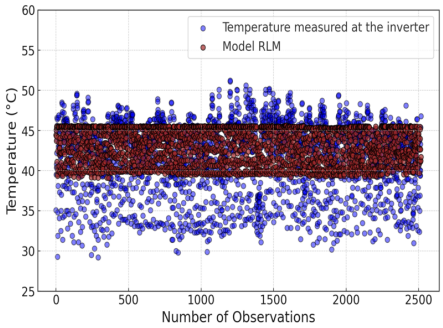


Fig. 10. Comparison of the temperature measured and the Model MLR with training data.

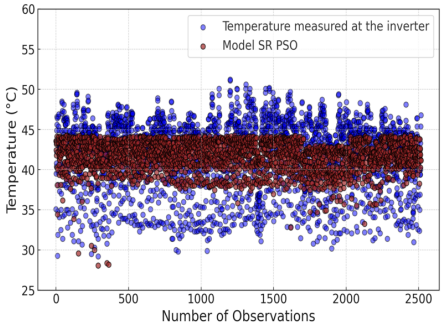


Fig. 11. Comparison of the temperature measured and the Model RS PSO with training data.

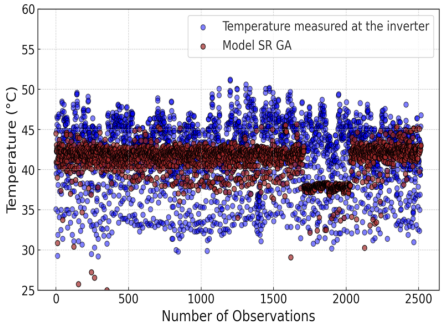


Fig. 12. Comparison of the temperature measured and the Model RS GA with training data.

Fig. 13, Fig. 14, and Fig. 15 show the behavior of Model MRL, Model SR PSO, and Model SR GA with training data. The model MLR represents a linear approximation with more substantial errors in scenarios where the relationships between variables and inverter temperature are not strictly linear, resulting in more significant deviations between predicted and actual values. The model SR PSO typically aligns more closely with actual inverter temperatures and exhibits reduced bias across multiple observations. Additionally, the Model SR GA model demonstrates inferior performance compared to the MLR and RS PSO models in evaluation metrics.

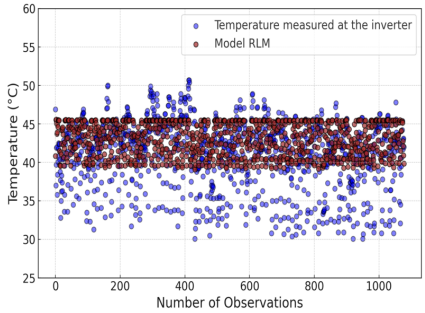


Fig. 13. Comparison of temperature measured and the MLR model with test data.

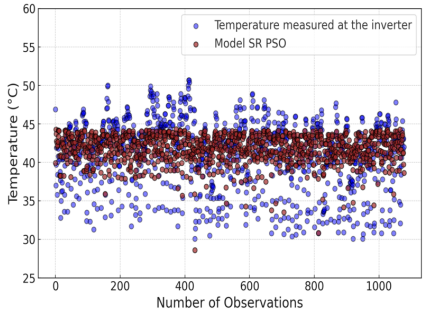


Fig. 14. Comparison of temperature measured and the Model RS PSO with test data.

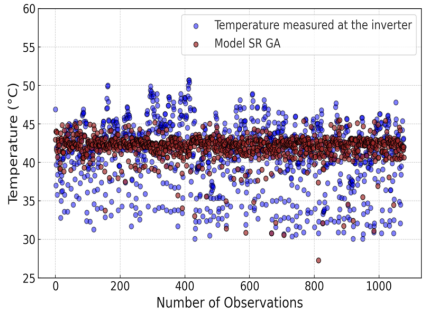


Fig. 15. Comparison of temperature measured and the Model RS PSO with test data.

Fig. 16, Fig. 17, and Fig. 18 show the performance of the MLR, RS PSO, and SR GA models based on the training data,

while Fig. 19, Fig. 20, and Fig. 21 illustrate the behavior with test data.

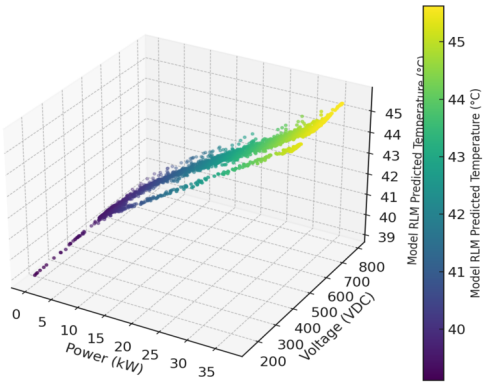


Fig. 16. Distribution of the data predicted by the model MLR using the training data.

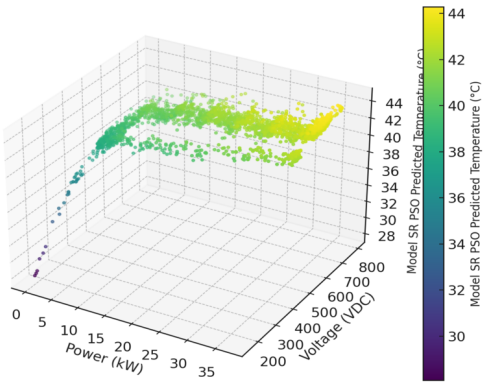


Fig. 17. Distribution of data predicted by the model RS PSO using the training data.

While multiple linear regression (MLR) presents a simplified linear trend with a limited prediction range (39-45 °C) that does not capture the full complexity of the actual data (30-50 °C), the symbolic regression model with PSO demonstrates superior capability in reproducing the non-linear nature of interactions among inverter power, voltage, and temperature, as evidenced by its improved fit to variable gradients and more realistic prediction range (28-44°C).

Both the symbolic regression model response based on PSO and the one based on GA exhibited similar RMSE values; however, the regression developed with PSO achieved the lowest value after 10 executions in both algorithms.

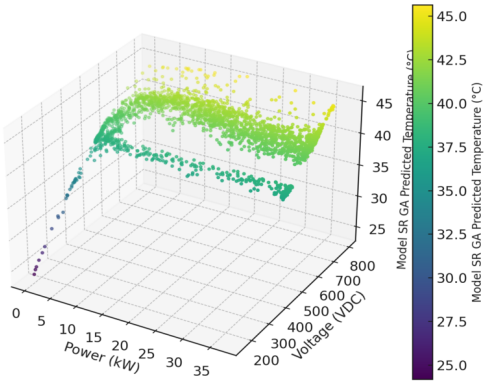


Fig. 18. Distribution of data predicted by the model RS GA using the training data.

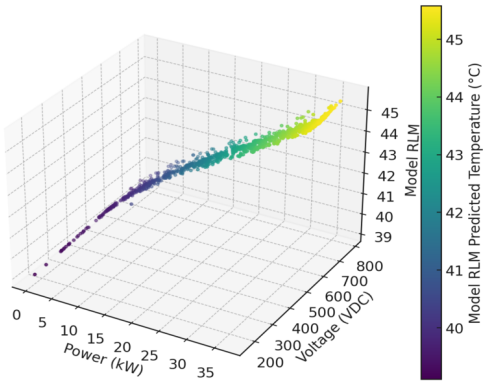


Fig. 19. Distribution of data predicted by the model MLR using test data.

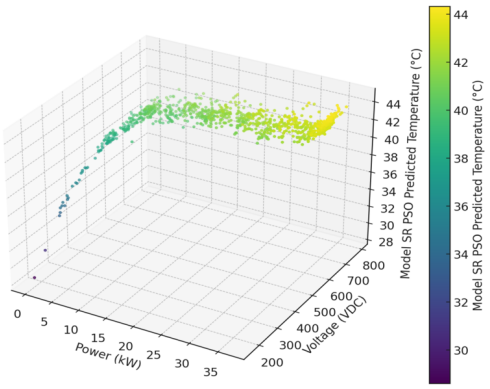


Fig. 20. Distribution of data predicted by the model RS PSO using the test data.

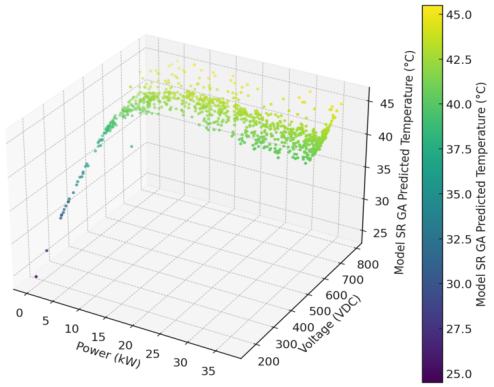


Fig. 21. Distribution of data predicted by the model RS GA using the test data.

IV. Discussion

The rapid convergence of the PSO-based model and its reduced computational cost compared to the GA-based approach render it an attractive option for symbolic regression tasks. Nevertheless, PSO may experience premature convergence and stagnation in complex search spaces, corroborating recent findings [32].

The response surface generated by PSO more accurately reflects the local variations and curvature observed in the experimental data. This suggests that this method is more appropriate for modeling complex physical systems where relationships between variables are inherently non-linear, as with temperature in solar inverters.

The results validate the efficacy of PSO in symbolic regression; however, its tendency to converge prematurely may result in suboptimal solutions if not managed appropriately.

V. Conclusion

The study shows that a symbolic regression model based on particle swarm optimization (PSO) outperforms multiple linear regression (MLR) and a Symbolic regression model based on GA in predicting the internal temperature of solar inverters, as evidenced by lower RMSE and MAE values. The SR PSO model's ability to capture nonlinear relationships between active power (W), DC bus voltage (VD), and temperature represents a significant advantage given the complex nature of solar inverter systems. However, the RS PSO model may require more computational resources.

According to the analyzed literature, this study investigated the application of the Particle Swarm Optimization (PSO) algorithm in symbolic regression, a domain predominantly utilizing Genetic Programming (GP) and neural networks. Although the surface modeling of the training and test data was not very close, the model highlighted the computational efficiency and potential adaptability of PSO to symbolic regression tasks.

Future research could examine the generalizability of the SR PSO model across different solar inverter types and environmental conditions, incorporate additional variables such as ambient temperature and humidity, and explore possible improvements through parameter optimization. In addition, comparing the SR PSO model with other advanced machine learning models and integrating it into real-time monitoring systems could further improve its practical application in photovoltaic systems.

Acknowledgment

This work is supported by Universidad Pontificia Bolivariana, Seccional Monteria. The authors also wish to express their gratitude for their collaboration in developing this research at Universitat Politècnica de València, Universidad de Guadalajara, and Universidad de Pamplona.

References

[1] M. Nasr, M. Ibrahim, and A. El Berry, "Solar Photovoltaic Panels Fault Detection Due to Thermal Effects," *Journal of International Society for Science and Engineering*, vol. 6, no. 2, pp. 26–35, 2024. [Online]. Available: www.isse.org.eg

[2] R. Madake and S. Dhanaraj, "Photovoltaic inverters experimentally validate power quality mitigation in electrical systems," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 36, no. 2, pp. 715–723, nov. 2024. doi: 10.11591/ijeecs.v36.i2.pp715-723.

[3] T. Liang et al., "Implementation and Applications of Grid-Forming Inverter with SiC for Power Grid Conditioning," *IEEJ Journal of Industry Applications*, 2023. doi: 10.1541/ieejjia.22006702.

[4] L. Kruitwagen, K. Story, J. Friedrich, L. Byers, S. Skillman, and C. Hepburn, "A global inventory of photovoltaic solar energy generating units," *Nature*, vol. 598, no. 7882, 2021. doi: 10.1038/s41586-021-03957-7.

[5] T. Ivanov and R. Stanev, "Mathematical model of photovoltaic inverters," in *11th Electrical Engineering Faculty Conference, BulEF 2019*, 2019. doi: 10.1109/BulEF48056.2019.9030705.

[6] Z. Wang, "A Thermal Management Strategy for Inverter System Based on Predictive Control," *International Journal of Heat and Technology*, vol. 41, no. 1, 2023. doi: 10.18280/ijht.410115.

[7] D. Mathew and R. Naidu, "A review on single-phase boost inverter technology for low power grid integrated solar PV applications," *Ain Shams Engineering Journal*, vol. 15, no. 2, 2024. doi: 10.1016/j.asej.2023.102365.

[8] A. Ribeiro et al., "Junction Temperature Prediction and Lifetime Assessment in a PV Inverter Using a 10-year Mission profile," *Eletrônica de Potência*, vol. 29, pp. e202438, oct. 2024. doi: 10.18618/REP.e202438.

[9] A. Luiz Perin, A. Krenzinger, and C. Massen Prieb, "Modelagem térmica de um inversor fotovoltaico conectado à rede," in *Anais Congresso Brasileiro de Energia Solar - CBENS*, 2016. doi: 10.59627/cbens.2016.1978.

[10] Z. Zheng, "Introduction to Solar Technology and its Environmental Importance," *Applied and Computational Engineering*, vol. 98, no. 1, pp. 47–56, nov. 2024. doi: 10.54254/2755-2721/98/20241116.

[11] A. Sheta, A. Abdel-Raouf, K. Fraihat, and A. Baareh, "Evolutionary Design of a PSO-Tuned Multigene Symbolic Regression Genetic Programming Model for River Flow Forecasting," 2023. [Online]. Available: www.ijacsa.thesai.org

[12] H. Dong and H. Liu, "Research on Photovoltaic Inverter Temperature Prediction Method Based on CNN-LSTM," in *9th International Conference on Intelligent Computing and Signal Processing, ICSP 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 1647–1651. doi: 10.1109/ICSP62122.2024.10743359.

[13] A. Sangwongwanich, H. Wang, and F. Blaabjerg, "Reduced-Order Thermal Modeling for Photovoltaic Inverters Considering Mission Profile Dynamics," *IEEE Open Journal of Power Electronics*, vol. 1, 2020. doi: 10.1109/OJPEL.2020.3025632.

[14] M. Karim et al., "Explainable AI for Bioinformatics: Methods, Tools and Applications," *Briefings in Bioinformatics*, vol. 24, no. 5, 2023. doi: 10.1093/bib/bbad236.

[15] B. Cohen, B. Beykal, and G. Bollas, "Dynamic System Identification from Scarce and Noisy Data Using Symbolic Regression," in *Proceedings of the IEEE Conference on Decision and Control*, 2023. doi: 10.1109/CDC49753.2023.10383906.

[16] Z. Bastiani, R. Kirby, J. Hochhalter, and S. Zhe, "Complexity-Aware Deep Symbolic Regression with Robust Risk-Seeking Policy Gradients," 2024. [Online]. Available: http://arxiv.org/abs/2406.06751.

[17] N. Jiang and Y. Xue, "Racing Control Variable Genetic Programming for Symbolic Regression," 2023. [Online]. Available: http://arxiv.org/abs/2309.07934.

[18] M. Aliwi, S. Aslan, and S. Demirci, "Firefly Programming for Symbolic Regression Problems," in *28th Signal Processing and Communications Applications Conference*, 2020. doi: 10.1109/SIU49456.2020.9302201.

[19] V. Khaidurov, "Methods and Algorithms of Swarm Intelligence for the Problems of Nonlinear Regression Analysis and Optimization of Complex Processes, Objects, and Systems: Review and Modification of Methods and Algorithms," *System Research in Energy*, 2024. doi: 10.15407/srenergy2024.03.046.

[20] G. Kronberger, F. de Franca, H. Desmond, D. Bartlett, and L. Kammerer, "The Inefficiency of Genetic Programming for Symbolic Regression – Extended Version," *GECCO '18: Proceedings of the Genetic and Evolutionary Computation Conference*, 2024. [Online]. Available: http://arxiv.org/abs/2404.17292.

[21] P. Orzechowski, W. La Cava, and J. Moore, "Where are we now? A large benchmark study of recent symbolic regression methods," in *Proceedings of the 2018 Genetic and Evolutionary Computation Conference*, 2018. doi: 10.1145/3205455.3205539.

[22] M. Taha, M. Tousizadeh, A. Deb, O. Alatise, L. Ran, and P. Mawby, "Real Time Estimation of Power Transistor Junction Temperature for Motor Drive Application," in *IET Conference Proceedings*, 2022. doi: 10.1049/icp.2022.1153.

[23] G. Xu, H. Wu, Z. Pan, W. Wei, and Y. Zhang, "A Novel Online Method for Monitoring the Junction Temperature Based on the Input and Output of the Inverter," in *8th International Conference on Electrical and Electronics Engineering, ICEEE 2021*, 2021. doi: 10.1109/ICEEE52452.2021.9415952.

[24] Q. Niu and L. Wang, "Real-time Inverter Semiconductor Die Temperature Estimation Using Inverter Operating Condition-based Gate Recurrent Unit," in *9th IEEE International Symposium on Power Electronics for Distributed Generation Systems, PEDG 2018*, 2018. doi: 10.1109/PEDG.2018.8447642.

[25] D. Angelis, F. Sofos, and T. Karakasidis, "Artificial Intelligence in Physical Sciences: Symbolic Regression Trends and Perspectives," 2023. doi: 10.1007/s11831-023-09922-z.

[26] A. Ziogas et al., "Productivity, portability, performance: data-centric Python," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, in SC '21. New York, NY, USA: Association for Computing Machinery, 2021. doi: 10.1145/3458817.3476176.

[27] J. Jiang, "Multiple Linear Regression," in *Applied Medical Statistics*, John Wiley & Sons, Ltd, 2022, ch. 15, pp. 345–367. doi: https://doi.org/10.1002/9781119716822.ch15.

[28] Y. Radwan, G. Kronberger, and S. Winkler, "A Comparison of Recent Algorithms for Symbolic Regression to Genetic Programming," 2024. [Online]. Available: http://arxiv.org/abs/2406.03585.

[29] B. Supriet al., "Asian Stock Index Price Prediction Analysis Using Comparison of Split Data Training and Data Testing," *Jurnal Ekonomi, Manajemen, dan Akuntansi*, vol. 9, no. 4, 2023. doi: 10.35870/jemsi.v9i4.1339.

[30] T. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," 2022. doi: 10.5194/gmd-15-5481-2022.

[31] Z. Dayev, A. Kairakbaev, K. Yetilmezsoy, M. Bahramian, P. Sihag, and E. Kıyan, "Approximation of the discharge coefficient of differential pressure flowmeters using different soft computing strategies," *Flow Measurement and Instrumentation*, vol. 79, 2021. doi: 10.1016/j.flowmeasinst.2021.101913.

[32] M. Montes-Rivera, C. Guerrero-Mendez, D. Lopez-Betancur, and T. Saucedo-Anaya, "Dynamical Sphere Regrouping Particle Swarm Optimization Programming: An Automatic Programming Algorithm Avoiding Premature Convergence," *Mathematics*, vol. 12, no. 19, pp. 3021, 2024. doi: 10.3390/math12193021.

Towards Effective Anomaly Detection: Machine Learning Solutions in Cloud Computing

Hussain Almajed, Abdulrahman Alsaqer, Abdullah Albuali
Department of Computer Networks and Communications
College of Computer Sciences and Information Technology
King Faisal University
Al-Ahsa, 31982 Saudi Arabia

***Abstract*—Cloud computing has transformed modern Information Technology (IT) infrastructures with its scalability and cost-effectiveness but introduces significant security risks. Moreover, existing anomaly detection techniques are not well equipped to deal with the complexities of dynamic cloud environments. This systematic literature review shows the advancements in Machine Learning (ML) solutions for anomaly detection in cloud computing. The study categorizes ML approaches, examines the datasets and evaluation metrics utilized, and discusses their effectiveness and limitations. We analyze supervised, unsupervised, and hybrid ML models showing their advantages in dealing with a certain threat vector. It also discusses how advanced feature engineering, ensemble learning and real-time adaptability can improve detection accuracy and reduce false positives. Some key challenges, such as dataset diversity and computational efficiency, are highlighted, along with future research directions to improve ML based anomaly detection for robust and adaptive cloud security. Hybrid approaches are found to increase the accuracy reaching up to 99.85% and reduces the number of false positives. This review provides a comprehensive guide to researchers aiming to enhance anomaly detection in cloud environments.**

***Keywords*—Anomaly; cloud; machine learning; detection**

I. Introduction

An important part of modern IT infrastructure today is cloud computing, which offers flexible, scalable and cost effective solutions for businesses and individuals over the internet [1],[2]. Offers an on demand access to computing resources such as servers, storage, databases, and applications, which can be rapidly provisioned and released with minimal management effort. Cloud computing has become widely adopted in different industries from healthcare to finance, entertainment to education due to its benefits. Despite these advantages, the use of cloud services has introduced security challenges that must be addressed to guarantee the reliability and trustworthiness of cloud systems [3].

Back in 2022, a leading health insurer in Australia called Medibank stored sensitive information about 10 million customer accounts in its cloud based systems, and the unlucky company suffered a massive data breach which revealed all their customer's data [4], [5]. While the company refused to pay a ransom, hackers then started to leak the stolen data on dark web forums. It highlights the critical requirement for better anomaly detection in cloud environments. In addition, cyberattacks on cloud computing environments are becoming more prevalent and cause significant data breaches. The number of breaches within the cloud environment also increased, from 35% of businesses in 2022 to 39% at 2023 [6].

Cloud environments are dynamic and elastic, which makes them vulnerable to attacks from malicious actors, therefore effective anomaly detection is crucial to keep cloud environment secure. The cloud infrastructure is by nature shared where many tenants may use the same physical resources, making the risk of potential security breaches even higher [7]. Furthermore, managing various virtualized environments, and the constant scaling of resources, makes it difficult to establish a stable security baseline. In these environments security threats can vary from external threat, like Distributed Denial of Service (DDoS) attacks to internal threat, losing control over an insider, unauthorized access, data breaching and configuration errors.

Anomaly detection is identifying unusual patterns or behaviour which may indicate security breach, failure of system or performance issues [8] [9]. An effective anomaly detection method is necessary to reduce the impact of these threats by providing timely and proactive responses. Traditional rule based detection methods based on predefined signatures or rules struggle to keep pace with the complexity and evolution of cloud environments. These methods are unable to detect previously unseen or novel threats, particularly with the diversity and scale of cloud services. Attack patterns evolve rapidly and the cloud environment is always changing, for that we need more adaptive and more intelligent approaches [10].

ML steps in when it provides advanced algorithmic tools to process huge amounts of data and react to new threats by identifying anomalies [11] [12]. Anomaly detection with ML has the capability to offer more secure cloud systems by means of automated, intelligent monitoring of cloud systems. Because ML techniques learn from data, find complex relationships and get better over time, they are particularly well suited to the cloud [13]. With these characteristics, ML is a promising approach to identifying security anomalies that would likely be missed by traditional approaches.

Therefore, cloud computing has revolutionized how the organizations manage and access to the IT resources by providing many advantages and at the same time introducing various security issues. Anomaly detection helps to identify abnormal activities in the cloud, which can indicate abnormal threats. Anomaly detection is a powerful capability of ML which improves the security posture of cloud environments with

adaptive, data driven techniques. In this systematic literature review, we will explore in detail the current state of the art of ML based anomaly detection in cloud computing, the challenges faced and future research directions to address the evolving threat landscape.

Following this introduction, Section II provides background information on cloud computing, anomaly detection, and a definition of ML. Section V describes the research methodology used in this study. Section VI reviews related work in the field. After that, Section VII will illustrate a case study regarding the research study. In Section VIII, the results of the review will be presented and discussed. Section IX then highlights key challenges and outlines open directions for future research. Finally, Section X synthesizes the key findings and concludes the study.

II. Background

A. Cloud Computing Overview

Cloud computing is an on demand network access to a shared pool of configurable computing resources such as a network, applications, servers, storage, or services, in which the providers deliver the resources on demand because they are-scalable, elastic and vary as per your need [14], [15]. The next section represents the cloud service models of the cloud and the cloud deployment models.

B. Cloud Service Model

There are three main service models for cloud computing which offer a varying level of control, flexibility and management. Fig. 1 shows the different cloud service models with examples.

- Infrastructure as a Service (IaaS): The model virtualizes these resources on demand. Users can deploy applications and the configuration settings, without managing or controlling the underlying cloud infrastructure.
- Platform as a Service (PaaS): This model provides a development setting where developers create and deploy applications without having to know how many processors.
- Software as a Service (SaaS): This model is used to deliver the applications over the web where they are consumed by the consumers through web portals.

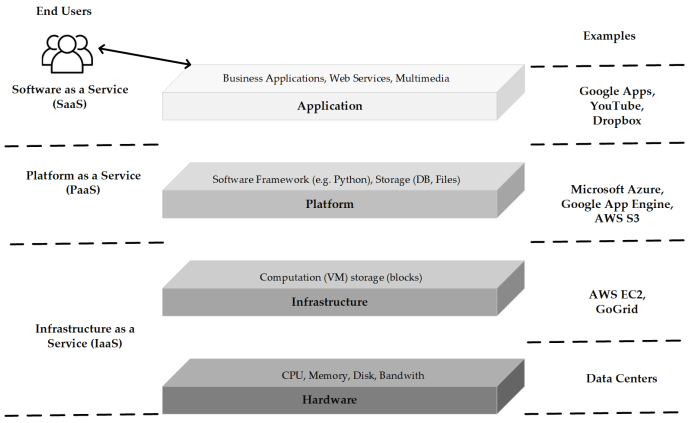


Fig. 1. Cloud service models with examples.

C. Cloud Deployment

The cloud deployment models as shown in Fig. 2 define the way by which cloud resources are managed and offered to the users. There are four main types:

- Private Cloud: The cloud infrastructure is delivered solely for use by one organization (a business unit) that consists of multiple consumers [15].
- Public Cloud: The cloud infrastructure is delivered for general public open use. It may be managed, owned, and operated by a corporation, academic or government organization, or a mix.
- Community Cloud: It is a cloud computing environment where multiple organizations with similar goals and security requirements share a cloud.
- Hybrid Cloud: Two or more distinct cloud infrastructures (private, community, or public) that have been independently operated and connected using standard or proprietary technology.

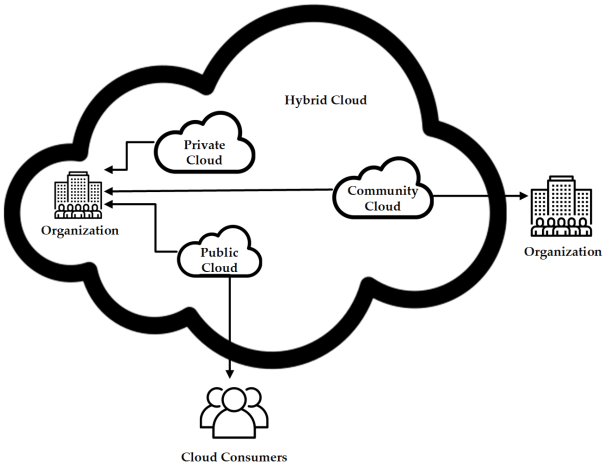


Fig. 2. Cloud deployment models.

D. Cloud Threat

It any event, situation or action which could lead to compromise of the confidentiality, integrity, or availability of cloud computing resources, data or services, including access, data breaches, service disruptions or other security violations, whether deliberate or accidental, that affect the security and trustworthiness of cloud computing environments [16]. The cloud threat can be categorized based Confidentiality,Integrity, and Availability (CIA) tried in the next section.

1) Classification of threat-based CIA:

- Confidentiality: There are many confidentiality threats in cloud computing environments. When unauthorized access of sensitive data stored in the cloud leads to data breaches and privacy violations, it is called Data Breaches [17], [18]. Moreover, Shared Technology Vulnerabilities e.g. hypervisor vulnerabilities and cross Virtual Machine (VM) side channel attacks are at risk as a result of shared infrastructure and multi tenancy [19].
- Integrity: Data tampering is one of the integrity threats in cloud computing, which refers to the unauthorized modifications of data stored in cloud which adversely affects the accuracy and reliability of data. Data integrity can also be compromised by Malicious Insiders, system administrators, and former employee. Application Programming Interface (APIs) with such vulnerabilities are insecure, which can easily be exploited in data manipulation.
- Availability: Denial of Service (DoS) Attacks threaten the availability of cloud computing by overwhelming cloud services with traffic making the cloud services unavailable to the legitimate users.

2) Attacks in cloud: The cloud can be attacked by diverse attacks and for better understanding security threats and vulnerabilities in cloud computing can be broadly classified into five main categories application-based,storage-based, VM-based network-based, Identity and Access Management. Fig. 3 shows list of some attacks based in the five categories.

1) Network-based attacks: Attacks related to network communications and configurations in cloud computing environments [20], [21]. Examples such as flooding attacks, Structured Query Language (SQL) injection attack, spoofing.
2) VM-Based Attacks: Attacks related to virtualization technology and hypervisor vulnerabilities. Examples such as Hypervisor Vulnerabilities, VM Escape, and VM Image Sprawl.
3) Storage-based Attacks: Attacks related to data confidentiality, integrity, and availability in cloud storage. Examples include Data Breaches, Data Isolation, and Data Backup and Redundancy.
4) Application-Based Attacks: The attacks are aimed at cloud infrastructure running applications. Examples include web services, malware infusion, and shared design vulnerabilities.
5) Identity and Access Management based attacks: Threats related to managing identities and providing secure and efficient access to data [18]. Examples include Identity Management, authorization, authentication, access control, and federation management.

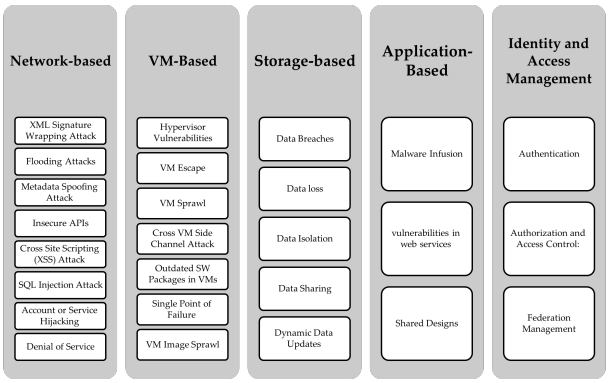


Fig. 3. List of cloud attacks based in category.

E. Anomaly Detection Definition

Finding data points, patterns or behaviors that are very different from the norm is known as anomaly detection [22]. When considering computing and cybersecurity, anomalies are indicative of possible issues, including security breaches, system failures and fraudulent activities or unusual behavior. Anomalies are outliers basically, data that doesn't conform to expected patterns or historical trends. Anomaly detection is used to detect these deviations early, so that organizations can take timely corrective action. Anomaly detection is important for reliability, security, and minimizing operational risk in modern systems and in the face of the variety of data and system complexity. Anomaly detection can be applied to a wide range of domains such as finance, healthcare, industrial monitoring and cloud computing in which real time and accurate anomaly detection can prevent a major loss or damage [13].

1) Anomaly Detection Techniques: The techniques can be categorized into three main types: statistical methods, ML techniques, and hybrid techniques.

- Statistical Methods: These techniques use statistical models to specify what normal behavior of a system is. Statistical methods build a baseline distribution from historical data, and flag any data point outside this distribution as an anomaly. Three common statistical techniques are z-score analysis, hypothesis testing and time series modeling [23]. These are easy to implement statistical methods, but can have difficulty with high dimensional data and in capturing complex patterns in dynamic environments.
- ML Approaches: Anomaly detection has gained popularity with ML due to its ability to learn directly from data while adjusting to evolving patterns. A variety of models are used in ML approaches to understand normal behavior and identify deviations which could indicate anomalies. These models are very useful in dynamic environments such as cloud environments where traditional models can fail to capture evolving patterns.
- Hybrid Techniques: Statistical, ML, or other domain specific anomaly detection methods could be combined to improve accuracy, robustness. Hybrid methods combine the strengths of different methods, filling the shortcomings of single techniques like increasing detection accuracy or reducing false positives. A hybrid approach may be a statistical model to identify potential anomalies, and apply a ML algorithm to validate it further. In complex environments like the cloud, these methods are effective where adaptive

learning and baseline modeling are both necessary to cope with dynamic changes.

Choice of anomaly detection technique is conditioned by data nature, availability of labeled data set, system complexity and desired detection accuracy and computational efficiency trade off. Each technique has pros and cons, and in most practical cases, a set of techniques are combined to obtain the best performance in real world settings.

F. Machine Learning

ML is a field of artificial intelligence concerned with enabling systems to learn and act based on data. The basic idea falls under training models to come up with patterns and predict without being explicitly programmed for the task that is required [24]. Due to its capacity to learn and get better with time, ML is now a must in many domains. ML is one of the important roles in cloud computing to improve security by automated detection of unusual or potentially malicious behaviors. ML models operating off large datasets can offer advanced and intelligent solutions to complicated issues such as anomaly detection.

G. Type of ML

Depending on their way of learning and the types of tasks, ML can be divided into various types.

1) Supervised Learning: Supervised learning is a type of ML where a model is trained using a labeled dataset, where input data is given along with corresponding outputs or labels [25]. By identifying patterns in the data, the model learns to map inputs to outputs. Supervised learning is very powerful in the case of anomaly detection in cloud computing, if there are a lot of labeled normal and anomalous behaviors [26]. For example, ML classification techniques commonly applied include classifications like Decision Trees (DT), Support Vector Machines (SVM), and Random Forests (RF), etc to classify activities as normal or anomalies. The major drawback of supervised learning, is that it is very hard to collect a large number of labeled data that are representative of rare events such as security breaches or insider threats [27]. Table I shows some of the supervised models.

2) Unsupervised Learning: It is a ML approach that does not demand labeled datasets. It does not look for normal behavior, nor does it look for anomalous behavior, instead it looks for patterns or groupings in the data without prior knowledge of what is normal and what is anomalous [25]. These techniques work by identifying deviation from known patterns, and are therefore particularly well suited to detecting new and previously unseen threats. One of the problems with unsupervised learning is that it is hard to tell the difference between benign deviations and actual security events without labeled data, and as a result, can have very high false positive rates. Table II shows some of the unsupervised models.

3) Reinforcement Learning (RL): It is learning what to do and how to map cases to actions to maximize a numerical reward signal. Unlike supervised and unsupervised learning, rather it is trial and error based sequential decision making with the agent's goal being to maximize cumulative rewards by taking the best possible actions [25]. RL can be used for developing adaptive security systems in cloud computing for anomaly detection, which models optimal reaction to threats as time elapsed. One of the reasons that RL is particularly appealing for cloud security is that it allows for real time adaptation to new and evolving threats [28]. RL has the potential to help improve the robustness and adaptability of cloud based anomaly detection systems.

TABLE I. LIST OF SUPERVISED ML MODELS SHOWING THEIR ADVANTAGES AND DISADVANTAGES

Model	Definition	Advantages	Disadvantages
RF [17]	It is algorithm used for classification and regression, where the data gets split into subsets and we train several DTs.	Automatic processing of missing values and no need to transform variables, good performance with many variables and large data, and high accuracy	Slow learning, large memory footprint, and difficult interpretation.
SVM [17]	A binary classification method that creates a hyperplane to divide two target values.	High accuracy, works well with high-dimensional data, and memory efficient	Poor performance with noisy data, and long training time.
Naïve Bayes (NB) [17]	Collection of classification algorithms based on Bayes theorem.	Easy to understand and configure, Fast and small memory footprint, and can learn from small data.	Failure to predict rare events, and possible overfitting.
Logistic Regression (LR) [17]	Statistical method for analyzing data with dichotomous outcomes.	Good outcomes with a small number of variables and easy to implement	Requires many samples for training, and not easy result interpretation.
DTs [17]	It used algorithm for classification and regression by dividing the data into areas of similar characteristics.	Easy to maintain and understand, good results with small data, and intuitive	Large memory footprint, tendency for overfitting, and, high variation in generated models.

TABLE II. LIST OF UNSUPERVISED ML MODELS SHOWING THEIR ADVANTAGES AND DISADVANTAGES

Model	Definition	Advantages	Disadvantages
K-Means Clustering [17]	Builds a hierarchy of clusters by connecting adjacent clusters.	Not sensitive to distance selection, accepts noisy data, does not require predetermination of the number of clusters.	Complex algorithm (O(n power 3)), cannot process large data volumes.
Hierarchical Clustering [17]	A binary classification method that creates a hyperplane to divide two target values.	High accuracy, works well with high-dimensional data, and memory efficient	Long training time, and poor performance with noisy data.

H. Ensemble Learning

It is a technique that uses multiple models to obtain better performance than any one model alone. Ensemble learning is an idea where predictions from several models of different knowledge and strategies are averaged, letting the resulting system handle overfitting and perform better [24]. Anomaly detection systems in cloud environments are improved by ensemble methods like bagging, boosting and stacking. Nevertheless, ensemble models are computationally expensive, and

their complexity makes them difficult to deploy in real time cloud environments[29].

III. ML's Role in Real-time Anomaly Detection in Cloud Computing

Fig. 4 shows several advantages of using traditional ML techniques for real time anomaly detection in cloud computing environments:

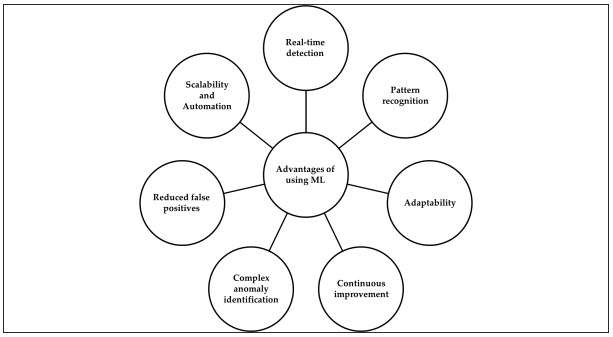


Fig. 4. Advantages of using traditional ML.

1) Real-time detection: ML is real time and immediately alerts when there are suspicious activities which is critical for responding to risks quickly [30].
2) Pattern recognition: On historical data, ML formulates complex patterns and trends and highlights normal activities against anomalies.
3) Adaptability: The ML models learn, update and change, almost continuously, to accommodate the evolving data patterns in a dynamic cloud environment.
4) Continuous improvement: Anomaly detection capabilities are made better and better by ML models through retraining with new data. This iterative improvement keeps the models effective at detecting new threats and adapting to ever changing cloud environment patterns.
5) Complex anomaly identification: Compared to other anomaly detection approaches, ML is unique because it can deal with multivariate anomalies from sources like unusual access patterns of users, different behavior in network traffic.
6) Reduced false positives: ML algorithms are trained to understand the unique behaviour of the organization's cloud environment so normal activities are less likely to trigger anomalies. With this precise tuning, security teams don't have to be overwhelmed with false alerts, but can concentrate on real threats.
7) Scalability and Automation: In cloud environment, ML consumes large volume of generated data and conducts an automated anomaly detection, reducing the manual intervention.

IV. Objective

The objectives of this research are as follows:

- To conduct a comprehensive literature review of existing research on anomaly-based ML detection in cloud computing.
- To investigate the ML techniques used, datasets used, and their accuracy.
- To examine the cloud computing models used.
- To develop a taxonomy for the systematic literature review to categorize and analyze the research findings.
- To identify the challenges and advantages of anomaly-based ML detection in cloud computing.
- To outline potential future directions for research in anomaly-based ML detection in cloud computing.

By addressing these objectives, this review aims to offer a clear understanding of the current landscape of anomaly detection using ML in cloud computing, highlight the barriers that need to be overcome, and propose directions for future innovations that can help secure cloud environments more effectively. The insights gained from this review can be valuable for both researchers and practitioners in the field of cloud security, guiding future research efforts and helping organizations implement effective anomaly detection solutions.

V. Research Methodology

We follow a systematic approach to review the existing literature on anomaly based ML detection in cloud computing, following the Preferred Reporting Items for Systematic Reviews and Meta Analyses (PRISMA) guidelines. It includes defining the research questions, selecting databases, developing search strings, establishing of inclusion exclusion criteria, and applying a quality assessment framework. The methodology is organized as follows:

A. Research Questions (RQ)

The following research questions (RQs) guide this study to provide a structured analysis of anomaly detection models in cloud computing environments:

- RQ1: What anomaly-based ML techniques are applied in cloud computing environments, and how are these models classified?
- RQ2: What datasets and evaluation metrics are used in the assessment of these models?
- RQ3: What are the primary challenges and benefits of using anomaly-based ML detection in cloud environments?
- RQ4: What gaps exist in the current literature, and what future research directions are suggested for advancing anomaly-based detection in cloud computing?

B. Data Sources and Search Strategy

To ensure comprehensive coverage of relevant studies, the search was conducted across the following academic databases:

- IEEE Xplore
- MDPI
- SpringerLink
- ScienceDirect

- ACM Digital Library

The keywords used for the selection based on the related research objectives:

("Anomaly Detection" OR "Anomaly") AND ("Machine Learning" OR "ML") AND ("Cloud Computing" OR "CC" OR "Cloud")

Only peer reviewed journal articles, conference papers published between 2020 and 2024 were considered to capture recent developments.

C. Inclusion and Exclusion Criteria

To filter search results for relevant studies, we established the following inclusion and exclusion criteria:

1) Inclusion Criteria:

- Studies that focus on anomaly detection using ML within cloud computing environments.
- Peer-reviewed journal articles, conference papers.
- Studies that provide empirical results or evaluations using datasets relevant to cloud settings.
- Publications written in English.

2) Exclusion Criteria:

- Studies not related to anomaly detection in ML applications for cloud computing.
- Publications that only provide theoretical models without empirical validation.
- Non-peer-reviewed sources such as theses, white papers, and editorials.

D. Study Selection Process

The study selection process adhered to the PRISMA framework, proceeding in three stages:

- Initial Screening: All retrieved articles were screened by titles and abstracts to exclude irrelevant studies and choose those meeting the inclusion criteria for full text review.
- Full-Text Review: Full texts of selected articles were reviewed to determine their relevance and quality. Excluded articles that did not provide detailed information on ML techniques, datasets or empirical evaluations.
- Data Extraction and Coding: A standardized form was used to extract data from the final set of articles, including anomaly detection techniques, datasets, evaluation metrics, as well as identified challenges and benefits.

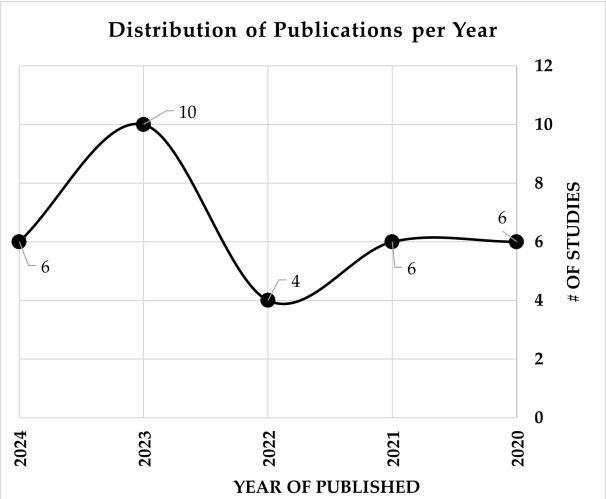


Fig. 5. Distribution of the selected studies per publication year.

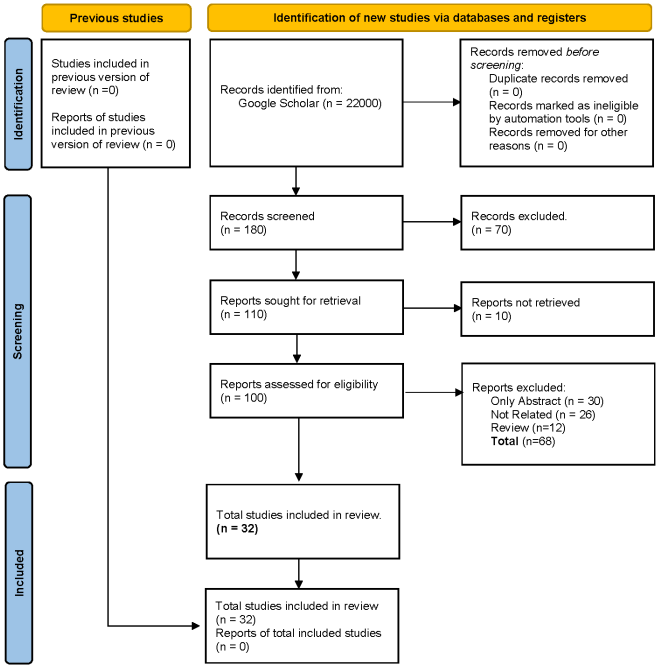


Fig. 6. PRISMA flow chart for the selection process.

E. Selection Results

With the application of selection criteria, 70 papers were excluded and 110 papers were selected for further review. Of these, 10 papers were not retrieved, and 100 articles were assessed for eligibility. At this stage, 68 articles were excluded leaving 32 articles in the final SLR. Fig. 5 shows distribution of the selected studies per publication year.

Fig. 6 shows the PRISMA flow diagram showing each stage of the process is presented in PRISAM.

VI. Literature Review

A. Supervised Models

Talpur et al. [31] presents a robust framework for DDoS attacks using evolutionary algorithms with ML models. They propose an innovative hybrid methodology that combines Extreme Gradient Boosting(XGBoost)-Genetic Algorithm(GA) Optimization, RF-GA Optimization, and SVM-GA Optimization with the Tree-based Pipelines Optimization Tool (TPOT). It automates the optimization of ML pipelines to enhance accuracy. Datasets such as KDD Cup 99 and CIC-IDS 2017 were used for the study, which achieved high accuracy scores of 99.99% for XGBoost-GA and SVM-GA, and 99.50% for RF-GA using 10-fold cross validation. Although effective, the methodology is limited by an increase in computational complexity resulting from multiple detection models and optimization phases.

Alduailij et al. [32] introduces an effective approach to DDoS attack detection in cloud environments. The main contribution is the use of Mutual Information (MI) and Random Forest Feature Importance (RFFI) for feature selection to reduce misclassification errors. The authors evaluate five ML models RF, Gradient Boosting (GB), Weighted Voting Ensemble (WVE), K-Nearest Neighbor (KNN), and LR. The method demonstrated to achieve high accuracy rates using the CICIDS 2017 and CICDDoS 2019 datasets, with RF reaching 99.997% accuracy and lowest misclassification errors when trained using 19 features. The study notes that KNN has a higher computational cost and models such as LR and GB need better parameter tuning. Although these limitations exist, the research proves that MI and RFFI feature selection can increase DDoS attack detection accuracy over different ML models.

DASARI and KALURI [33] suggests a hierarchical ML approach optimized for hyperparameter tuning to boost intrusion detection in networks due to DDoS attacks. The research uses the CICIDS 2017 dataset and preprocesses the data with normalization and balancing techniques such as Min-Max scaling and SMOTE. Feature selection is executed via the LASSO method, and the selected features are fed into five ML classifiers XGBoost, Light Gradient-Boosting Machine (LightGBM), CatBoost, RF, and DT. Model accuracy was improved through hyperparameter optimization. Of all these, LightGBM had the highest classification accuracy of 99.77%, better than all other models. It also mentions future areas of improvements in handling real time data and adaptive attacks. The work contributes to improving Intrusion Detection System (IDS) capabilities with hierarchical ML techniques for high precision and recall.

Mishra et al. [34] introduce a perplexed Bayes classifier model for identifying and mitigating DDoS attacks in cloud computing environments. This classifier uses the NSL-KDD dataset which contains DDoS attack scenarios and features. The innovation is in using correlation based feature selection to improve classification accuracy to a 99%. This method is benchmarked against the established algorithms of NB and RF and found to be more accurate, sensitive, and specific. Furthermore, they compare perplexed Bayes classifier against nature inspired feature selection techniques such as GA and Particle Swarm Optimization (PSO) and show that perplexed Bayes classifier achieves 2% to 8% higher accuracy.

Parameswarappa et al. [35] propose a new intrusion detection system for cloud computing based on ML and deep learning techniques to boost security. UNSW-NB15 dataset was used by the authors for developing and testing their model which consists of multiple classifiers, LR, KNN, DT, RF, Extra Trees, GB, and Multilayer Perceptron (MLP). It focuses on preprocessing by using K best feature selection for optimizing classification tasks. Models detected cloud anomalies and attacks with a detection rate of 97.68% by RF. The model improves precision and reduce false positives, but is limited by its dependence on labeled datasets and its use in broader cloud environments. This suggests further work of integrating advanced data mining, deep learning techniques into the existing anomaly detection process to increase its accuracy on various anomalies.

Advanced ML techniques are developed by M et al. [36] to enhance data security in cloud computing environment. They evaluate three ML models such as RF, Deep Neural Network (DNN), and Q-Learning across different experiments. RF model was reliable in categorizing security threats, with 95% accuracy and balanced precision of 0.92, recall of 0.96, and F1 score of 0.94. Also, DNN model demonstrated good performance with an accuracy of 97%, a recall of 0.98 and an F1 score of 0.96 and it could recognize complex patterns in cloud data. The research used cloud system data sets of logs, network traffic, and access patterns for anomaly detection and adaptive security response. The resource intensive nature of the deep learning models, difficulties to reduce false positives in Q-Learning and ethical issues, such as privacy preservation are also limitations. In particular, this work calls for the continuous optimization of ML models for cloud security to cope with the ever changing threat landscape.

ABUBAKAR et al. [37] propose a hybrid DDoS detection and mitigation mechanism using an optimized SVM is combined with SNORT Intrusion Prevention System (IPS). This integrated approach identifies malicious traffic early and mitigates attacks by rerouting or dropping suspicious packets. The methodology uses the KDDCup99 and DARPA datasets, while the abnormalities in real time network traffic are analyzed. The results show that the system achieves superior average detection accuracy of 97.9% compared to traditional SNORT IPS, Probabilistic Neural Networks (PNN) and Back Propagation methods. While the method has high accuracy and low false positives, it suffers from two limitations: multithreading and zero-day attack detection. The model focuses on supervised learning configurations for traffic behavior analysis and protocol validation, which is effective but limited by dataset quality and scope.

A sophisticated cloud IDS is introduced by BAKRO et al. [38] which utilizes a hybrid feature selection method combined with a RF classifier. The proposed methodology integrates Information Gain (IG), Chi-Square and PSO for selecting relevant features which increases the model accuracy and reduces the data dimensionality. The Synthetic Minority Over-sampling Technique (SMOTE) is used to address data imbalance, and robust performance in multi-class attack detection is ensured. The proposed system shows its effectiveness while detecting different attack types on the UNSW NB15 and Kyoto datasets with 98% and 99% accuracy rates, respectively. However, it

suffer from reliance on resource-intensive feature selection methods. The system is shown to significantly improve detection rates but may not necessarily generalize to real world cloud environments without further dataset diversity.

BAKRO et al. [39] presents a novel cloud IDS. The hybrid feature selection technique Grasshopper Optimization Algorithm (GOA) and GA aims to optimize feature selection and enhance IDS performance by improving classification accuracy and reducing computational complexity. This hybrid approach helps optimize feature selection while increasing accuracy of the classfied data and reducing amounts of computation. The model uses an RF classifier trained on the selected features with an ADAptive SYNthetic (ADASYN) algorithm for minority oversampling and RUS for majority class balancing. The proposed system is evaluated on three datasets including UNSW-NB15, CIC-DDoS2019, and CIC Bell DNS EXF 2021 and achieves accuracy of 98%, 99% and 92% respectively. However, because it is dependent on the specific datasets to evaluate on, and may not be generalizable. It achieved improvements in True Positive Rate (TPR) and False Positive Rate (FPR) along with better performance than state of the art classifiers like SVM, AlexNet and XGBoost.

A novel framework for cloud anomaly detection using a Secure Packet Classifier (SPC) is proposed by Chkirbene et al. [40] The SPC combines two ML algorithms selected based on accuracy and computational efficiency and leverages collaborative filtering. The main focus of the model is anomaly detection and classifying different types of attack which is crucial for targeted counter measures. Using the UNSW-NB15 dataset, the model delivered an impressive improvement on accuracy, detecting 81% of anomalies with a lower FPRthan the traditional methods. The work is constrained by the fact that the model relies on specific dataset properties and does not generalize to other datasets without significant retraining.

Aldallal and Alisa [41] propose to develop a hybrid IDS for cloud computing environments, which integrates GA for feature selection and SVM for classification. A novel fitness function is used by the system to measure the performance of the intrusion detection system, combining F1-score, accuracy and TPR to ensure balance and accuracy of the detection system. They used CICIDS2017 dataset for evaluation, results show that the proposed model provided up to 5.74% improvement on detection accuracy over benchmarks, while demonstrating its effectiveness. Although the system performed well, it required data preprocessing, including cleaning missing or corrupted entries, and relied on predefined datasets for evaluation rather than real-time data.

Jaber and Rehman [42] propose an IDS for cloud computing environments by combining Fuzzy C-Means (FCM) clustering with SVM. The hybrid FCM–SVM model proposed can overcome the limitations of conventional IDS including high false alarm rates and poor accuracy. The proposed system is implemented using the NSL-KDD dataset, with clustering used to group data points for improving SVM performance in anomaly detection. The system shows a capability to classify different types of network attacks with an accuracy of 97.37% for User to Root (U2R) attacks, 98.46% for Remote to Local (R2L) attacks and 98.85% for Probe attacks.

AlSaleh et al. [43] proposes a novel ML approach for detecting DDoS attacks in cloud computing settings using a Bayesian Convolutional Neural Network (BaysCNN). To achieve significant improvements in DDoS detection accuracy, BaysCNN uses a 19 layer architecture with an average accuracy of 99.66% across 13 multi class attacks. The study also improves model performance using the Bayesian-based Convolutional Neural Network with Data Fusion (BaysFusCNN) approach, which combines features from different sources, yielding a better accuracy of 99.79%. This research demonstrates that these models can tackle challenges including distinguishing application-layer attacks and real time detection. Bayesian methods are also used in the models to estimate uncertainties to improve reliability. The limitation is the dependence on the CICDDoS2019 dataset for which the results are not generalizable to other datasets and environments.

Sherubha et al. [44] propose a novel anomaly detection mechanism through an auto-encoder for feature selection and a NB classifier for classification. The approach improves the ability of existing IDS to deal with unlabeled data and reduces redundancy and noise in datasets. On NSL-KDD dataset, the model shows a detection accuracy of 93% which is better than traditional methods like J48 and RF. The main contribution of this study is the combination of unsupervised learning for dimensionality reduction and NB for robust classification, which achieves superior performance in detecting network anomalies. However, the approach is tested only on a static dataset, which limits its real time applicability and ability to address zero day attacks. The results of this research highlight the possibility of application of hybrid methodologies in order to improve intrusion detection in cloud computing environments.

Moreira et al. [45] propose ISAD; an intelligent system for anomaly detection in smart environments based on the integration between Fog and cloud computing. The system uses ML techniques to process network traffic and detect unusual behavior, offloading the data processing overhead to the Fog and cloud environments. A fog layer is used to perform raw network traffic data processing, feature extraction, and transmit filtered data to the cloud for dynamic anomaly detection using ML models. The system achieves high accuracy especially with RF, achieving 98.7% accuracy with Microsoft Azure. The CICIDS dataset is used, which represents real network traffic scenarios. The system shows robust performance, but it is reliant on fog and cloud environment computational infrastructure and has reduced recall in some ML configurations, which poses challenges in generalizing anomalies.

The authors, Alshammari and Aldribi, presents a lightweight ML based framework to boost IDS for detecting network anomalies [46]. In order to assess its performance, the framework utilizes the ISOT-CID dataset by incorporating novel features, more specifically the 'rambling feature', in classification. Also, In this study, six ML models such as DT, RF, and KNN are evaluated by using cross-validation and split validation techniques. Therefore, results show that DT and RF models are the most accurate, with 100% accuracy in both validation strategies. The novel feature addition and data preprocessing make the dataset better in quality, making the training of the models effective. While the model worked well, it requires large datasets and is not ready for real time deployment because of latency issues. This work points to future work, where deep learning approaches will be integrated to

overcome these limitations and better refine anomaly detection in real-time cloud environments.

Al-jumaili and Bazzi investigate the use of ML models to improve IDS in cloud environments, that suffer from dynamic threats and false positives [47]. The research compares and analyzes algorithms like DT, RF, XGBoost and SVM and finds XGBoost as the best effective model with 99.63% accuracy as it has a great GB capability. They use NSL KDD dataset, which is well known for its rich network intrusion patterns and perform robust preprocessing such as label encoding and data scaling to improve the performance of the model. However, since the research is based on synthetic datasets, it is not applicable in real world scenarios and more realistic cloud traffic should be validated. The results help explain how ML can be used to develop robust and scalable IDS solutions for the ever changing cloud landscape.

Naiem et al. [48] proposes a new framework to optimize the Gaussian Naïve Bayes (GNB) classifier for DDoS detection in cloud environments. The research acknowledges the GNB's limitations, dependency on feature independence, and sensitivity to the zero frequency problem by addressing them with an iterative feature selection process and advanced preprocessing. Also, feature selection techniques such as the Pearson Correlation Coefficient (PCC), MI and Chi-squared tests are used, and the SMOTE algorithm is used to address data imbalance. They demonstrate a 2% improvement in accuracy and substantial gains in precision, recall and F1-score. In addtion , the approach improves GNB's performance on the CICD2018 dataset to the level of other classifiers such as RF and SVM with the simplicity and computational efficiency.

Aslan et al. [49] propose a new cloud based malware detection system focusing on intelligent behavior analysis. The proposed Cloud-Based Behavior-Centric Model (CBCM) collects execution traces of suspicious files in VMs, identifies relevant behaviors, and extracts discriminative features. Both learning based and rule based detection agents process these features. They used ML classifiers such as RF and logistic model trees with 99.83% accuracy and a 0.6% FPRon a dataset of 10,000 samples for RF. Also, real time detection is further augmented by rule based agent. In addtion, the work uses cloud scalability to efficiently analyze malware, showing the high accuracy and speed of detection compared to traditional methods. Limitations, however, exist in the form of difficulty in detecting advanced obfuscated malware, and the requirement for broader dataset diversity. This research greatly enhances the malware detection efficiency in cloud computing environments.

Mehmood et al. [50] offer an advanced framework of privilege escalation attacks detection and mitigation within cloud computing environments, which is based on ensemble learning. The research works on a customized dataset from multiple CERT dataset files, using ML algorithms RF, Adaptive Boosting (AdaBoost), XGBoost and LightGBM to classify and mitigate insider threats. Moreover, the highest accuracy (97%) was achieved by LightGBM, which was better than RF (86%), AdaBoost (88%) and, XGBoost (88.27%). This was achieved by pre processing the dataset, training the models, and tuning the hyperparameters to solve the specific attack scenarios thereby leading to a robust detection mechanism. Insider threat research concludes that insider threats, in particular those resulting from privilege abuse, are especially critical and suggests ensemble learning for increased classification accuracy. Although the study yields promising accuracy, it is limited to a single dataset and is challenged in recognizing subtle attack patterns, which suggests future exploration of the diversity of datasets.

Bamasag et al. [51] introduce the Real-Time DDoS flood Attack Monitoring and Detection (RT– AMD) model which is aimed at mitigating the effects of DDoS attacks on cloud computing environments. The RT-AMD model works to utilize ML algorithms such as RF, KNN, NB, and DT, with high accuracy, to detect abnormal network traffic. The model is evaluated on the DDoS-2020 dataset, which contains balanced attack and normal traffic records for Transmission Control Protocol (TCP), Domain Name System (DNS), and Internet Control Message Protocol (ICMP) protocols, and achieves an accuracy of 99.38% in real time detection. The incremental learning capability further improves real time adaptation and detection without a costly retraining process. The study reveals the scope of expansion in its impressive accuracy and performance, including the incorporation of various DDoS types and evaluating on other cloud environments. This is key to advancing secure, real time cloud operations.

Chkirbene et al. [52] introduce a novel ML based weighted class classification scheme to tackle the challenges of anomaly detection in cloud computing, in the context of class imbalance problem. The system improves the classification accuracy of rare attack classes by integrating supervised learning with past node behavior and a weight optimization algorithm. The approach involves training a DT classifier on the UNSW-NB15 data set and using historical data to determine decision weights and obtains 95% accuracy. This framework significantly enhances multi class detection capabilities and is resilient to underrepresentation of minority attack classes. However, the reliance on historical data and the computational cost of weight optimization restrict the model's adaptability in dynamic real time environments.

In their work, Sambangi and Gondi investigate the use of MLR to detect DDoS attacks in cloud environments [53]. The study uses the CICIDS2017 dataset, focusing on the Friday afternoon traffic logs and applies an IG based feature selection technique to select critical attributes, reducing the dimensionality from 79 to 16 and then to 6. It is shown that the MLR model achieves 73.79% accuracy with 16 attributes. In addition, residual plots and fit charts are used by the authors to visualize the model's ability to differentiate between benign and attack traffic. The research is restricted to single day log data and does not explore ensemble or deep learning methods. This work contributes to enhancing the DDoS detection efficiency by simplifying the feature selection process and focusing on the regression analysis to handle the attack classification issues.

In response to DDoS attacks in cloud environments, Wani et al. [54] propose a robust IDS. They created a unique dataset with 21 attributes by using the CloudStack platform for experimentation and using Tor Hammer for generating malicious traffic. Also, the researchers evaluated six ML models: DT, NB, RF, C4.5, and SVM, K-Means. The SVM algorithm showed the best accuracy of 99.7% among the algorithms such as C4.5 (98.7%) and RF (97.6%). Furthermore, to evaluate the performance of the system, the dataset was analyzed by using Weka tool and evaluating the performance using precision,

recall metric. However, the study shows that data imbalance is a challenge, but does not consider more general attack scenarios or feature selection optimization. They contribute by showing that SVM is effective in detecting anomalies but their work is limited to specific tools and attack types.

P et al. [55] propose a ML framework for detection of phishing attacks in distributed cloud systems. In particular, the authors use supervised learning algorithms such as NB, SVM, and DT to detect phishing attacks. The study evaluates and compares the performances of these algorithms in terms of accuracy by using IDS generated dataset and shows that DT is the best method with slightly slower than other algorithms. The work aims at solving the critical challenge of resource management in cloud systems, which is often exploited for phishing attacks. Although the proposed model has a good detection accuracy, its response time is slower than the current standards and it depends on pre processed datasets which does not allow it to adapt to real time. The results highlight the importance of feature reduction and classification methodology for improving detection rates and reducing false alarms in cloud computing environments.

Kushwah and Ranga propose a novel system based on a Voting based Extreme Learning Machine (V-ELM) to detect DDoS attacks in cloud computing environments [56]. Unlike conventional single layer neural network approaches, the system uses multiple Extreme Learning Machines (ELM) with a majority vote mechanism to improve detection accuracy and reduce false alarms. The proposed model is evaluated using NSL-KDD and ISCX datasets and the accuracies of 99.18% and 92.11% respectively are shown. Accuracy and false positive rates for the system are shown to be superior to traditional models such as RF and Adaboost. But, this requires great amounts of labeled training data which can constrain it. Finally, the research mainly focuses on the challenges of high detection accuracy, false positive rates, and fast training and offers an efficient scalable DDoS attack detection solution.

Guezzaz et al. [57] propose a cloud based intrusion detection model by using the RF algorithm and feature engineering for improved anomaly detection. The research discusses the ever growing security challenges in cloud environments including unauthorized intrusions and real time attack detection. For the proposed framework, the feature set is reduced to only two important attributes of the NSL-KDD and Bot-IoT datasets while using data visualization to ease the feature engineering process. It achieved accuracy rates of 98.3% and 99.99% on these datasets. Although the model was able to achieve high precision and accuracy, it has low recall, meaning that it cannot detect all attacks of some types. It shows that RF can be a better classifier than SVM and DNN with a minimal feature set, and thus the work demonstrates the potential for using a minimal feature set for effective classification. Yet, there is room for improvement in recall and evaluation has yet to be done across various datasets. Table III presents a summary of supervised ML models, datasets and their accuracy rates as demonstrated by the related works.

The attacks, advantages, and disadvantages of supervised models for anomaly detection in cloud computing are highlighted in Table IV.

TABLE III. SUMMARY OF SUPERVISED MODELS AND THEIR ACCURACY RATES BASED ON RELATED WORK

Author	Year	ML Models Used	Dataset	Accuracy Rate
Talpur et al. [31]	2024	XGBoost-GA, SVM-GA	KDD Cup 99, CIC-IDS 2017	99.99%
Alduailij et al. [32]	2022	RF	CICIDS 2017, CI-CDDoS 2019	99.997%
DASARI and KALURI [33]	2024	LightGBM	CICIDS 2017	99.77%
Mishra et al. [34]	2022	Perplexed Bayes Classifier	NSL-KDD+	99%
Parameswarappa et al. [35]	2023	RF	UNSW-NB15	97.68%
M et al. [36]	2024	DNN	Cloud-based	97%
ABUBAKAR et al. [37]	2020	SVM	KDDCup99 and DARPA	97.9%
BAKRO et al. [38]	2023	RF	Kyoto	99%
BAKRO et al. [39]	2024	GOA-GA with RF	CIC-DDoS2019	99%
Chkirbene et al. [40]	2021	SPC	UNSW-NB15	81%
Aldallal and Alisa [41]	2021	GA with SVM	CICIDS2017	99.65%
Jaber and Rehman [42]	2020	FCM with SVM	NSL-KDD	98.85%
AlSaleh et al. [43]	2024	BaysFusCNN	CICDDoS2019	99.79%
Sherubha et al. [44]	2023	NB	NSL-KDD	93%
Moreira et al. [45]	2021	RF	CICIDS	98.7%
Alshammari and Aldribi [46]	2021	RF	ISOT-CID	100%
Al-jumaili and Bazzi [47]	2023	XGBoost	NSL-KDD	99.63%
Naiem et al. [48]	2023	GNB	CICD2018	97.57% with PCC-IM
Aslan et al. [49]	2021	RF	Various sources and forming (7000 malware, 3000 benign)	99.83% with cross-validation
Mehmood et al. [50]	2023	LightGBM	customized dataset derived from multiple CERT dataset files	97%
Bamasag et al. [51]	2022	RF	DDoS-2020	99.38%
Chkirbene et al. [52]	2020	DT	UNSW-NB15	95%
Sambangi and Gondi [53]	2020	MLR	CICIDS2017(Friday afternoon traffic logs)	73.79% using 16 features
Wani et al. [54]	2020	SVM	Custom dataset with 21 attributes	99.7%
P et al. [55]	2023	DT	IDS-generated dataset	87%
Kushwah and Ranga [56]	2020	V-ELM	NSL-KDD	99.18%
Guezzaz et al. [57]	2023	RF	Bot-IoT	99.99%

B. Unsupervised Models

Shanthi and Maruthi [58] introduced a new method to build anomaly-based IDS in cloud computing environment with the combination of Isolation Forest and SVM models. The proposed system aims to improve the efficiency and accuracy of detecting anomalous activities in large and complex network datasets. The study uses the NSL-KDD dataset to evaluate the performance of both models with Isolation Forest attaining an accuracy of 99% and SVM of 95%. Isolation Forest isolates anomalies via recursive random splits, and a supervised binary classifier SVM learns to identify anomalies by learning the normal vs anomaly distinction. Although Isolation Forest

TABLE IV. SUMMARY OF THE ATTACKS FOCUS, ADVANTAGES, AND DISADVANTAGES OF SUPERVISED MODELS IN ANOMALY DETECTION IN CLOUD COMPUTING

Ref	Attacks	Advantages	Disadvantages
[31]	DDoS	Superior pipeline optimization with TPOT.	Increased computational complexity.
[32]	DDoS	Integration of MI and RFFI for feature selection.	LR and GB require tuning to reduce errors.
[33]	DDoS	Hyperparameter tuning significantly improved overall performance metrics.	Limited to CICIDS dataset.
[34]	DDoS	Efficient feature selection method.	Dependent on pre-processed and structured datasets.
[35]	DDoS	Integration of historical and real-time decisions.	Dependency on labeled datasets.
[36]	Data breaches, unauthorized access	Ability to detect complex patterns.	Resource-intensive computation, privacy concerns.
[37]	DDoS	Early detection and mitigation of attacks.	Limited handling of zero-day attacks.
[38]	DoS, worms, and exploits.	High accuracy, robust detection rates.	Resource-intensive feature selection.
[39]	DoS, DDoS, and DNS attacks	Enhanced classification accuracy.	Limited evaluation datasets.
[40]	Analysis, Backdoor, DoS, Exploits, and others	Effective classification of attack types.	Requires significant retraining for new environments.
[41]	Brute-force attacks, SQL Injection, and others	Effective feature selection	Requires significant preprocessing for corrupted data.
[42]	U2R, R2L, Probe and DoS	High accuracy, low false alarm rates.	Dependency on pre-defined datasets.
[43]	Multi-class DDoS attacks	High accuracy.	Potential computational overhead for real-time scenarios.
[44]	DoS, probe, R2L and U2R	Effective feature selection	Ineffective against zero-day attacks
[45]	Network anomalies, including DDoS	Reduces data processing overhead by fog preprocessing.	Recall performance varies across models.
[46]	Malicious network traffic	Novel features like "rambling" improved detection performance.	Limited real-time deployment capability.
[47]	Network intrusions, anomalies, Unknown and zero-day attacks	Balanced precision and recall.	High computational cost for XGBoost.
[48]	DDoS	Iterative feature selection improves accuracy and reduces overfitting.	GNB still underperforms compared to more advanced classifiers.
[49]	Malwares	Combines ML and rule-based approaches for robustness.	Imbalanced dataset distribution of malware types.
[50]	Privilege escalation attacks (horizontal and vertical) and insider threats.	Efficient in identifying privilege escalation.	Limited dataset diversity restricts generalizability.
[51]	DDoS flood attacks targeting the network/transport layer (TCP, DNS, ICMP)	Real-time detection, high accuracy, uses incremental learning, and adaptive model.	Limited to specific attack types.
[52]	Analysis, Backdoor, DoS, Exploits, Fuzzers, Reconnaissance, Shellcode, Worms.	Enhances detection accuracy for rare attack classes.	Relies heavily on historical data; computationally intensive weight optimization.
[53]	DDoS	Demonstrates effectiveness of MLR	Limited to single-day traffic logs.
[54]	DDoS	Custom dataset tailored to cloud-specific scenarios.	Dataset imbalance issues.
[55]	Phishing attacks	Highlighted the importance of feature reduction to enhance detection speed and accuracy.	Slower response time.
[56]	DDoS	High detection accuracy.	Requires large labeled datasets.
[57]	General anomaly detection.	Low computational overhead due to reduced feature set.	Poor recall, and limiting generalizability.

has high accuracy and adaptability, it is also dependent on strong parameter tuning like contamination factor, and SVM is dependent on data portioning. The work suggests that careful feature selection and dimensionality reduction are necessary to enhance detection abilities.

Ntambo and Adeshina present a proactive anomaly detection model for detecting anomalies in VM resource usage in cloud environments with emphasis on the multitenancy vulnerabilities in public clouds [59]. The model uses Isolation Forest and One-Class Support Vector Machine (OCSVM) algorithms to detect anomalies in deviations of VM metrics such as CPU, memory usage and disk throughput. The dataset used is from the Grid Workload Archive, consisting of time series VM resource data. The results show that OCSVM gains the best accuracy with F1 = 0.97 for hourly and F1 = 0.89 for daily series compared to Isolation Forest. However, these results rely on specific datasets and only consider a limited set of real world scenarios. In this work, they contribute a scalable approach to augmenting cloud security with real time anomaly detection capabilities for VM resources.

Table V presents a summary of unsupervised ML models, datasets and their accuracy rates as demonstrated by the related works.

The attacks, advantages, and disadvantages of unsupervised models for anomaly detection in cloud computing are highlighted in Table VI.

TABLE V. SUMMARY OF UNSUPERVISED MODELS AND THEIR ACCURACY RATES BASED ON RELATED WORK

Author	**Year**	**ML Models Used**	**Dataset**	**Accuracy Rate**
Shanthi and Maruthi [58]	2023	Isolation Forest	NSL-KDD	99%
Ntambo and Adeshina [59]	2021	OCSVM	Grid Workload Archive	not specified.

C. Hybrid Models

Wang et al. [60] proposes a hybrid anomaly detection system for cloud computing environments using a Stacked Contractive Autoencoder (SCAE) and a SVM. A deep learning based SCAE for unsupervised feature extraction is proposed which transforms raw network traffic data into low dimensional robust representations. An SVM is used to classify these features for malicious activities. The methodology is evaluated on two benchmark datasets, KDD Cup 99 and NSL-KDD. Experimental results show that the model achieves good detection rates and has an accuracy of 87.33% in multi class classification tasks. The research also points out the limitations

TABLE VI. SUMMARY OF THE ATTACKS FOCUS, ADVANTAGES, AND DISADVANTAGES OF UNSUPERVISED MODELS IN ANOMALY DETECTION IN CLOUD COMPUTING

Ref	Attacks	Advantages	Disadvantages
[58]	Network-based such as intrusions	High accuracy, robust feature extraction, efficient handling of large datasets.	Dependence on hyperparameter tuning, limited response to encrypted traffic.
[59]	Anomalies in VM resource usage, including stealthy attacks exploiting multitenancy in public clouds	Combines VM metrics for improved anomaly detection.	Limited generalizability due to specific datasets.

of the classifier in identifying less common attack types and identifies directions for further optimizing the classifier. The results of this work suggest that hybrid models are a viable solution to the scalability and precision problems inherent in cloud IDS.

A hybrid clustering and classification-based approach to intrusion detection in distributed cloud computing environments is proposed by Samunnisa et al. [61]. They introduces a ML-based anomaly detection system that combines K-means clustering with RF classifiers to classify malicious activities across five types DoS, Probe, U2R, R2L, and normal. The model is tested using NSL-KDD and KDDCup99 datasets and shows high accuracy and low false alarm rates with 99.78% detection rate and 0.09% false alarm rate for NSL-KDD dataset. The methodology uses threshold-based functions and measures accuracy, detection rate and the area under the curve. The study points out that the datasets used are out dated and do not necessarily reflect the current network threats. It demonstrates that hybrid models can improve IDS but the use of the outdated datasets may limit their applicability to modern, evolving attack patterns..

Megouache et al. [62] present a new framework for intrusion detection that combines clustering and classification in cloud environments. It uses the K-means clustering to label previously unlabeled datasets and is able to use the ELM classifier to quickly identify and prevent malicious activities. Also, the proposed system is tested on the KDD99 dataset, where an accuracy of 99.2% in detecting non legitimate users is achieved. The main innovation is in the integration of clustering and classification to perform data segmentation and intrusion detection in an optimized manner, dealing with issues like scalability and false positives. In addition, the approach uses probabilistic methods to minimize data loss risks and improve real time cloud security. Still, the method is limited by the time needed to perform matrix operations to train large scale datasets. In summary, the work provides a high speed, accurate, and scalable solution for intrusion detection in the cloud, which is of significant importance to cloud security, but requires further work to improve scalability and processing efficiency.

Table VII presents a summary of hybrid models, datasets and their accuracy rates as demonstrated by the related works.

The attacks, advantages, and disadvantages of hybrid models for anomaly detection in cloud computing are highlighted in Table VIII.

D. Taxonomy of the Research

The research taxonomy Fig. 7 has been systematically organized in order to categorize ML methods used in the cloud computing and hence give a clear picture and structure to the field. The taxonomy categorized ML as supervised, unsupervised and hybrid techniques. Furthermore, this taxonomy benefits researchers and practitioners to select the specific dataset and to select the most appropriate technique for their use cases. Also, the purpose of this taxonomy ultimately is to provide a framework for driving innovation in and making better decisions about using ML to solve cloud computing issues.

TABLE VII. SUMMARY OF SUPERVISED MODELS AND THEIR ACCURACY RATES BASED ON RELATED WORK

Author	**Year**	**ML Models Used**	**Dataset**	**Accuracy Rate**
Wang et al. [60]	2022	SCAE and SVM	NSL-KDD	87.33%
Samunnisa et al. [61]	2023	K-means with RF	NSL-KDD	99.85%
Megouache et al. [62]	2024	K-means Clustering, ELM	KDD99	99.2% for detecting non-legitimate users

VII. CASE STUDY: PRIVILEGE ESCALATION ATTACK DETECTION IN CLOUD COMPUTING USING ML

A. Analyzing Real-World Case Studies

The study by Mehmood et al. [50] is a good example of providing ML solutions to a real world inspired scenario that addresses the issue of privilege escalation attacks in cloud computing. An example of privilege escalation attacks is cases where attackers abuse the faulty system vulnerability, misconfiguration or inadequate access control to receive increased access to resources or information. These attacks can be classified across horizontal privilege escalation or where an attacker gains access to another user's privileges, as well as vertical privilege escalation or where the attacker gains the higher level of access like administrative or root privileges. Based in that and without a doubt, privilege escalation attacks can grant an attacker unauthorized access to sensitive information or cause disruption of the critical system operations leading to through severe data breaches.

For this study, a customized CERT dataset was used to aggregate user behavior logs from various sources to emulate real world insider activities. Furthermore, malicious activities like unauthorized file copy to deleting files and abnormal system access patterns were also part of these activities.

The proposed methodology is to design a ML enabled insider threat detection system to classify and address privilege escalation attacks. The dataset was then preprocessed using carefully designed strategies to remove outliers, manage missing values and select appropriate features. The dataset was able

TABLE VIII. SUMMARY OF THE ATTACKS FOCUS, ADVANTAGES, AND DISADVANTAGES OF HYBRID MODELS IN ANOMALY DETECTION IN CLOUD COMPUTING

Ref	Attacks	Advantages	Disadvantages
[60]	R2L, U2R, Probe, and DoS.	Efficient feature extraction; high accuracy in detecting major attack types.	Limited detection of less-represented attack types.
[61]	R2L, U2R, Probe, and DoS.	High accuracy, low false alarm rates.	Relies on outdated datasets.
[62]	Anomalies and malicious user identification.	High accuracy in intrusion detection.	Computational inefficiency with large datasets due to matrix operations.

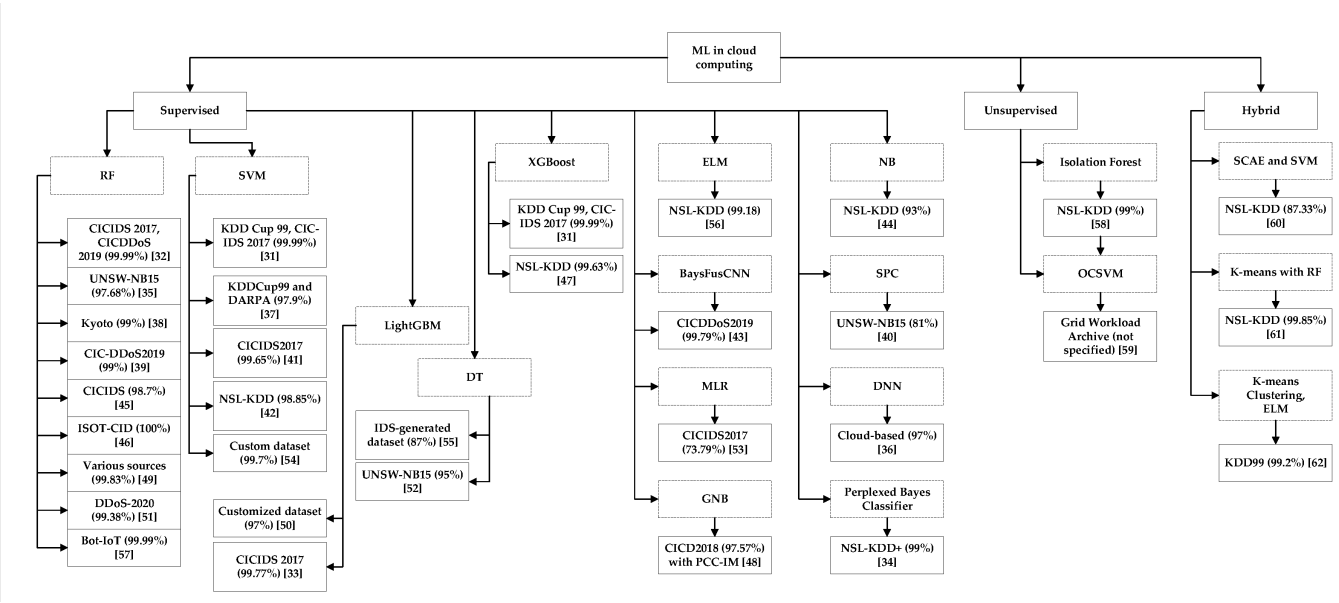


Fig. 7. Taxonomy of the Research

to simulate real world activities to provide an effective foundation in evaluating ML algorithms in real-world scenarios. In addition to ensuring the practical applicability of the models, this approach brought to light common challenges, including imbalanced data and effective feature engineering.

B. Insights from Actual Implementations

For detecting and classifying insider threats, the study employed four ML algorithms: AdaBoost, RF, XGBoost and LightGBM. The results of these algorithms were evaluated based on accuracy, precision, recall and F1 score for each algorithm. It is to note that, for the largest dataset, the LightGBM algorithm achieved the highest accuracy of 97%, due to its good leaf wise growth technique and also its performance on large dataset. Key insights from the implementation include:

- Algorithm Selection and Optimization: The results illustrated that using an ensemble approach is important to improve prediction accuracy by combining multiple models. The boosting techniques, especially LightGBM, were very good at catching complex patterns in high dimensional data, and are very effective for real world insider threat detection.
- Feature Engineering: Feature selection and preprocessing were emphasized as being critical. To improve model performance, features, that were irrelevant like ”employee” or ”file tree” were removed. The algorithms further gained stability on the learning of data by having data normalized and aggregated.
- Challenges in Data Quality: A large part of the first preprocessing phase involved handling missing values and outliers in the dataset. For example, while imputing the missing values in ‘File Copy’ feature, missing values were filled with the value which is found in the dataset patterns, hence not affecting the training process.
- Performance Metrics and Comparative Analysis: The performance metrics and confusion matrices helped them locate weaknesses and strengths of each algorithm. The high accuracy and low false alarm rate of LightGBM indicate that this solution can be used as a robust solution in real world deployments. Table IX shows the comparative analysis of performance metrics for the ML models in case study.
- Mitigation Strategies: The study not only discusses how detection of privilege escalation attacks is possible, but also suggests how such attacks can be suppressed to make them infeasible. As such, these include developing multifactor authentication and models of behavioral biometrics and secure access controls.

TABLE IX. LIST OF SUPERVISED ML MODELS SHOWING THEIR ADVANTAGES AND DISADVANTAGES

Algorithm	Accuracy	Precision	Recall	F1-Score	False Alarm Rate
RF	86%	86%	85%	85%	0.19
AdaBoost	88.27%	88%	86%	86%	0.16
XGBoost	89%	88.27%	87%	87%	0.13
LightGBM	**97%**	**97%**	**95%**	**95%**	**0.11**

C. Lessons Learned and Best Practices

The successful application of ML models in this case study highlights several best practices for implementing anomaly detection in cloud computing:

- Datasets are adapted to reflect real world scenarios to keep things practical and makes for better algorithm training.
- Using ensemble methods like LightGBM and XGBoost, combining the power of multiple methods, we have much higher accuracy and better reliability in threat detection.
- Preprocessing impact, They address the data quality issue of missing values and irrelevant features since these issues decrease model performance.
- The detection capabilities are complimented with preventive measures to include multifactor authentication and access control to provide additional security.
- Employees should be educated about cybersecurity best practice to reduce insider threats.

As indicated in this case study, ML algorithms have the potential to adequately deal with these complex cloud security challenges. These implementations provide valuable insights to guide anomaly detection projects in cloud environments in the future.

VIII. RESULTS AND DISCUSSION

The reviewed papers highlight several ML models that excel in anomaly detection for cloud computing environments. RF is shown to be the dominant supervised model as shown in Fig. 8, capable of reaching accuracy rates of 99.997% when complemented with powerful feature selection techniques such as MI and RFFI [32]. Other top performers include SVM, which may achieve accuracy of 99.99%, using techniques such as GAs or clustering [31], [41]. LightGBM and XGBoost, which are tree-based ensemble methods, have competitive accuracy 99.77% and higher when they are hyperparameter tuned [33], [47].

Anomaly detection associated with unsupervised and hybrid models is also critical. In unsupervised scenarios, Isolation Forest is very effective with 99% accuracy using recursive random splits [58]. Other hybrid models such as K-means clustering with a RF classifier deliver remarkable performance, and achieved an accuracy up to 99.85% [61]. Other innovative approaches, for example using autoencoders with a classifier such as SVM, were explored that can extract robust features and still achieve detection accuracy as high as 87.33% [60].

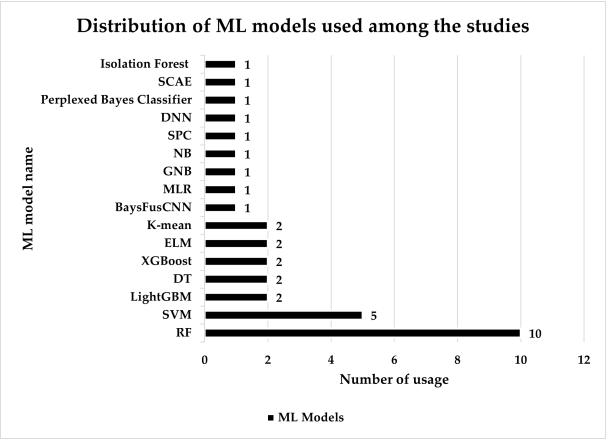


Fig. 8. Distribution of ML models used among the studies.

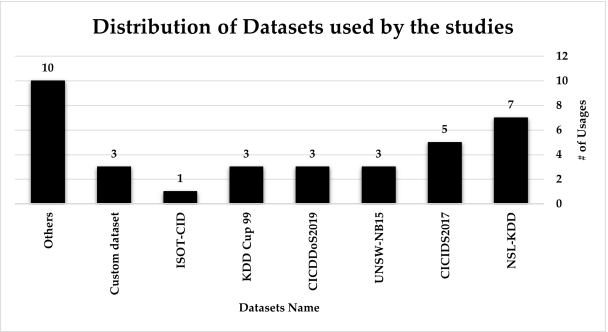


Fig. 9. Distribution of datasets used by the studies.

The challenges in imbalanced datasets and multi class anomaly detection are addressed by these hybrid and unsupervised models [52], [38].

Several advantages of using ML for anomaly detection in cloud environments are observed from the reviewed studies. Data dimensionality is reduced effectively, providing higher precision to a model using advanced feature selection techniques like LASSO, Chi-Square tests, and GAs [33], [38]. Often, hybrid approaches, combining clustering and supervised classification or unsupervised feature extraction with supervised classification are found to increase the accuracy and reduce the number of false positives [61], [62]. A second strength is scalability, and some studies have exploited cloud and fog computing to offload heavy computations and realize real time anomaly detection [45], [51].

However, the studies also identify notable limitations. Techniques such as genetic optimization and Bayesian networks improve accuracy add computational complexity, restricting their real time applicability [31], [43]. In many models, the datasets are highly dependent and based on CICIDS2017 and NSL KDD as shown in Fig. 9, which may limit their applicability in various and evolving cloud environments [41], [42]. Even high performing models such as DNNs continue to struggle with false positives [36]. Additionally, the experimental results are promising, but most systems are only tested on static datasets, so they are unable to handle real time detection or adapt to zero-day attacks [44], [53].

Feature handling is a critical factor for improving ML performance. Class imbalance and datasets are normalized using preprocessing techniques like Min Max scaling, SMOTE and ADASYN, that help improve model accuracy [33], [38]. LASSO and clustering (e.g., FCM–SVM hybrids) are adopted to perform feature sets optimization [42], [62]. In addition, several studies introduce new features, including rambling features and behavioral analytics, which can be used to enhance anomaly detection in a particular cloud scenario [46], [49].

RF and SVM models generally achieve high precision and recall, but perform poorly in detecting complex attack patterns [36], [43]. Techniques such as probabilistic methods and ELMs highlight speed and computational efficiency but the scalability is a concern with large scale datasets [62], [34]. The integration of fog and cloud improves system scalability at the cost of additional infrastructure [45], [51].

From these studies, valuable insights are drawn that emphasize the need for real time detection capability, including incremental learning in RT-AMD models that enable the models to adapt to dynamic threats [51]. However, the usage of outdated or synthetic datasets like KDD Cup 99 shows a need to have various and realistic benchmarks [31], [42]. Adaptive models, such as Bayesian methods and RL, are also studied by some to respond to changing threats, but these solutions are generally not optimized [43], [36]. Moreover, [36] emphasize ethical considerations especially on privacy preservation, in deep learning based systems.

IX. Challenges and Open Directions

- Development of Realistic Datasets: Future research should focus on creating and utilizing realistic, diverse, and up to date datasets that follow the real-world traffic patterns, including zero-day attacks and multi vector threats. This gap can be addressed by collaborative datasets derived from actual cloud systems.
- Scalable and Low-Latency Models: By optimizing ML models for computational efficiency, especially in the context of deep learning frameworks, it can help to deploy them in real time, high traffic environments. Incremental learning techniques, like those of RT-AMD, are promising areas for expansion [51].
- Advanced Feature Engineering: Additionally, incorporating novel features, behavioral analytic or domain specific characteristics rambling features proposed by [46] improve anomaly precision and reduce false positives.
- Integration of Hybrid Approaches: Addressing challenges such as data imbalance, generalization across datasets are possible by combining the strengths of unsupervised and supervised models such as the use of autoencoders for feature extraction and classes like SVM for detection.
- Enhanced Validation: Several, very different, datasets need to be validated across multiple models to promote generalizability. By benchmarking against real-time traffic from cloud providers [47], it can better evaluate their effectiveness in practical applications.
- Reduction of False Positives: Future studies include transferring ensemble methods or an advanced voting method (V-ELM) with the objective of reducing false positives and improving detection reliability [56].

X. Conclusion

Anomaly detection on cloud computing has emerged as an important application of ML and has provided us with robust tools to address evolving security challenges. RF and LightGBM models are very accurate on structured data and Isolation Forest capable of handling novel threats. Hybrid models that combine clustering with classification can deal with imbalanced datasets and with more complex attack patterns. Despite these advancements, limitations persist include the dependency on out-of-date dataset, computational inefficiencies, and high false positive rates. Future work must focus on developing realistic datasets that capture the dynamic nature of cloud environments, design of scalable models for real-time detection, and improvement of hybrid approaches to tradeoff between accuracy and adaptability. Techniques like incremental learning, advanced feature engineering, and ensemble methods should be prioritized to minimize false positives and improve model generalizability. Addressing these challenges will solidify the role of ML in protecting cloud infrastructures, towards making computing secure and resilient for a wide range of applications.

Funding

This work was funded by King Faisal University, Saudi Arabia. [Project No. GRANT KFU250561].

Acknowledgment

This work was supported through the Annual Funding track by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Project No. GRANT KFU250561].

Conflicts of Interest

All authors declare no conflict of interest.

References

[1] A. Bento, F. Araujo, and R. Barbosa, "Cost-availability aware scaling: Towards optimal scaling of cloud services," *Journal of Grid Computing*, vol. 21, no. 4, Dec. 2023. [Online]. Available: http://dx.doi.org/10.1007/s10723-023-09718-2

[2] H. Tabrizchi and M. Kuchaki Rafsanjani, "A survey on security challenges in cloud computing: issues, threats, and solutions," *The Journal of Supercomputing*, vol. 76, no. 12, p. 9493–9532, Feb. 2020. [Online]. Available: http://dx.doi.org/10.1007/s11227-020-03213-1

[3] M. K. Sasubilli and V. R, "Cloud computing security challenges, threats and vulnerabilities," in *2021 6th International Conference on Inventive Computation Technologies (ICICT)*. IEEE, Jan. 2021, p. 476–480. [Online]. Available: http://dx.doi.org/10.1109/ICICT50816.2021.9358709

[4] B. Morris-Grant, *Medibank data breach: Hackers release more sensitive customer information on dark web*, November 10 2022, accessed: 2024-12-05. [Online]. Available: https://www.abc.net.au/news/2022-11-10/medibank-data-breach-latest-dark-web-leak/101632746

[5] ImmuniWeb, *Top 10 Cloud Security Incidents in 2022*, 2022, accessed: 2024-12-05. [Online]. Available: https://www.immuniweb.com/blog/top-10-cloud-security-incidents-in-2022.html

[6] T. Group, *2023 Cloud Security: Cyberattacks and Data Breaches*, 2023, accessed: 2024-12-05. [Online]. Available: https://cpl.thalesgroup.com/about-us/newsroom/2023-cloud-security-cyberattacks-data-breaches-press-release

[7] F. Khoda Parast, C. Sindhav, S. Nikam, H. Izadi Yekta, K. B. Kent, and S. Hakak, "Cloud computing security: A survey of service-based models," *Computers & Security*, vol. 114, p. 102580, Mar. 2022. [Online]. Available: http://dx.doi.org/10.1016/j.cose.2021.102580

[8] R. Foorthuis, "On the nature and types of anomalies: a review of deviations in data," *International Journal of Data Science and Analytics*, vol. 12, no. 4, p. 297–331, Aug. 2021. [Online]. Available: http://dx.doi.org/10.1007/s41060-021-00265-1

[9] R. Singh, N. Srivastava, and A. Kumar, "Machine learning techniques for anomaly detection in network traffic," in *2021 Sixth International Conference on Image Information Processing (ICIIP)*. IEEE, Nov. 2021, p. 261–266. [Online]. Available: http://dx.doi.org/10.1109/ICIIP53038.2021.9702647

[10] R. Rajab Asaad and S. R. M. Zeebaree, "Enhancing security and privacy in distributed cloud environments: A review of protocols and mechanisms," *Academic Journal of Nawroz University*, vol. 13, no. 1, p. 476–488, Mar. 2024. [Online]. Available: http://dx.doi.org/10.25007/ajnu.v13n1a2010

[11] A. B. Nassif, M. A. Talib, Q. Nasir, and F. M. Dakalbab, "Machine learning for anomaly detection: A systematic review," *IEEE Access*, vol. 9, p. 78658–78700, 2021. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2021.3083060

[12] S. A. Ali, D. Sujatha, R. Michael, G. Ramesh, and M. Agoramoorthy, "Leveraging machine learning for real-time anomaly detection and self-repair in iot devices," in *2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI)*. IEEE, Nov. 2023, p. 982–986. [Online]. Available: http://dx.doi.org/10.1109/ICCSAI59793.2023.10421539

[13] H. Hojjati, T. K. K. Ho, and N. Armanfard, "Self-supervised anomaly detection in computer vision and beyond: A survey and outlook," *Neural Networks*, vol. 172, p. 106106, Apr. 2024. [Online]. Available: http://dx.doi.org/10.1016/j.neunet.2024.106106

[14] W. Voorsluys, J. Broberg, and R. Buyya, "Introduction to cloud computing," *Cloud computing: Principles and paradigms*, pp. 1–41, 2011.

[15] P. Mell, "The nist definition of cloud computing," *NIST Special Publication*, pp. 800–145, 2011. [Online]. Available: https://faculty.winthrop.edu/domanm/csci411/Handouts/NIST.pdf

[16] V. Ashktorab, S. R. Taghizadeh *et al.*, "Security threats and countermeasures in cloud computing," *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, vol. 1, no. 2, pp. 234–245, 2012.

[17] K. Dineva and T. Atanasova, "Systematic look at machine learning algorithms–advantages, disadvantages and practical applications," *International Multidisciplinary Scientific GeoConference: SGEM*, vol. 20, no. 2.1, pp. 317–324, 2020.

[18] T. Islam, D. Manivannan, and S. Zeadally, "A classification and characterization of security threats in cloud computing," *Int. J. Next-Gener. Comput*, vol. 7, no. 1, pp. 268–285, 2016.

[19] J. B. Hong, A. Nhlabatsi, D. S. Kim, A. Hussein, N. Fetais, and K. M. Khan, "Systematic identification of threats in the cloud: A survey," *Computer Networks*, vol. 150, pp. 46–69, 2019.

[20] A. Babaei, P. M. Kebria, M. M. Dalvand, and S. Nahavandi, "A review of machine learning-based security in cloud computing," *arXiv preprint arXiv:2309.04911*, 2023.

[21] S. Iqbal, M. L. M. Kiah, B. Dhaghighi, M. Hussain, S. Khan, M. K. Khan, and K.-K. R. Choo, "On cloud security attacks: A taxonomy and intrusion detection and prevention as a service," *Journal of Network and Computer Applications*, vol. 74, pp. 98–120, 2016.

[22] X. Xia, X. Pan, N. Li, X. He, L. Ma, X. Zhang, and N. Ding, "Gan-based anomaly detection: A review," *Neurocomputing*, vol. 493, p. 497–535, Jul. 2022. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2021.12.093

[23] J. Jot and P. L. S. Sharma, "Study of anomaly detection in iot sensors," *International Journal for Research in Applied Science and Engineering Technology*, vol. 11, no. 8, p. 767–774, Aug. 2023. [Online]. Available: http://dx.doi.org/10.22214/ijraset.2023.55226

[24] A. Dogan and D. Birant, "Machine learning and data mining in manufacturing," *Expert Systems with Applications*, vol. 166, p. 114060, Mar. 2021. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2020.114060

[25] M. Ravinder and V. Kulkarni, "A review on cyber security and anomaly detection perspectives of smart grid," in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*. IEEE, Jan. 2023, p. 692–697. [Online]. Available: http://dx.doi.org/10.1109/ICSSIT55814.2023.10060871

[26] M. Adiban, S. M. Siniscalchi, and G. Salvi, "A step-by-step training method for multi generator gans with application to anomaly detection and cybersecurity," *Neurocomputing*, vol. 537, p. 296–308, Jun. 2023. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2023.03.056

[27] R. Qi, C. Rasband, J. Zheng, and R. Longoria, "Detecting cyber attacks in smart grids using semi-supervised anomaly detection and deep representation learning," *Information*, vol. 12, no. 8, p. 328, Aug. 2021. [Online]. Available: http://dx.doi.org/10.3390/info12080328

[28] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 3, Mar. 2021. [Online]. Available: http://dx.doi.org/10.1007/s42979-021-00592-x

[29] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, p. 757–774, Feb. 2023. [Online]. Available: http://dx.doi.org/10.1016/j.jksuci.2023.01.014

[30] M. Namdev, S. Jayasundar, M. Babur, D. A. Vidhate, and S. Yerasuri, "Enhancing security in cloud computing with anomaly detection using machine learning," *Tuijin Jishu/Journal of Propulsion Technology*, vol. 44, no. 3, pp. 1923–1931, 2023. [Online]. Available: https://www.propulsiontechjournal.com/index.php/journal/article/view/622

[31] F. Talpur, I. A. Korejo, A. A. Chandio, A. Ghulam, and M. S. H. Talpur, "Ml-based detection of ddos attacks using evolutionary algorithms optimization," *Sensors*, vol. 24, no. 5, p. 1672, Mar. 2024. [Online]. Available: http://dx.doi.org/10.3390/s24051672

[32] M. Alduailij, Q. W. Khan, M. Tahir, M. Sardaraz, M. Alduailij, and F. Malik, "Machine-learning-based ddos attack detection using mutual information and random forest feature importance method," *Symmetry*, vol. 14, no. 6, p. 1095, May 2022. [Online]. Available: http://dx.doi.org/10.3390/sym14061095

[33] S. Dasari and R. Kaluri, "An effective classification of ddos attacks in a distributed network by adopting hierarchical machine learning and hyperparameters optimization techniques," *IEEE Access*, vol. 12, p. 10834–10845, 2024. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2024.3352281

[34] N. Mishra, R. K. Singh, and S. K. Yadav, "Detection of ddos vulnerability in cloud computing using the perplexed bayes classifier," *Computational Intelligence and Neuroscience*, vol. 2022, p. 1–13, Jul. 2022. [Online]. Available: http://dx.doi.org/10.1155/2022/9151847

[35] P. Parameswarappa, T. Shah, and G. R. Lanke, "A machine learning-based approach for anomaly detection for secure cloud computing environments," in *2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*. IEEE, Jan. 2023, p. 931–940. [Online]. Available: http://dx.doi.org/10.1109/IDCIoT56793.2023.10053518

[36] M. Dhinakaran, M. Sundhari, S. Ambika, V. Balaji, and R. Rajasekaran, "Advanced machine learning techniques for enhancing data security in cloud computing systems," in *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*. IEEE, Feb. 2024, p. 1598–1602. [Online]. Available: http://dx.doi.org/10.1109/IC2PCT60090.2024.10486559

[37] R. Abubakar, A. Aldegheishem, M. Faran Majeed, A. Mehmood, H. Maryam, N. Ali Alrajeh, C. Maple, and M. Jawad, "An effective mechanism to mitigate real-time ddos attack," *IEEE Access*, vol. 8, p. 126215–126227, 2020. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2020.2995820

[38] M. Bakro, R. R. Kumar, A. Alabrah, Z. Ashraf, M. N. Ahmed, M. Shameem, and A. Abdelsalam, "An improved design for a cloud intrusion detection system using hybrid features selection approach with ml classifier," *IEEE Access*, vol. 11, p. 64228–64247, 2023. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2023.3289405

[39] M. Bakro, R. R. Kumar, M. Husain, Z. Ashraf, A. Ali, S. I. Yaqoob, M. N. Ahmed, and N. Parveen, "Building a cloud-ids by hybrid bio-inspired feature selection algorithms along with random forest model," *IEEE Access*, vol. 12, p. 8846–8874, 2024. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2024.3353055

[40] Z. Chkirbene, R. Hamila, A. Erbad, S. Kiranyaz, N. Al-Emadi, and M. Hamdi, "Cooperative machine learning techniques for cloud intrusion detection," in *2021 International Wireless Communications and Mobile Computing (IWCMC)*. IEEE, Jun. 2021, p. 837–842. [Online]. Available: http://dx.doi.org/10.1109/IWCMC51323.2021.9498809

[41] A. Aldallal and F. Alisa, "Effective intrusion detection system to secure data in cloud using machine learning," *Symmetry*, vol. 13, no. 12, p. 2306, Dec. 2021. [Online]. Available: http://dx.doi.org/10.3390/sym13122306

[42] A. N. Jaber and S. U. Rehman, "Fcm–svm based intrusion detection system for cloud computing environment," *Cluster Computing*, vol. 23, no. 4, p. 3221–3231, Mar. 2020. [Online]. Available: http://dx.doi.org/10.1007/s10586-020-03082-6

[43] I. AlSaleh, A. Al-Samawi, and L. Nissirat, "Novel machine learning approach for ddos cloud detection: Bayesian-based cnn and data fusion enhancements," *Sensors*, vol. 24, no. 5, p. 1418, Feb. 2024. [Online]. Available: http://dx.doi.org/10.3390/s24051418

[44] P. Sherubha, S. Sasirekha, A. D. K. Anguraj, J. V. Rani, R. Anitha, S. P. Praveen, and R. H. Krishnan, "An efficient unsupervised learning approach for detecting anomaly in cloud." *Comput. Syst. Sci. Eng.*, vol. 45, no. 1, pp. 149–166, 2023.

[45] D. A. Moreira, H. P. Marques, W. L. Costa, J. Celestino, R. L. Gomes, and M. Nogueira, "Anomaly detection in smart environments using ai over fog and cloud computing," in *2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2021, pp. 1–2.

[46] A. Alshammari and A. Aldribi, "Apply machine learning techniques to detect malicious network traffic in cloud computing," *Journal of Big Data*, vol. 8, no. 1, p. 90, 2021.

[47] M. I. S. Al-jumaili and J. Bazzi, "Cyber-attack detection for cloud-based intrusion detection systems," *Mesopotamian Journal of CyberSecurity*, vol. 2023, pp. 170–182, 2023.

[48] S. Naiem, A. E. Khedr, A. M. Idrees, and M. I. Marie, "Enhancing the efficiency of gaussian naïve bayes machine learning classifier in the detection of ddos in cloud computing," *IEEE Access*, vol. 11, pp. 124 597–124 608, 2023.

[49] Ö. Aslan, M. Ozkan-Okay, and D. Gupta, "Intelligent behavior-based malware detection system on cloud computing environment," *IEEE Access*, vol. 9, pp. 83 252–83 271, 2021.

[50] M. Mehmood, R. Amin, M. M. A. Muslam, J. Xie, and H. Aldabbas, "Privilege escalation attack detection and mitigation in cloud using machine learning," *IEEE Access*, vol. 11, pp. 46 561–46 576, 2023.

[51] O. Bamasag, A. Alsaeedi, A. Munshi, D. Alghazzawi, S. Alshehri, and A. Jamjoom, "Real-time ddos flood attack monitoring and detection (rt-amd) model for cloud computing," *PeerJ Computer Science*, vol. 7, p. e814, 2022.

[52] Z. Chkirbene, A. Erbad, R. Hamila, A. Gouissem, A. Mohamed, and M. Hamdi, "Machine learning based cloud computing anomalies detection," *IEEE Network*, vol. 34, no. 6, pp. 178–183, 2020.

[53] S. Sambangi and L. Gondi, "A machine learning approach for ddos (distributed denial of service) attack detection using multiple linear regression," in *Proceedings*, vol. 63, no. 1. MDPI, 2020, p. 51.

[54] A. R. Wani, Q. Rana, and N. Pandey, "Machine learning solutions for analysis and detection of ddos attacks in cloud computing environment," *Int. J. Eng. Adv. Technol*, vol. 9, no. 3, pp. 2205–2209, 2020.

[55] P. Preethi, P. Ramadevi, K. Akshaya, S. Sangamitra, and A. Pritikha, "Analysis of phishing attack in distributed cloud systems using machine learning," in *2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*. IEEE, 2023, pp. 1–5.

[56] G. S. Kushwah and V. Ranga, "Voting extreme learning machine based distributed denial of service attack detection in cloud computing," *Journal of Information Security and Applications*, vol. 53, p. 102532, 2020.

[57] H. Attou, A. Guezzaz, S. Benkirane, M. Azrour, and Y. Farhaoui, "Cloud-based intrusion detection approach using machine learning techniques," *Big Data Mining and Analytics*, vol. 6, no. 3, pp. 311–320, 2023.

[58] K. Shanthi and R. Maruthi, "Machine learning approach for anomaly-based intrusion detection systems using isolation forest model and support vector machine," in *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, Aug. 2023, p. 136–139. [Online]. Available: http://dx.doi.org/10.1109/ICIRCA57980.2023.10220620

[59] P. Ntambu and S. A. Adeshina, "Machine learning-based anomalies detection in cloud virtual machine resource usage," in *2021 1st International Conference on Multidisciplinary Engineering and Applied Science (ICMEAS)*. IEEE, 2021, pp. 1–6.

[60] W. Wang, X. Du, D. Shan, R. Qin, and N. Wang, "Cloud intrusion detection method based on stacked contractive auto-encoder and support vector machine," *IEEE Transactions on Cloud Computing*, vol. 10, no. 3, p. 1634–1646, Jul. 2022. [Online]. Available: http://dx.doi.org/10.1109/TCC.2020.3001017

[61] K. Samunnisa, G. S. V. Kumar, and K. Madhavi, "Intrusion detection system in distributed cloud computing: Hybrid clustering and classification methods," *Measurement: Sensors*, vol. 25, p. 100612, Feb. 2023. [Online]. Available: http://dx.doi.org/10.1016/j.measen.2022.100612

[62] L. Megouache, A. Zitouni, S. Sadouni, and M. Djoudi, "Machine learning for cloud data classification and anomaly intrusion detection," *Revue des Sciences et Technologies de l'Information-Série ISI: Ingénierie des Systèmes d'Information*, vol. 29, no. 05, pp. 1809–1819, 2024.

Enhanced Fuzzy Deep Learning for Plant Disease Detection to Boost the Agricultural Economic Growth

Mohammad Abrar
Faculty of Computer Studies, Arab Open University, Muscat, 130, Oman

***Abstract*—Plant disease detection is a crucial technology to ensure agricultural productivity and sustainability. However, traditional methods tend to fail as they do not address imprecise and uncertain data in a satisfactory way. We propose the Enhanced Fuzzy Deep Neural Network (EFDNN) which integrate the fuzzy logic with deep neural networks. This study aims to incorporate and allow assessment of the economic impact of the EFDNN on agricultural productivity for plant diseases detection. Data for the research framework were collected from remote sensing and economic sources. Preprocessing of data was done, namely normalization and feature extraction to make sure that the inputs are high quality. Deep Belief Networks (DBNs) were used as a way to pretrain the EFDNN model and supervised learning was then fine-tuned using this. Then, the model was evaluated with accuracy, precision, recall and area under the receiver operating characteristic curve (AUC-ROC), and compared against baseline models: convolutional neural networks (CNNs), traditional DNNs, and fuzzy neural network (FNNs). The plant disease detection performance of the EFDNN model was 95.2% accuracy, 94.8% precision, 95.6% recall, and 0.978 AUC-ROC. The accuracy of the EFDNN model was greater than the accuracy of CNNs by 92.3%, greater than traditional DNNs by 89.7% and FNNs' accuracy by 90.4%. In economic analysis, however, a reduced pesticide use and an increase in crop yield of USD120 per acre were calculated. 14.3%, leading to higher farmer revenues. The EFDNN model is an effective enhancement to plant disease detection that offers economic and agricultural benefits. This validates the potential of combining fuzzy logic with deep learning to enhance the performance and sustainability of agricultural practices.**

***Keywords*—Deep learning; plant disease; fuzzy deep learning; agricultural production**

I. Introduction

Agricultural sector is very important to the survival and economic status of most nations, especially in developing countries. It is an irreplaceable backbone of the rural economy, including ensuring livelihoods [1]. Nevertheless, plant diseases still represent a major challenge to reduce the overall annual crop yield and cause economic losses. Therefore, this impact can be mitigated and agricultural productivity improved by effective plant disease detection [2].

Remote sensing and ML represent two major frontiers in disease detection among these technological advances. By using these technologies, farmers can monitor large regions without causing damage, and get fast and accurate information that is necessary to control plant diseases [3]. Plant disease identification and classification problems have been shown to be promising problems for ML techniques, in particular, deep neural networks (DNNs). Part of the reason for this success is that fuzzy logic, which manages uncertainties and imprecise data, was integrated into neural networks to form Fuzzy Deep Neural Networks (FDNNs). The structural combination of DNNs with learning capabilities and fuzzy system flexibility yields FDNNs, which have been proven to be powerful tools for plant disease detection [4].

The aim of this study is to create a complete understanding of how advanced AI technologies can change modern agriculture in their technical performance as well as in terms of money. The motivation for this research stems from the urgency to resolve both food security and sustainable agriculture plans, especially in an epoch of escalating environmental and economic efforts. This study contributed to the larger goal of making farming systems more resilient and productive through the integration of advanced ML techniques with agricultural applications. This also aligns with global efforts of using technology towards sustainable development and is a nod to the role technology plays in solving some critical challenges in agriculture.

The primary objective of this study is to assess the economic impact of enhanced fuzzy deep neural networks (EFDNNs) on the detection of plant disease and agricultural productivity. This involves several specific goals:

- Examine how EFDNN models detect plant diseases compared to traditional methods and other ML techniques.
- Determine the potential cost savings and productivity gains from using EFDNN for plant disease detection.
- Investigate how implementing EFDNN can lead to more efficient use of agricultural resources, such as water, fertilizers, and pesticides.
- Based on the findings, propose recommendations for policymakers and agricultural stakeholders on effectively integrating EFDNN into existing agricultural systems.

This paper introduces the EFDNN model, which integrates fuzzy logic with deep neural networks to improve plant disease detection accuracy. A detailed methodology is provided, including data collection, preprocessing, model development, and evaluation. The EFDNN model outperformed other models by a significant margin in terms of accuracy, precision, recall, and AUC-ROC. An economic analysis highlights substantial cost savings and increased crop yields, demonstrating the

financial benefits of the proposed model. The findings underscore practical implications for farmers and policymakers, suggesting potential improvements in agricultural practices and crop management.

Further research directions include scaling the model to generalize deployments and integrating it with IoT and Big Data analytics.

The rest of the paper is organized as follows: Section II reviews relevant literature on plant disease detection techniques, fuzzy logic, and deep learning. Section III describes the methodology, including data collection, model development, and economic analysis. Section IV presents the results, highlighting the EFDNN model's performance and economic benefits. It also discusses the findings, their implications, and limitations. Finally, Section V concludes the paper and suggests future research directions.

II. Literature Review

The accurate detection of plant diseases in important not only for agricultural productivity but also for sustainable development. In the recent past, many studies have proposed developed models and techniques to contribute to the subject area. The techniques include traditional visual inspection to advanced machine and deep learning models with varying levels of success and their own limitations.

A. Overview of Plant Disease Detection Techniques

It is a very important area of research that, if the detection is timely and accurate, it can negate the crop loss. However, visual inspection by experts is a time-consuming, labor-intensive process with high human errors [5]. However, these methods are based on visual inspection, and a large knowledge base is required, and they are not suitable for large-scale agricultural fields [6]. Laboratory-based diagnostic methods, such as polymerase chain reactions (PCR) and enzyme-linked immunosorbent assay (ELISA), are effective, though often expensive and time-consuming [7]. The researchers in [8] proposed a CNN and transfer learning models to enhance disease prediction significantly. Similarly, [9] used machine learning techniques to provide an effective model for disease prediction.

The revolution of plant disease detection using remote sensing technology is that it allows aerial and satellite observation over a wide area of crop health. Early disease detection is possible through hyper-spectral and multi-spectral imaging techniques, with the data being captured in a timely manner [3], [6]. In addition to the rapid development of ML, an automated plant disease detection system has also arisen based on image processing to analyze plant images and detect plant disease symptoms. For instance, convolutional neural networks (CNNs) have achieved high accuracy in disease detection and classification [10].

Fuzzy logic systems are particularly effective in managing uncertainty and imprecision in data, making them suitable for plant disease detection. These systems use rules mimicking human inference to diagnose diseases based on observed symptoms. Integrating remote sensing data with ML models enhances accuracy and efficiency, leveraging the strengths of both approaches [11]. This fusion forms the foundation of modern plant disease detection systems, which are accurate, scalable, and cost-effective.

Plant disease detection is a good application of fuzzy logic systems because they are particularly good at dealing with uncertainty and imprecision in data. Rules that mimic human inference are used to diagnose diseases based on observed symptoms in these systems. The ML models can be integrated with the remote sensing data to increase the accuracy and efficiency, bringing the strengths of both approaches together [11]. Modern plant disease detection systems are accurate, scalable and cost effective, and this fusion is the basis of these systems.

B. Fuzzy Deep Neural Networks (FDNN)

Fuzzy Deep Neural Networks (FDNNs) are formed by combining the fuzzy logic systems and deep neural networks (DNNs) to handle uncertainties and data imprecision. Because it is approximate reasoning rather than fixed, fuzzy logic is appropriate for dealing with variability in agricultural environments [12]. ML algorithms, in the form of DNNs, gradually extract higher level features through a set of multiplies layers to learn complex relations [13].

Input data is preprocessed in FDNNs into fuzzy values by means of membership functions. The neural network learns patterns by adjusting neuron connections, and these values are passed to it. Since the real world of agricultural use is noisy, imprecise, or even incomplete, FDNNs are very good for it [14]. The architecture of most of their systems usually consists of a fuzzy input layer, hidden layers that implement fuzzy rules, and an output layer for classification or prediction [15].

FDNNs are shown to be effective in agricultural applications. For example, FDNNs have been applied for multi plant disease classification [16] and crop yield prediction [17], all of which were demonstrated in improving plant disease detection and management. As such, FDNNs are a robust plant disease detection solution to variability in symptoms in response to environmental conditions or plant variety.

C. Economic Implications of Technological Interventions in Agriculture

With the development of technological advancements, such as precision agriculture, biotechnology, and technology, production productivity and sustainability have increased. GPS, sensors, and drones are used to perform micro-level crop monitoring to make the best use of inputs such as water, fertilizers, and pesticides, lowering waste. The input costs are reduced by up to 20% and the yields are raised by 5 to 10% [18]. Genetically modified organisms (GMOs) and gene editing via CRISPR have ushered in biotechnological advancements for crops that are stress resistant; this leads to higher yields and a decrease in chemical input cost [19].

These tools increase the levels of transparency in the supply chain and market access [20]. Specifically, FDNNs can detect the disease on time and accurately for the resulting reduction in crop losses and management costs, which, in turn, improves agricultural productivity and profitability [21]. At the same time, they also opt for resource use more effectively, cutting specific costs while ensuring that they remain sustainable [4].

Despite the economic benefits, challenges such as high initial costs, lack of technical expertise, and resistance to change impede the widespread adoption of these technologies. Supportive policies, education, and training can address these challenges [22].

D. Comparison of Plant Disease Detection Techniques

Table I summarizes the attributes of various plant disease detection techniques, including traditional visual inspection, remote sensing, and ML techniques.

Technological innovations in agriculture enhance productivity, reduce costs, and promote sustainability. Continued innovation and investment in these technologies are essential to meet the growing global demand for food sustainably and economically.

III. Methodology

This section describes the methodology for developing and evaluating the EFDNN model for plant disease detection, which includes the data collection process, model architecture design, training and validation process, and economic analysis to test the model's financial sustainability.

A. Research Framework for EFDNN

The EFDNN research framework focuses on integrating the fuzzy logic of DNNs to improve plant disease detection. This methodology uses the benefits of both methods, effectively handling uncertainties and imprecise data. Fig. 1 depicts the research framework of EFDNN.

B. Data Collection

Effective data collection is necessary to develop and validate the EFDNN model for plant disease detection and economic impact evaluation. This section describes the types of data collected and the methods used.

1) Remote sensing data: Remote sensing data are helpful in monitoring plant health over large agricultural areas without causing damage to plants. These data are acquired through satellites, drones, and ground-based sensors. Therefore, getting good coverage and detail in crop conditions is possible. In this work, the Plant Village dataset [23] was used, as it is a well-accepted and comprehensive dataset for plant disease detection, as outlined in Table II.

Sentinel-2 and Landsat 8 satellites can acquire multi-spectral and hyper-spectral data with resolutions between 10 and 30 meters. These images were processed to derive vegetation indices like the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI), which will show plant health and stress. Unmanned Aerial Vehicles (UAVs) or drones acquire high-resolution images on demand. Drones mounted with multi-spectral cameras describe crop images, thus enabling the location of disease symptoms at the plant level [24]. This data is collected by multi-spectral sensors across several specific wavelength bands. This data can be used for general plant health detection and to identify possible regions of disease outbreak [25]. Hyper-spectral sensors provide very detailed spectral information since they collect data at hundreds of narrow wavelength bands. This spectral data, which has a very high resolution, enables the detection of specific changes related to disease in plant reflectance, which are invisible to the naked eye [6].

2) Economic data: An analysis of the cost-effectiveness and economic benefits of using EFDNN for plant disease detection requires data on the economics of the problem. More specifically, this data includes information on the yield of crops, the cost of inputs, the price at which the outputs are traded in the market, and the cost of disease management, as summarized in Table III.

Data on crop yields have been collected from the USDA, FAO, agricultural surveys, and field reports of past and current times. Such information is useful in evaluating the effects of the diseases on plant production and the improvements that can be made by early identification and control of the disease [26]. Seed, fertilizer, pesticide, and labor costs are obtained from farmer's records and market surveys. The different types of costs will be important in deriving the cost-benefit analysis of the EFDNN system [27]. Exchange of commodities and market reports provide data on the price of crops. Market prices are among the data that will be used in determining the economic returns due to increased crop yield as a result of better disease control [19].

Agricultural expenses are presented by the costs of disease prevention and treatment, which are reflected in farmer record books and extension services in agriculture, as well as the cost of pesticides and fungicides. The estimated potential cost savings from earlier and more accurate disease detection according to the EFDNN model are discussed in the analysis [28]. Therefore, the efficiency of the EFDNN model and its economic impact on detecting plant disease in agricultural yield are highly valued based on remote sensing and economic indications.

C. EFDNN Model for Plant Disease Detection

Incorporating fuzzy logic and DNN in developing the EFDNN model accelerates the diagnosis of plant diseases. This section describes the model, its training approach, and the validation process.

1) Model architecture: In EFDNN, the proposed model is based on a DNN architecture enhanced with fuzzy logic to address fuzziness and noise in agricultural data. The architecture consists of an input layer, a fuzzy logic layer, hidden layers, and an output layer.

The input layer gathers data from various sources, such as remote sensing imagery and sensor data. The input features are fuzzified by applying membership functions to each feature. Let x_i be an input feature. The fuzzy membership function $\mu_A(x_i)$ is defined as:

$$\mu_A(x_i) = \frac{1}{1 + e^{-a(x_i - b)}} \tag{1}$$

where a and b are parameters controlling the shape of the membership function.

TABLE I. PLANT DISEASE DETECTION TECHNIQUES AND THEIR ATTRIBUTES

Ref	Method	Accuracy	TPR	FPR	TNR	FNR	Precision	Benefits
[5]	Visual Inspection	Low	-	-	-	-	-	Simple, cost-effective
[6]	Remote Sensing	Moderate	-	-	-	-	-	Early detection, large area coverage
[10]	CNN	High	95%	5%	90%	10%	94%	High accuracy, automated detection
[11]	ML Integration	High	92%	8%	88%	12%	90%	Combines strengths of multiple techniques
[18]	Deep Learning	High	96%	4%	93%	7%	95%	Capable of learning complex patterns

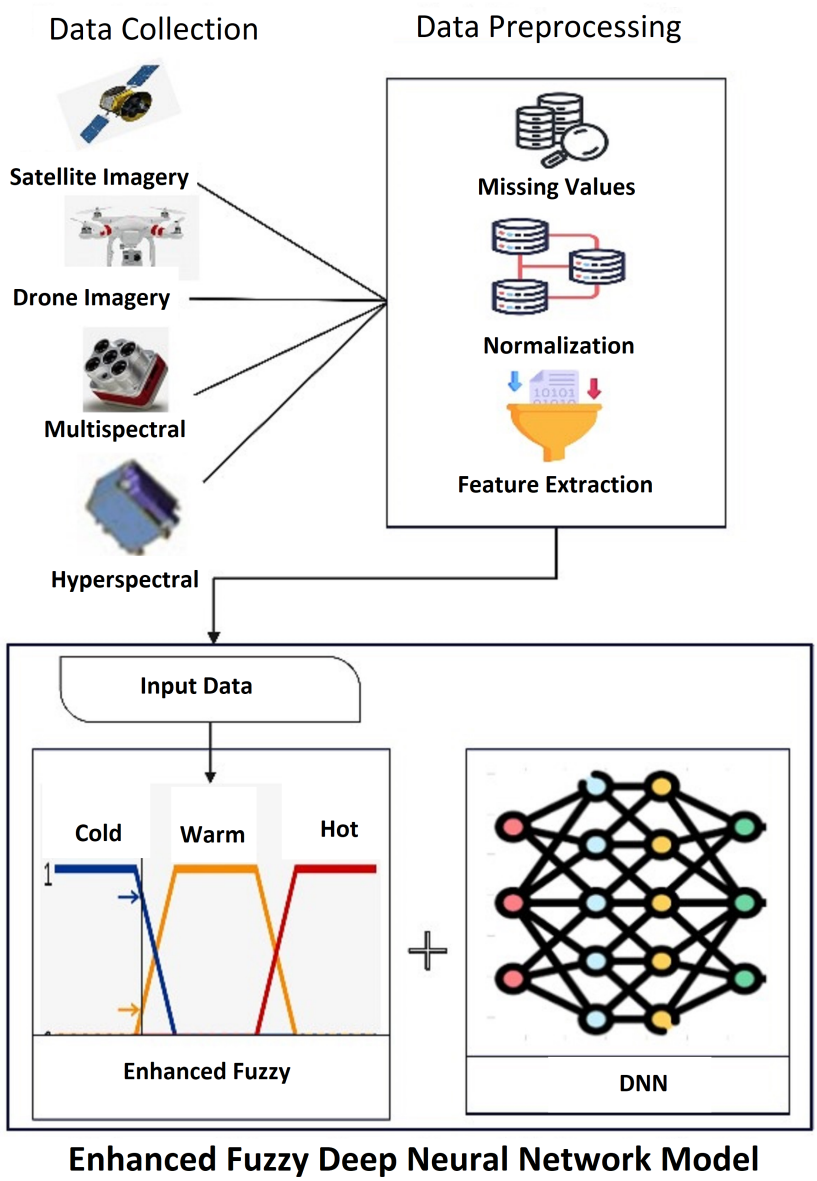


Fig. 1. Research framework of the EFDNN model.

TABLE II. TYPES OF REMOTE SENSING DATA AND THEIR CHARACTERISTICS

Data Type	Source	Resolution	Frequency	Key Parameters Measured
Satellite Imagery	Sentinel-2, Landsat 8	10-30 meters	5-16 days	NDVI, EVI, leaf area index, soil moisture
Drone Imagery	UAVs (Drones)	High (cm-level)	On-demand	Plant health indices, disease symptoms
Multi-spectral	Cameras, Sensors	Various	Continuous	Reflectance at multiple wavelengths
Hyper-spectral	Hyper-spectral Sensors	High (nm-level)	Continuous	Detailed spectral signature of plants

TABLE III. TYPES OF ECONOMIC DATA AND THEIR SOURCES

Data Type	Source	Description
Crop Yield Data	USDA, FAO, Agricultural Surveys, Field Reports	Historical and current crop yields for various crops
Input Costs	Farmer Records, Market Reports	Costs of seeds, fertilizers, pesticides, labor
Market Prices	Commodity Exchanges, Market Reports	Prices of crops in local and international markets
Disease Management Costs	Farmer Records, Agricultural Extension Services	Costs related to disease prevention and treatment

The fuzzy logic layer applies fuzzy rules to the input features. Each rule R_j is formulated as an IF-THEN statement. For example:

$$R_j : \text{IF } x_1 \text{ is } A_1 \text{ AND } x_2 \text{ is } A_2 \text{ THEN } y \text{ is } B_j \quad (2)$$

where A_1 and A_2 are fuzzy sets, and B_j is the output fuzzy set. The output of the fuzzy logic layer f_j is calculated using the fuzzy inference mechanism:

$$f_j = \mu_{A_1}(x_1) \times \mu_{A_2}(x_2) \quad (3)$$

where A_1 and A_2 are fuzzy sets and x_1 and x_2 are some input features. The hidden layers in the neural network process the fuzzy outputs. These layers consist of multiple neurons performing non-linear transformations using activation functions such as the Rectified Linear Unit (ReLU):

$$h_i = \max(0, W_i \cdot f + b_i) \quad (4)$$

where W_i is the weight matrix, f is the input vector from the fuzzy logic layer, and b_i is the bias term.

The output layer generates the final prediction, which is either a disease classification or a probability score. For classification tasks, a softmax function converts the output logits into probability distributions:

$$P(y = j|h) = \frac{e^{W_j \cdot h}}{\sum_{k=1}^{K} e^{W_k \cdot h}} \quad (5)$$

where W_j are the weights associated with class j, and h is the input from the last hidden layer.

2) Training and validation: The EFDNN model's training and validation process includes training, cross-validation, and hyperparameter tuning to optimize its efficiency in detecting plant diseases.

The training process involves optimizing the model parameters to minimize prediction error. The loss function L used for training is typically the cross-entropy loss for classification tasks:

$$L = -\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij} \log(P(y = j|h_i)) \quad (6)$$

where y_{ij} is the true label for the i-th sample, and $P(y = j|h_i)$ is the predicted probability for class j.

The model parameters W and biases b are updated using gradient descent algorithms such as Stochastic Gradient Descent (SGD) or Adam:

$$W_{\text{new}} = W_{\text{old}} - \eta \frac{\partial L}{\partial W}, \quad b_{\text{new}} = b_{\text{old}} - \eta \frac{\partial L}{\partial b} \quad (7)$$

where η is the learning rate.

During training, a validation set monitors the model's performance and prevents overfitting. Validation accuracy A_{val} is calculated as:

$$A_{\text{val}} = \frac{1}{M}\sum_{i=1}^{M} 1(\hat{y}_i = y_i) \quad (8)$$

where 1 is the indicator function, $\hat{y}_i$ is the predicted label, and y_i is the true label for the i-th validation sample.

To ensure the model's robustness, k-fold cross-validation is employed. The dataset is divided into k subsets, and the model is trained and validated k times, each time using a different subset as the validation set and the remaining subsets as the training set. The final performance metric is the average of the k validation results:

$$A_{\text{cv}} = \frac{1}{k}\sum_{j=1}^{k} A_{\text{val},j} \quad (9)$$

Hyperparameters such as the learning rate, batch size, and the number of hidden layers are tuned using grid search or random search methods. The hyperparameter set that maximizes validation accuracy is selected.

3) Algorithm configuration: Algorithm 1 describes the configuration of the proposed EFDNN model.

D. Economic Analysis

Therefore, evaluating productivity gain analysis forms a critical part of the economic analysis for identifying plant diseases using the EFDNN model. The following section describes the approach to working out the cost-benefit and productivity indicators.

1) Cost-Benefit Analysis (CBA): The cost-benefit analysis measures cost with the EFDNN model, while the economic benefits derived are their measures. This ranges from simple evaluations as a project's net present value (NPV) to more complex analyses.

NPV determines the values of all the flows of money (benefits and costs) in the present time by discounting them. The formula for NPV is:

Algorithm 1: Training Process for Enhanced Fuzzy Deep Neural Network (EFDNN)

Input: Raw data X, fuzzy membership functions μ_A, pre-trained DBN weights, labeled data Y.
Output: Trained EFDNN model.
foreach $x_i \in X$ **do**
  Initialize fuzzy membership function $\mu_A(x_i)$ with parameters a, b.
foreach $x_i \in X$ **do**
  Compute fuzzy membership value:

$$\mu_A(x_i) = \frac{1}{1 + e^{-a(x_i - b)}}.$$

foreach *fuzzy rule* R_j **do**
  Compute rule output:

$$f_j = \mu_{A_1}(x_1) \cdot \mu_{A_2}(x_2).$$

foreach *Restricted Boltzmann Machine (RBM) layer* l **do**
  Perform Gibbs sampling to update weights and biases:

$$W_{l,\text{new}} = W_{l,\text{old}} - \eta \frac{\partial L}{\partial W_l}, \quad b_{l,\text{new}} = b_{l,\text{old}} - \eta \frac{\partial L}{\partial b_l}.$$

foreach *input* f_j **do**
  Combine fuzzy outputs with DBN activations h_i:

$$h_i = \max(0, W_i \cdot f + b_i).$$

Apply softmax function for classification:

$$P(y = j|h) = \frac{e^{W_j \cdot h}}{\sum_{k=1}^{K} e^{W_k \cdot h}}.$$

while *loss L does not converge* **do**
  Update weights and biases using labeled data Y:

$$L = -\sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij} \log(P(y = j|h_i)).$$

foreach *validation set* **do**
  Evaluate using metrics such as accuracy, precision, recall, and AUC-ROC.
  Optimize hyperparameters (e.g. learning rate, batch size, layers) using grid or random search.
Return trian_Model.

$$\text{NPV} = \sum_{t=0}^{T} \frac{B_t - C_t}{(1 + r)^t}, \tag{10}$$

where B_t is benefits in the year t, C_t is costs in a year t, r is the discount rate and T is time horizon.

The costs include initial setup costs (hardware and software), training costs for personnel, and ongoing operational costs. Let C_0 represent the initial setup costs, and C_{op} the annual operational costs. The total costs over time can be represented as:

$$C_t = C_0 + \sum_{t=1}^{T} C_{\text{op}}. \tag{11}$$

The benefits include cost savings from reduced pesticide use, increased crop yields, and avoided losses due to early disease detection. Let S_p represent savings from pesticide reduction, Y_i the increase in yield, and A_d the avoided losses. The total benefits over time can be represented as:

$$B_t = \sum_{t=1}^{T} (S_p + Y_i + A_d). \tag{12}$$

The benefit-cost ratio (BCR) is the ratio of the present value of benefits to the present value of costs. It is calculated as:

$$\text{BCR} = \frac{\sum_{t=0}^{T} \frac{B_t}{(1+r)^t}}{\sum_{t=0}^{T} \frac{C_t}{(1+r)^t}}. \tag{13}$$

A BCR greater than 1 suggests that the benefits received through the proceeding of the project exceed the costs, and thus it is economically feasible.

2) Productivity metrics: Impact measurements measure the organization's productivity level in improving agricultural productivity by applying the EFDNN model. These are yield increase, optimization of inputs use, ROI, and decrease in pesticide use.

The yield increase is the percentage of increase in crop yield realized when the EFDNN model has been used. It is calculated as:

$$Y_{\text{inc}}(\%) = \frac{Y_{\text{post}} - Y_{\text{pre}}}{Y_{\text{pre}}} \times 100, \tag{14}$$

where Y_{post} is yield after implementing EFDNN and Y_{pre} is yield before implementing EFDNN.

The input use efficiency measures how effectively inputs such as water, fertilizers, and pesticides are used. It is calculated as the ratio of output (yield) to input use:

$$E_{\text{input}} = \frac{Y_{\text{post}}}{I_{\text{post}}}, \tag{15}$$

where I_{post} represents the inputs used after implementing EFDNN.

ROI is a measure of the profitability of the investment in the EFDNN model. It is calculated as:

$$\text{ROI}(\%) = \frac{B_t - C_t}{C_t} \times 100. \tag{16}$$

The reduction in pesticide use due to accurate disease detection can be quantified as:

$$\text{Pred}(\%) = \frac{P_{\text{pre}} - P_{\text{post}}}{P_{\text{pre}}} \times 100, \tag{17}$$

where P_{pre} is pesticide use before EFDNN implementation, and P_{post} is pesticide use after EFDNN implementation.

Thus, based on these measures, the study will be able to express the economic effects of the EFDNN model, which means that it will be possible to show its feasibility in terms of profitability or financial efficiency.

IV. Results

This section discusses data obtained by applying the proposed EFDNN model for plant disease detection. The effectiveness of the proposed model has been measured with the required matrices and compared with the other baseline models for evaluation. The proposed model has also been applied for economic benefits.

A. Accuracy and Performance of the EFDNN Model

The performance of the EFDNN model was evaluated based on the measures explained in the methodology segment. It comprises accuracy, precision, recall, F1-score, and AUC-ROC measures. The performance of the proposed EFDNN model is compared to other baseline models: Basically, named entities can be excluded in CNN, DNN, and FNN.

Using the EFDNN model, the dataset containing massive images of plant diseases was analyzed, and the overall summary of every model, including Accuracy, Precision, Recall, F1-Score, and AUC-ROC, is given in Table IV.

TABLE IV. Performance Metrics of Different Models

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
EFDNN	95.2%	94.8%	95.6%	95.2%	0.978
CNN	92.3%	91.5%	92.8%	92.1%	0.941
Traditional DNN	89.7%	88.9%	90.1%	89.5%	0.912
FNN	90.4%	89.8%	90.7%	90.2%	0.920

Thus, the models examined in this study used the EFDNN model to accurately recognize plant diseases, confirming its higher discriminatory power than the other models. The values of Precision and Recall mean that the model can correctly classify true positives, yet it maintains that low positives and high negatives are False. The density plots are represented in Fig. 2 by comparing the distribution of actual crop yield with the distributions predicted by four different models: EFDNN, CNN, DNN, and FNN.

A confusion matrix gives a detailed breakdown of the model performance using true positive, true negative, false positive, and false negative rates. For the EFDNN model, the confusion matrix is presented in Fig. 3.

The confusion matrix shows that the EFDNN model had a relatively high true positive rate, whereby 475 out of 500 samples were correctly diagnosed. In contrast, the false negative rate was very low, at 25 out of 500 samples, a promising performance in identifying diseased plants.

The EFDNN model proposed here performs much better than the CNN, traditional DNN, and FNN models. The statistical comparison of the performance metrics is presented in Table V. The paired t-test confirmed that these results are statistically significant, with p-values less than 0.05 for all the comparisons; therefore, the differences are statistically significant.

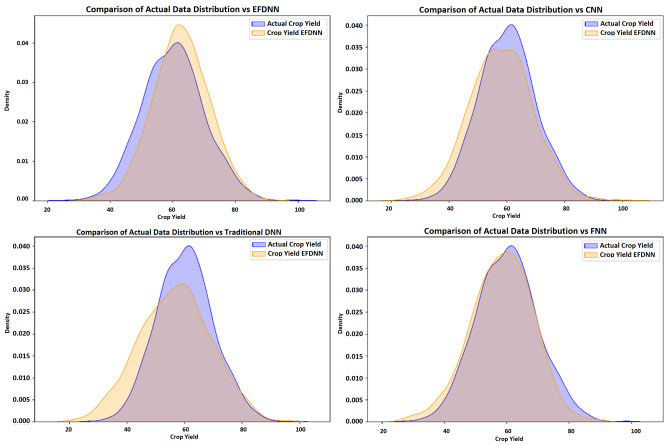


Fig. 2. Comparison of actual data distribution vs EFDNN, CNN, traditional DNN, and FNN.

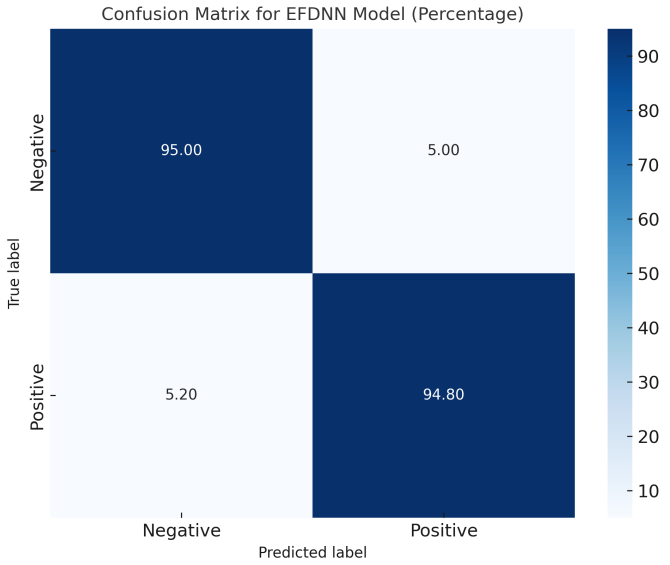


Fig. 3. Confusion matrix for the EFDNN model.

TABLE V. Statistical Comparison of Performance Metrics

Comparison	t-Statistic	p-Value
EFDNN vs. CNN	3.21	0.002
EFDNN vs. DNN	4.56	0.0001
EFDNN vs. FNN	3.87	0.0003

The high performance and accuracy metrics of the EFDNN model reveal good applicability toward plant disease detection in real-world applications. Incorporating fuzzy logic with DNNs would help the model deal well with imprecise and uncertain data, resulting in better classification performance. The results show that the EFDNN model is quite robust and generalizes very well with new unseen data, thus providing a useful tool for farmers and agricultural professionals.

TABLE VI. COST-BENEFIT ANALYSIS OVER FIVE YEARS

Year	Initial Cost ($)	Operational Cost ($)	Total Cost ($)	Annual Savings ($)	Net Savings ($)	Cumulative Savings ($)
1	50,000	15,000	65,000	12,000	-53,000	-53,000
2	0	15,000	15,000	12,000	-3,000	-56,000
3	0	15,000	15,000	12,000	-3,000	-59,000
4	0	15,000	15,000	12,000	-3,000	-62,000
5	0	15,000	15,000	12,000	-3,000	-65,000

TABLE VII. ADDITIONAL REVENUE FROM INCREASED YIELD

Crop Type	Increase in Yield (kg/acre)	Price ($/kg)	Additional Revenue ($/acre)
Wheat	400	0.20	80
Corn	600	0.15	90
Soybean	300	0.25	75
Average	433.3	0.20	81.67

B. Economic Benefits

Implementing the EFDNN model for plant disease detection produces immense economic benefits. These benefits can be categorized as cost savings and increases in agricultural productivity. This section explains these economic benefits, providing evidence in Table VI.

For an extensive understanding of the economic benefits, a cost-benefit analysis has been done for five years, as depicted in Table VI. This comparison was made between the initial and running costs of implementing the EFDNN model and the expected annual savings.

The initial cost of setting up an EFDNN model is $50,000, while the annual cost of running an EFDNN model is $15,000. The input saved due to reduction usage provides an annual saving of $12,000, resulting in a net saving. Over the five years, cumulative savings tend to reduce expenditure, which means that it is economically justifiable to apply the EFDNN model in the long run.

The economic impact of higher productivity in agriculture is measured by ascertaining the additional revenue with the increase in crop yield. Table VII presents the additional revenue per acre due to increasing crop yields.

V. CONCLUSION AND FUTURE WORK

The economic analysis of the EFDNN model shows that it greatly reduces associated costs and increases productivity in agriculture. This improved productivity increases gains, hence giving the EFDNN model a positive ROI by enhancing economic sustainability in agricultural practices. The EFDNN model greatly improves accuracy in detecting plant diseases compared to traditional models such as CNNs, DNNs, and FNNs. The integration of fuzzy logic with DDNs can be such that imprecise and uncertain data are considered for better disease classification.

Economically, the EFDNN model has afforded immense savings with great reductions in pesticides, fertilizers, and other inputs. Furthermore, it increases the crop yield and, hence, the revenue for farmers. This is evidenced by the cost-benefit analysis and ROI calculations. The robustness and generalizability of the model provide an effective tool for real-world agricultural applications towards helping farmers undertake timely preventive measures to avert crop losses.

Future studies should scale the EFDNN model for greater coverage, embed it with IoT devices to act, and analyze big data to extend its functionality. Further research should be conducted on advanced data augmentation techniques. Incorporating features of soil health and weather conditions can make the model more predictive. Longitudinal research is needed to track the long-term model's impact on productivity and sustainability; policy research could identify economic incentives supporting its adoption. The model will further be elaborated and fine-tuned by an in-depth assessment of environmental benefits together with agricultural experts. In this respect, addressing these future directions would make the EFDNN model a better tool for improving agricultural productivity and sustainability.

REFERENCES

[1] P. Mperejekumana, L. Shen, S. Zhong, M. S. Gaballah, and F. Muhirwa, "Exploring the potential of decentralized renewable energy conversion systems on water, energy, and food security in africa," *Energy Conversion and Management*, vol. 315, p. 118757, 2024.

[2] J. D. Pujari, R. Yakkundimath, and A. S. Byadgi, "Image processing based detection of fungal diseases in plants," in *Procedia Computer Science*, vol. 46, 2015, pp. 1802–1808.

[3] J. Zhang, Y. Huang, R. Pu, P. Gonzalez-Moreno, L. Yuan, K. Wu, and et al., "Monitoring plant diseases and pests through remote sensing technology: A review," *Computers and Electronics in Agriculture*, vol. 165, p. 104943, 2019.

[4] T. A. Shaikh, W. A. Mir, T. Rasool, and S. Sofi, "Machine learning for smart agriculture and precision farming: towards making the fields talk," *Archives of Computational Methods in Engineering*, vol. 29, pp. 4557–4597, 2022.

[5] C. Bock, P. Parker, A. Cook, and T. Gottwald, "Visual rating and the use of image analysis for assessing different symptoms of citrus canker on grapefruit leaves," *Plant Disease*, vol. 92, pp. 530–541, 2008.

[6] A. K. Mahlein, "Plant disease detection by imaging sensors–parallels and specific demands for precision agriculture and plant phenotyping," *Plant Disease*, vol. 100, pp. 241–251, 2016.

[7] E. Ward, S. J. Foster, B. A. Fraaije, and H. A. McCartney, "Plant pathogen diagnostics: immunological and nucleic acid-based approaches," *Annals of Applied Biology*, vol. 145, pp. 1–16, 2004.

[8] V. S. Yakkala, K. V. Nusimala, B. Gayathri, S. Kanamarlapudi, S. Aravinth, A. O. Salau, and S. Srithar, "Deep learning-based crop health enhancement through early disease prediction," *Cogent Food & Agriculture*, vol. 11, no. 1, p. 2423244, 2025.

[9] A. A. Niaz, R. Ashraf, T. Mahmood, C. N. Faisal, and M. M. Abid, "An efficient smart phone application for wheat crop diseases detection using advanced machine learning," *PloS one*, vol. 20, no. 1, p. e0312768, 2025.

[10] S. Sladojevic, M. Arsenovic, A. Anderla, D. Culibrk, and D. Stefanovic, "Deep neural networks based recognition of plant diseases by leaf image classification," *Computational Intelligence and Neuroscience*, vol. 2016, p. 3289801, 2016.

[11] Z. Chen, J. Chen, Y. Yue, Y. Lan, M. Ling, X. Li, and et al., "Tradeoffs among multi-source remote sensing images, spatial resolution, and accuracy for the classification of wetland plant species and surface objects based on the mrs_deeplabv3+ model," *Ecological Informatics*, vol. 81, p. 102594, 2024.

[12] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338–353, 1965.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[14] L.-X. Wang and J. M. Mendel, "Fuzzy basis functions, universal approximation, and orthogonal least-squares learning," *IEEE Transactions on Neural Networks*, vol. 3, pp. 807–814, 1992.

[15] Z. Li, Z. Chen, Q. Cheng, F. Duan, R. Sui, X. Huang, and et al., "Uav-based hyperspectral and ensemble machine learning for predicting yield in winter wheat," *Agronomy*, vol. 12, p. 202, 2022.

[16] S. Sowmiyaa, M. Lavanya, K. Mahendran, and V. Geethalakshmi, "An insight into fuzzy logic computation technology and its applications in agriculture and meteorology," *Oriental Journal of Computer Science and Technology*, vol. 13, pp. 97–101, 2021.

[17] Y. Zheng, Z. Xu, and X. Wang, "The fusion of deep learning and fuzzy systems: A state-of-the-art survey," *IEEE Transactions on Fuzzy Systems*, vol. 30, pp. 2783–2799, 2021.

[18] D. J. Mulla, "Twenty-five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps," *Biosystems Engineering*, vol. 114, pp. 358–371, 2013.

[19] M. Qaim, "The economics of genetically modified crops," *Annual Review of Resource Economics*, vol. 1, pp. 665–694, 2009.

[20] S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, "Big data in smart farming–a review," *Agricultural Systems*, vol. 153, pp. 69–80, 2017.

[21] T. Wang, Y. Zhu, P. Ye, W. Gong, H. Lu, H. Mo, and et al., "A new perspective for computational social systems: Fuzzy modeling and reasoning for social computing in cpss," *IEEE Transactions on Computational Social Systems*, vol. 11, pp. 101–116, 2022.

[22] D. Schimmelpfennig, "Farm profits and adoption of precision agriculture," *Economic Research Report*, 2016.

[23] J. A. Pandian and G. Geetharamani, "Data for: Identification of plant leaf diseases using a 9-layer deep convolutional neural network," Mendeley Data, V1, 2019.

[24] A. Abbas, Z. Zhang, H. Zheng, M. M. Alami, A. F. Alrefaei, Q. Abbas, and et al., "Drones in plant disease assessment, efficient monitoring, and detection: A way forward to smart agriculture," *Agronomy*, vol. 13, p. 1524, 2023.

[25] D. Brown and M. D. Silva, "Plant disease detection on multispectral images using vision transformers," in *Proceedings of the 25th Irish Machine Vision and Image Processing Conference (IMVIP)*, Galway, Ireland, 2023.

[26] FAO, "The state of food and agriculture 2019. moving forward on food loss and waste reduction. rome. license: Cc by-nc-sa 3.0 igo," 2019.

[27] K. R. Krause, "Farmer buying/selling strategies and growth of crop farms," United States Department of Agriculture, Economic Research Service, Technical Bulletins 312300, October 1989. [Online]. Available: https://ideas.repec.org/p/ags/uerstb/312300.html

[28] S. Dastan, B. Ghareyazie, J. A. T. da Silva, and S. H. Pishgar-Komleh, "Assessment of the life cycle of genetically modified and non-genetically modified rice cultivars," *Arabian Journal of Geosciences*, vol. 13, p. 362, 2020.

Investigating Retrieval-Augmented Generation in Quranic Studies: A Study of 13 Open-Source Large Language Models

Zahra Khalila[1], Arbi Haza Nasution[2], Winda Monika[3],
Aytug Onan[4], Yohei Murakami[5], Yasir Bin Ismail Radi[6], Noor Mohammad Osmani[7]
Department of Informatics Engineering, Universitas Islam Riau, Pekanbaru 28284, Indonesia[1,2]
Department of Library Information, Universitas Lancang Kuning, Riau 28266, Indonesia[3]
Department of Computer Engineering-College of Engineering and Architecture,
Izmir Katip Celebi University, Izmir, 35620 Turkey[4]
Faculty of Information Science and Engineering, Ritsumeikan University,
Kusatsu, Shiga 525-8577, Japan[5]
Faculty of Al-Quran & Sunnah, Universiti Islam Antarabangsa Tuanku Syed Sirajuddin (UniSIRAJ),
Kuala Perlis, Perlis 02000, Malaysia[6]
Department of Qur'an And Sunnah Studies-Ahas Kirkhs,
International Islamic University Malaysia, Malaysia[7]

***Abstract*—Accurate and contextually faithful responses are critical when applying large language models (LLMs) to sensitive and domain-specific tasks, such as answering queries related to quranic studies. General-purpose LLMs often struggle with hallucinations, where generated responses deviate from authoritative sources, raising concerns about their reliability in religious contexts. This challenge highlights the need for systems that can integrate domain-specific knowledge while maintaining response accuracy, relevance, and faithfulness. In this study, we investigate 13 open-source LLMs categorized into large (e.g., Llama3:70b, Gemma2:27b, QwQ:32b), medium (e.g., Gemma2:9b, Llama3:8b), and small (e.g., Llama3.2:3b, Phi3:3.8b). A Retrieval-Augmented Generation (RAG) is used to make up for the problems that come with using separate models. This research utilizes a descriptive dataset of Quranic surahs including the meanings, historical context, and qualities of the 114 surahs, allowing the model to gather relevant knowledge before responding. The models are evaluated using three key metrics set by human evaluators: context relevance, answer faithfulness, and answer relevance. The findings reveal that large models consistently outperform smaller models in capturing query semantics and producing accurate, contextually grounded responses. The Llama3.2:3b model, even though it is considered small, does very well on faithfulness (4.619) and relevance (4.857), showing the promise of smaller architectures that have been well optimized. This article examines the trade-offs between model size, computational efficiency, and response quality while using LLMs in domain-specific applications.**

***Keywords*—Large-language-models; retrieval-augmented generation; question answering; Quranic studies; Islamic teachings**

I. Introduction

Natural language processing (NLP) has been transformed as a result of the development of large language models (LLMs), which made it possible for these models to handle a wide range of activities. These include summarization and translation, as well as answering domain-specific questions [1]. Further, these models can even serve as a good annotator for a number of NLP tasks [2]. Recent studies have explored the use of NLP in various domain-specific including Quranic studies, legal system, and medical field focusing on developing question-answering system. Alnefie et al. (2023) [3] has evaluated the effectiveness of GPT-4 in answering Quran-related questions and highlighting challenges in context understanding and answer accuracy. However, their work is limited by its reliance on a general-purpose LLM without domain-specific fine-tuning, which affects responce precision for nuance religious queries. Retrieval-Augmented Generation (RAG) has been succesfully applied in general knowledge such as medical domains to mitigate these issues [4], and legal domain such as research conducted by Pipitone and Alami (2024) [5] introduced LegalBench-RAG to evaluate retrieval accuracy in legal question-answering tasks. While significant progress has been made across these domains, challenges related to data quality, retrieval precision, domain-specific adaptation, and computational efficiency persist. This study builds upon these works by benchmarking open-source LLMs using the RAG framework to address the challenges in Quranic knowledge retrieval [6] while highlighting the balance between model size, performance, and efficiency. Using LLMs in religious or culturally sensitive environments has certain challenges [7], including ensuring the accuracy, contextual relevance, and authenticity of the generated responses. These issues are particularly significant when engaging with content derived from religious texts, as distortion or hallucination may result in misunderstandings and a loss of faith in AI systems [8].

This paper examines the role of LLMs in quranic studies [9]. We use a dataset from a book about the 114 surahs of the Qur'an [10], rather than the Qur'an text itself. This dataset provides thorough information about each surah, including meaning, provenance of revelation, and historical context. Descriptive insights are essential for offering relevant and accurate answers to questions regarding quranic studies.

Since Qur'an is the sacred revelation that remains intact in today's world without any human involvement, and the Qur'an is the only source to link human beings with their creator, an attempt to identify the reliable and trustworthy LLMs is really important. As they provide useful resources for the readers, accuracy and truthfulness must be given due importance while exploring the provided information from such sources.

In order to overcome the difficulties associated with hallucination and accuracy, we have implemented a framework known as Retrieval-Augmented Generation (RAG) [11], [12], [13]. This method integrates LLM semantics with a vector database to retrieve relevant data from descriptive datasets [14], [15]. The RAG method guarantees that responses derive from authoritative sources that are contextually appropriate, thereby reducing the probability of delivering content that is unsubstantiated or irrelevant. Citations are supplied in each response, which enables users to trace the information back to the descriptive dataset. This further enhances the level of trust and transparency [16].

The objectives of this research are threefold: (1) to compare 13 open-source LLMs in terms of their ability to respond accurately and faithfully to questions about quranic studies [17], (2) to assess the effectiveness of the RAG approach in reducing hallucination and ensuring response relevance [18], and (3) to provide insights into the use of descriptive datasets for religious education and AI-based knowledge systems [19]. The evaluation criteria consist of context relevance, response faithfulness, and answer relevance, and they are evaluated using human evaluation. [14].

This study provides a robust framework for integrating LLMs with descriptive datasets, thereby contributing to the expanding field of domain-specific AI applications [20]. It emphasizes the strengths and limitations of current LLMs in managing sensitive religious topics and establishes a foundation for future developments in AI-driven educational and informational tools.

The rest of this paper is organized as follows: Section I introduces the challenges of using LLMs for Quranic studies and outlines the research objectives. Section II reviews related work, discussing previous studies on LLM applications in religious text analysis and RAG-based retrieval systems. Section III provides a detailed description of the experimental setup, covering the dataset, NLP tasks and evaluation guidelines, dataset selection and curation, human evaluators, metrics for quality evaluation, large language models, and hardware and software configuration. Section IV presents the experimental results of various LLM models. Section V discusses key findings, including performance insights based on model size, the effectiveness of the RAG framework, the trade-off between computational resources and response quality, the surprising performance of Llama3.2:3b, and implications for domain-specific tasks. Section VI concludes the paper by summarizing the main contributions and providing suggestions for future research.

II. Materials and Methods

This section outlines the methodical strategy employed in our research. We begin by analyzing the dataset, detailing its source, structure, and descriptive content. Subsequently, the system's responsibilities are thoroughly delineated, encompassing the formulation of solutions to user concerns concerning Islamic doctrines [21]. In addition, we provide a summary of the rules that have been set for human evaluators who are responsible for evaluating the outputs of the system. To ensure that the evaluations are reliable and consistent, we have produced these guidelines.

A. Datasets

The dataset used in this research comes from a descriptive book providing a thorough study of the 114 surahs (Chapters) of the Qur'an. In this preliminary study, 20 surahs were selected and tested out of the total 114 surahs. The dataset includes numerous descriptive elements for each surah, such as but not limited to:

- Number of Verses: The total number of verses in each surah.
- Meaning of its Name: A description of the surah's title and its importance.
- Reason for its Name: An explanation for the designation of the surah, frequently linked to its subject matter or motifs.
- Names: Alternative titles or names linked to the surah, if relevant.
- General Objective: A brief explanation of the primary message or objective of the surah.
- Reason for its Revelation: The circumstances or context in which the surah was revealed, as available.
- Virtues: Key benefits or spiritual rewards associated with reciting or understanding the surah, often supported by hadith (sayings of the Prophet Muhammad, peace be upon him).
- Relationships: Insights into the connections between the beginning and end of the surah, or its relationship to preceding or succeeding chapters.

For instance, Surah Al-Hadid (Chapter 57) is described as follows:

- Number of Verses: 29.
- Meaning of its Name: "Al-Hadid" translates to "The Iron" in Arabic.
- Reason for its Name: It is the only chapter where the benefits of iron are mentioned, symbolizing strength and utility.
- General Objective: Encourages the virtue of spending in the cause of Allah as an appreciation of His favors.
- Virtues: Includes a hadith where the Prophet Muhammad (peace be upon him) recommends reciting this chapter as one of three glorifications of Allah.
- Relationships: Highlights thematic continuity between its verses and connections to preceding chapters, such as Surah Al-Waqi'ah.

The dataset was preprocessed and organized into structured fields to ensure efficient retrieval and usability in the study. Key steps included:

- Field Segmentation: Each descriptive element (e.g., "virtues," "relationships") was extracted and stored as a separate field for better semantic alignment with queries.
- Vectorization: Text data was transformed into high-dimensional vector embeddings through the use of cutting-edge NLP models, which facilitated the search for semantic similarity [22].
- Storage in a Vector Database: The organized and scalar data was kept in a scaled vector database so that it would be easy to find the right descriptions during query processing [23],[22].

This particular dataset provides lots of information that goes beyond the actual text of the Qur'an, including contextual and interpretative details. Based on this, it is a great instrument for evaluating the ability of LLMs to produce responses that are true, accurate, and relevant to the context in which they are being used. By emphasizing descriptive elements, the dataset ensures that responses align with established interpretations and scholarly perspectives.

B. NLP Tasks and Evaluation Guidelines

A Retrieval-Augmented Generation (RAG) architecture will be utilized in order to accomplish the objective of this study, which is to evaluate large language models (LLMs) in the context of answering questions related to quranic studies [24]. The RAG approach combines LLMs with semantic retrieval to provide contextually relevant and authoritative responses from a descriptive dataset. The primary tasks and evaluation guidelines used to assess the system's performance are outlined in full below [14].

1) NLP Tasks: The research's primary NLP task is to generate semantically pertinent and contextually accurate responses to inquiries regarding quranic studies. The system employs a Retrieval-Augmented Generation (RAG) architecture, combining retrieval-based and generative methodologies, to ensure that responses are both dataset-based and linguistically coherent. The system executes the following tasks:

- Semantic Search and Retrieval: Upon a user's query submission, the system does a semantic similarity search over the vectorized dataset obtained from Qur'anic surah descriptions [19]. This procedure determines the most contextually pertinent entries from the dataset to respond to the query.
- Response Generation: The retrieved descriptions are submitted to the LLMs, which produce a comprehensive response [14]. This response integrates the retrieved information and provides explanatory content to address the query.
- Citations and Contextualization: Each generated response includes references to the original dataset entries (e.g., surah descriptions or specific virtues), allowing users to trace the information back to its source [25].

2) Evaluation guidelines: To assess the quality of the responses generated by the system, human evaluators followed a structured set of evaluation guidelines. These guidelines provided a consistent framework for scoring responses across three key dimensions: Context Relevance, Answer Faithfulness, and Answer Relevance [14]. Each dimension is explained below, along with its calculation method and examples.

- Context Relevance evaluates how precisely the retrieved and generated responses align with the user query while avoiding irrelevant or extraneous information. The relevance score is calculated using the precision@k metric, where k represents the number of top retrieved results considered as shown in Equation 1.

$$\text{Precision@k} = \frac{\text{No. of relevant results in the top-k responses}}{k} \tag{1}$$

  Example:
  - Query: "What is the reason for Surah Al-Fatihah being named Umm Al-Kitab?"
  - Retrieved Information:
    1) Surah Al-Fatihah is named Umm Al-Kitab because it summarizes the essence of the Qur'an (relevant).
    2) It is recited in every unit of prayer (relevant).
    3) Surah Al-Baqarah discusses laws and stories (irrelevant).
    4) Surah Al-Fatihah has seven verses (relevant).
    5) Surah An-Nas is the last chapter of the Qur'an (irrelevant).

  If $k = 5$, then 3 out of the 5 retrieved results are relevant:

$$\text{Precision@5} = \frac{3}{5} = 0.6$$

  The context relevance score for this response is therefore 0.6.

- Answer Faithfulness ensures that the generated responses accurately represent the retrieved information without introducing unsupported content or hallucinations. Evaluators compare the generated response with the dataset to verify factual consistency.
  Example:
  - Query: "What does Surah Al-Fatihah emphasize?"
  - Retrieved Information: Surah Al-Fatihah emphasizes monotheism, gratitude, and seeking guidance from Allah.
  - Faithful Response: Surah Al-Fatihah highlights the themes of monotheism, gratitude, and the importance of seeking Allah's guidance.
  - Non-Faithful Response: Surah Al-Fatihah emphasizes the stories of past prophets.

  The faithful response adheres strictly to the retrieved information, while the non-faithful response introduces unsupported content.

- Answer Relevance measures whether the response directly addresses the query while maintaining se-

mantic and theological appropriateness. It assesses completeness, clarity, and alignment with the question. Example:

- Query: "Why is Surah Al-Fatihah called Umm Al-Kitab?"
- Relevant Response: Surah Al-Fatihah is called Umm Al-Kitab because it summarizes the central teachings of the Qur'an and is recited in every unit of prayer.
- Irrelevant Response: Surah Al-Fatihah has seven verses and is the first chapter of the Qur'an.

The relevant response directly answers the query, providing reasoning, while the irrelevant response, though factually correct, fails to address the specific question.

3) Evaluation process: Human evaluators assessed the system-generated responses through a web-based platform, where they were presented with prompts, the corresponding responses, and an interface for scoring. The evaluation process included the following steps:

- Reviewing Responses: The process of reviewing responses was a critical step in evaluating the quality of the outputs generated by the large language models (LLMs). Human evaluators carried out this task through a web-based platform specifically designed to facilitate structured and unbiased assessments.
- Scoring System: For each response, evaluators assigned scores on a Likert scale (1 to 5) for the three evaluation criteria: Context Relevance, Answer Faithfulness, and Answer Relevance. In addition to numerical scores, evaluators could provide written feedback to justify their evaluations [16]. This qualitative feedback emphasized specific faults or applauded features of the response, providing more insight into the system's performance.
- Reevaluation and Calibration: Since numerous replies were provided for each query, evaluators could compare the quality of outputs from various LLMs. This comparative approach was instrumental in identifying the models' relative strengths and limitations, thereby enabling a more thorough evaluation [26]. To verify the dependability of their judgments, evaluators returned to a subset of previously evaluated responses on a regular basis and reassessed them. This consistency check allowed evaluators to reflect on their scoring processes and ensure they were in line with the rating criteria.

The structured evaluation guidelines guaranteed that the assessment process was meticulous, consistent, and transparent. The guidelines established a comprehensive framework for assessing the system's performance by emphasizing context relevance, answer faithfulness, and answer relevance [27]. This method enable a thorough comparison of several LLMs and provided vital insights into their performance in answering Islamic queries with contextual precision and faithfulness [15], [27]. The evaluations were submitted through the platform after all responses to a specific query were evaluated, scored, and commented on. In order to provide a comprehensive dataset for the purpose of investigating the LLMs, the platform logged and stored the data for research [28].

C. Dataset Selection and Curation

The dataset used in this study was carefully selected and organized to guarantee it aligns with the goals of assessing large language models (LLMs) within the framework of quranic studies. The process of selection and curation included identifying a reliable source, organizing the data, and confirming its alignment with the research goals [15], [28].

1) Selection criteria: The dataset was chosen according to these specific criteria:

- Authenticity: The source underwent a thorough review to confirm its compliance with recognized Islamic scholarship and the absence of speculative interpretations [26], [27].
- Descriptive Richness: The dataset must deliver comprehensive, contextually rich descriptions that can be effectively employed for semantic search and response generation [27], [29].
- Clarity and Accessibility: The content needed to be created in a structured and clear manner, facilitating both manual review and computational processing [15], [28].
- Relevance: The dataset was meticulously curated to facilitate the process of addressing inquiries related to quranic studies, emphasizing themes that are frequently observed in these discussions [26]

2) Curation process: A comprehensive curation procedure was carried out on the dataset in order to get it ready for integration with the retrieval-augmented generation (RAG) system and LLMs:

- Data Digitization: The text from the source book was digitized to generate a dataset that can be accessed by machines. Optical Character Recognition (OCR) tools were employed where necessary to convert printed material into digital text [15], [28].
- Data Structuring: The content was segmented into surah name, number of verses, reason for the name, general objective, virtues, and relationships. Each field was carefully labeled to facilitate precise retrieval [27].
- Content Validation: The digitized and structured dataset was reviewed by experts in Islamic studies to verify its accuracy and alignment with the original source [26].
- Preprocessing: Unnecessary or redundant information was removed, and inconsistencies were corrected. Tokenization was performed to split the text into smaller, manageable units for processing by the semantic search system.
- Vectorization: The structured data was transformed into high-dimensional vector embeddings using pretrained language models [30]. This step allowed for

efficient and accurate semantic similarity searches within the dataset.

- Storage in a Vector Database: The vectorized dataset was stored in a scalable and efficient vector database, enabling quick retrieval of relevant entries based on user queries [22].

3) Dataset integrity: To ensure the integrity and reliability of the dataset, multiple layers of validation were employed, including manual review and automated consistency checks, regular audits of the data were conducted to identify and rectify any errors or discrepancies, and a backup of the raw and processed datasets was maintained for reproducibility and future reference.

4) Strengths of the dataset:

- Richness in Context: The dataset goes beyond literal translations, providing thematic, historical, and theological insights.
- High Relevance: The information directly supports answering user queries about quranic studies.
- Scalability: The vectorized format enables integration with modern NLP systems and future upgrades.

This curated dataset ensures that the responses generated by the system are accurate, faithful, and contextually relevant, thereby serving as the foundation for the investigating and evaluation of the LLMs

D. Human Evaluators

In this research, human evaluators were instrumental in evaluating the responses produced by the large language models (LLMs) [31]. The evaluations were carried out using a specially designed website that focused on optimizing the evaluation process and maintaining consistency. The website offered evaluators with questions and responses from several LLMs, allowing for a systematic assessment using preset criteria which are context relevance, answer faithfulness, and answer relevance.

1) Evaluator selection: The evaluators were selected with careful consideration to guarantee that they had the proper knowledge and comprehension of quranic studies, given that the study centers on inquiries pertaining to Islamic content. Criteria for selection included:

- Knowledge of quranic studies: Evaluators who had either formal education or significant experience in Islamic studies were prioritized.
- Analytical Skills: In order to evaluate the quality of responses across multiple dimensions, evaluators were required to possess strong analytical skills.
- Familiarity with Evaluation Tasks: It was considered beneficial to have prior experience analyzing textual data or utilizing NLP technologies.

Evaluators from a variety of backgrounds were included to make sure the replies were evaluated fairly and without bias.

2) Evaluation platform: The evaluation process was conducted through a dedicated website designed to facilitate efficient and user-friendly assessments. The platform included the following features:

- Query-Response Display:
  - Each evaluation session displayed a prompt (query) along with responses generated by different LLMs.
  - Responses were anonymized to prevent bias, ensuring that evaluators were not influenced by the identity of the LLM responsible for generating a response.
- Scoring Interface: Evaluators rated each response based on the three evaluation criteria:
  - Context Relevance: Precision and alignment of the response with the query.
  - Answer Faithfulness: Accuracy of the response in relation to the retrieved dataset content.
  - Answer Relevance: Appropriateness and direct pertinence of the response to the query.

  A Likert scale (1–5) was used for scoring, where 1 indicated poor performance and 5 indicated excellent performance.
- Feedback Mechanism: Evaluators could provide written comments to justify their scores or highlight specific issues in the responses. This feature allowed the identification of nuanced errors that might not be captured by numerical scores alone.

3) Significance of human evaluations: The use of human evaluators provided an essential layer of validation for the study, ensuring that the generated responses were assessed not only for technical accuracy but also for their theological and contextual integrity [31]. By leveraging a well-structured platform and robust evaluation criteria, the study ensured that the investigation of LLMs was both rigorous and comprehensive, offering valuable insights into their performance in responding to Islamic queries [32].

E. Metric for Quality Evaluation

This study employs Inter-Evaluator Agreement (IEA) as the primary metric to ensure the quality, reliability, and consistency of human evaluations [33]. Since human evaluators, rather than annotators, were tasked with assessing the generated responses, IEA provides a robust measure of agreement across evaluators, validating the credibility of the evaluation process and the results.

IEA measures the level of consistency between evaluators when scoring responses based on predefined criteria: context relevance, answer faithfulness, and answer relevance. An elevated IEA score signifies uniform application of evaluation criteria by assessors, whereas a diminished score reveals inconsistencies that may necessitate recalibration.

Fleiss' Kappa, as presented in Eq. (2), was employed to calculate IEA due to its appropriateness for evaluating agreement among multiple evaluators concurrently [33]. The consistent monitoring of IEA scores allowed the research team to detect discrepancies promptly and facilitate recalibration

sessions as needed, ensuring evaluators' comprehension and interpretation of the scoring guidelines were aligned. This methodical approach guaranteed that the evaluation process was dependable, replicable, and aligned with the study's goals.

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{2}$$

where:

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i \quad \text{and} \quad \bar{P}_e = \sum_{k=1}^{K} p_k^2 \tag{3}$$

with:

$$P_i = \frac{1}{n(n-1)} \sum_{k=1}^{K} n_{ik}(n_{ik} - 1), \quad p_k = \frac{\sum_{i=1}^{N} n_{ik}}{Nn} \tag{4}$$

where:

- κ: Fleiss' Kappa value.
- $\bar{P}$: Average observed agreement across all items.
- $\bar{P}_e$: Expected agreement based on chance.
- P_i: Proportion of agreement for item i.
- p_k: Proportion of ratings in category k.
- n: Total number of ratings per item.
- n_{ik}: Number of raters who assigned category k to item i.
- N: Total number of items.
- K: Total number of categories.

F. Large Language Models

The efficacy of numerous large language models (LLMs) in responding to inquiries regarding quranic studies is assessed in this study using a Retrieval-Augmented Generation (RAG) framework [34]. A comprehensive comparison of the capabilities of the LLMs selected for investigation is possible due to the fact that they represent a variety of architectures and parameter scales. More information about the models, how they are put together, and how they relate to this study is given below.

1) Llama: Meta AI has developed the Llama (Large Language Model Meta AI) family of models, which are state-of-the-art transformer-based architectures that are optimized for natural language understanding and generation [35]. Llama models are trained on vast, diversified corpora to do various NLP tasks such as contextual reasoning, question answering, and text summarization. They are available in a variety of parameter values, which provides a degree of flexibility in terms of computational requirements and performance. The Llama models were incorporated in this investigation due to their adaptability and superior performance across various parameter sizes. A thorough investigation of how model size affects the capacity to produce faithful, accurate, and contextually relevant replies was made possible by the range of configurations. The comparison of Llama generations (e.g., Llama3 with Llama3.1) yielded insights on the impact of incremental architectural enhancements on performance.

2) Gemma: Google's DeepMind developed the Gemma family of large language models, which are a set of transformer-based designs that are best for understanding and creating natural language. The Gemma models are engineered to provide superior performance while ensuring efficiency, rendering them adaptable for various jobs. These models have undergone pre-training on varied and comprehensive datasets, enabling them to generalize effectively across multiple domains [36]. Gemma models are offered in many parameter scales, including 27b, 9b, and 2b, providing flexibility to optimize performance and computational demands. Their versatility renders them appropriate for applications from resource-intensive jobs to real-time implementations in limited surroundings.

3) QwQ: The QwQ model developed by Alibaba Cloud, which has 32 billion parameters, is a large-scale transformer-based language model designed to handle complex natural language processing tasks [37]. While specific information about the QwQ model's architecture or pre-training details is limited in comparison to more established models like Llama and Gemma, its parameter scale positions it as a powerful model capable of capturing complex relationships in textual data. Its large size enables it to perform effectively on tasks requiring nuanced comprehension, contextual reasoning, and content generation across a wide range of subjects.

4) Phi: Microsoft created the Phi family of language models, which are a group of lightweight transformer-based models that work best for jobs that involve processing natural language. Even though the Phi models have lower parameter sizes compared to other large-scale models such as Llama and Gemma, they are engineered to exceed expectations, providing robust performance while ensuring computational efficiency [38]. They are pretrained on rigorously selected datasets that prioritize high-quality information, enabling effective generalization across tasks despite a reduced number of parameters. This method guarantees that Phi models provide an exceptional equilibrium between performance and resource demands, rendering them especially appropriate for resource-limited settings and real-time applications.

5) Key Features of the models: From smaller-scale models (e.g., 1 billion parameters) to large-scale ones (e.g., 70 billion parameters), the chosen models encompass a wide spectrum of parameter sizes. This variation enables a study of the trade-offs between response quality and computing efficiency. Smaller models might lose accuracy and contextual depth, yet producing faster responses with reduced processing expenses. On the other hand, it is anticipated that larger models will produce more nuanced and high-fidelity outputs, although at the expense of increased computational demands.

The models use transformer architectures, which are efficient in comprehending and producing natural language content. Their architecture allows to capturing the intricate patterns, correlations, and contextual subtleties within the dataset. This work emphasizes the models' capacity to adjust to specific domains, such as quranic studies, despite being

trained on varied datasets. The assessment analyzes the efficacy of these models when enhanced with domain-specific data through the RAG methodology. The study evaluates the models in a zero-shot context, without any fine-tuning on the unique dataset. This method emphasizes the models' intrinsic capacity to generalize and appropriately respond by utilizing external information obtained from the descriptive dataset.

6) Integration with the RAG framework: The selected LLMs were integrated with the RAG framework, allowing them to generate contextually relevant responses based on the dataset. The RAG framework enhances the performance of the models by providing contextual input, reducing hallucination, and facilitating citations.

G. Hardware and Software Configuration

The experiments were conducted using a high-performance computing system equipped with an Intel(R) Xeon(R) Gold 5318Y CPU operating at 2.10 GHz with 24 cores. The system featured four NVIDIA RTX A6000 GPUs, each providing 48 GB of VRAM, enabling efficient handling of computationally intensive tasks, particularly those involving deep learning models. Additionally, the system was supported by 128 GB of RAM, ensuring smooth execution of memory-intensive operations and facilitating large-scale data processing. This configuration provided the computational resources necessary to run and evaluate the models effectively.

III. Results

The results of this study are based on the implementation of a RAG architecture, designed to evaluate the performance of 13 LLMs in answering questions related to Quranic studies. By leveraging a descriptive dataset of Quranic surahs, the RAG system facilitates the integration of external knowledge to address the limitations of standalone models. The comparative analysis focuses on assessing the relevance, faithfulness, and contextual accuracy of the responses generated by the LLMs within this framework.

A. Experimental Results

The experimental evaluation assessed the capacity of a variety of large language models (LLMs) to respond to queries related to quranic studies. The models were assessed based on three critical metrics: context relevance, answer faithfulness, and answer relevance. The models were categorized into three categories based on their parameter sizes: large models (marked in red), medium models (marked in yellow), and small models (marked in green). Therefore, the results were analyzed. The following is a comprehensive analysis of the performance of each category.

1) Context relevance: Context relevance as shown in Fig. 1 evaluates how well the generated responses align with the query's context.

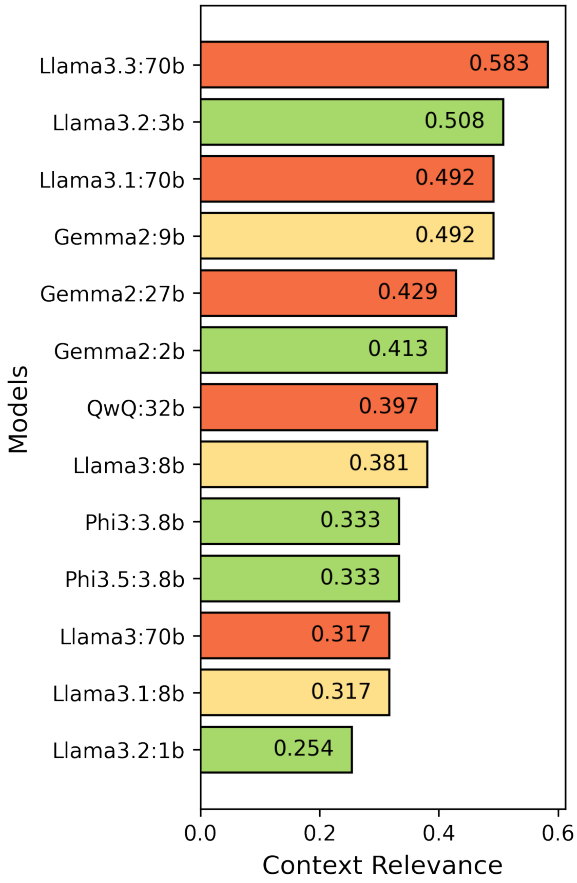


Fig. 1. Context relevance by the 13 LLMs.

- Large Models (Red): The large models outperformed other categories, with Llama3.3:70b achieving the highest score of 0.583, followed by Llama3.2:3b (0.508), despite being categorized as a small model. Both Llama3.1:70b and Gemma2:27b achieved competitive scores of 0.492 and 0.429, respectively, while QwQ:32b recorded 0.397. These models excel at retrieving relevant information and aligning responses with the query intent due to their larger parameter size.
- Medium Models (Yellow): Among the medium models, Gemma2:9b performed best with a score of 0.492, comparable to some large models. Llama3:8b and Llama3.1:8b followed with scores of 0.381 and 0.317, respectively. These models demonstrated decent performance but lagged behind the large models in handling complex or nuanced queries.
- Small Models (Green): The small models struggled overall, with Llama3.2:1b achieving the lowest score of 0.254. Phi3.5:3.8b and Phi3:3.8b performed moderately with scores of 0.333, while Gemma2:2b achieved 0.413. Notably, Llama3.2:3b outperformed all expectations with a score of 0.508, surpassing even some medium and large models.

2) Answer faithfulness: Answer faithfulness as shown in Fig. 2 measures whether the responses remain consistent with the retrieved content, avoiding inaccuracies or hallucinations.

- Large Models (Red): The large models dominated this metric, with Llama3.2:3b (exceptionally performing despite its small category) achieving the top score of 4.619. Llama3:70b and Llama3.1:70b scored

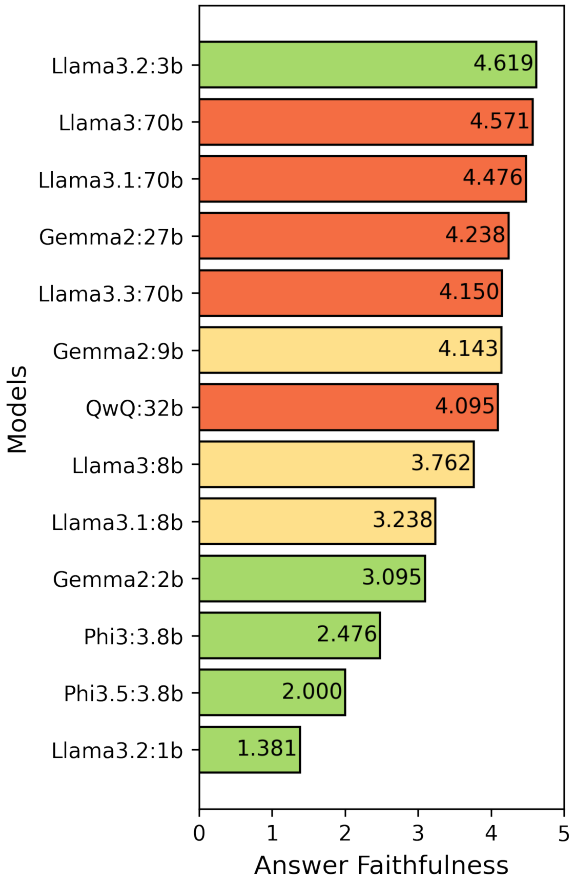


Fig. 2. Answer faithfulness by the 13 LLMs.

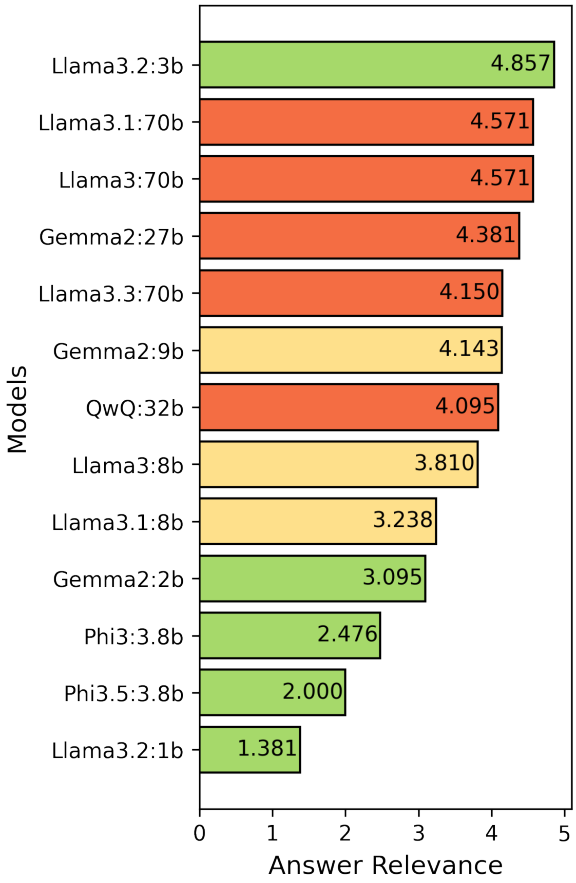


Fig. 3. Answer relevance by the 13 LLMs.

4.571 and 4.476, respectively, while Gemma2:27b and QwQ:32b followed with scores of 4.238 and 4.095. Their ability to maintain faithfulness highlights the advantages of larger parameter sizes.

- Medium Models (Yellow): The medium models showed reliable performance, with Gemma2:9b scoring 4.143 and Llama3:8b achieving 3.762. Llama3.1:8b, while consistent, lagged slightly behind with 3.238. These models balanced faithfulness and efficiency but struggled with queries requiring deep contextual reasoning.
- Small Models (Green): The small models faced significant challenges. Llama3.2:1b recorded the lowest faithfulness score of 1.381, and Phi3.5:3.8b achieved 2.000, indicating frequent inconsistencies. While Phi3:3.8b scored slightly higher at 2.476, Llama3.2:3b stood out with a score of 4.619, showcasing exceptional faithfulness that rivaled larger models.

3) Answer relevance: Answer relevance as shown in Fig. 3 assesses whether the generated responses address the query's intent effectively.

- Large Models (Red): The Llama3.2:3b, while classified as small, achieved the highest relevance score of 4.857, followed closely by Llama3:70b and Llama3.1:70b, both scoring 4.571. Gemma2:27b and QwQ:32b continued to perform well, with scores of 4.381 and 4.095, respectively. These models consistently delivered relevant responses aligned with the intent behind complex queries.
- Medium Models (Yellow): The medium models provided strong performance, particularly Gemma2:9b, which achieved 4.143. Llama3:8b followed with 3.810, and Llama3.1:8b recorded a slightly lower score of 3.238. These models addressed moderately complex queries effectively but occasionally lacked depth.
- Small Models (Green): The small models exhibited varied performance, with Llama3.2:1b scoring the lowest at 1.381. Phi3.5:3.8b and Phi3:3.8b recorded scores of 2.000 and 2.476, respectively, indicating challenges in providing fully relevant responses. However, Llama3.2:3b once again stood out, achieving the highest score of 4.857, performing on par with the best large models.

4) Intercoder agreement: We assessed intercoder agreement, which measures the extent to which evaluators within the same group report the same evaluation for a given instance. To compute this metric, we examined the percentage of instances where both evaluators assigned identical evaluation scores independently. This measure enabled us to evaluate the degree

of concordance between evaluators and assess their consistency in evaluation.

TABLE I. THE KAPPA VALUES FOR THE EVALUATION

Creator	Model	Evaluator
Meta	Llama 3.3 70B	0.80
Meta	Llama 3.2 3B	0.90
Meta	Llama 3.2 1B	0.82
Meta	Llama 3.1 70B	0.82
Meta	Llama 3.1 8B	0.85
Meta	Llama 3 70B	0.92
Meta	Llama 3 8B	0.80
Google	Gemma 2 27B	0.90
Google	Gemma 2 9B	0.92
Google	Gemma 2 2B	0.83
Microsoft	Phi 3.5 3.8B	0.83
Microsoft	Phi 3.3 3.8B	0.93
Alibaba	QwQ 32B	0.89

In Table I, we delve into the inter-annotator agreement analysis, which measures the level of agreement between human evaluators across different models using Fleiss' Kappa. This indicates a similar level of agreement between the evaluators.

IV. DISCUSSION

This section discusses the experimental findings, analyzing the performance of various LLMs categorized into large, medium, and small models across the three evaluation metrics: context relevance, answer faithfulness, and answer relevance [14]. This study examines the relationship among model size, response quality, and computational trade-offs, as well as the behavior of models within the Retrieval-Augmented Generation (RAG) framework.

A. Performance Insights Based on Model Size

The experimental results show that the quality of the responses across all three evaluation criteria is significantly affected by the model size.

- Large Models (Red): Large models, including Llama3:70b, Llama3.1:70b, Llama3.3:70b, Gemma2:27b, and QwQ:32b, consistently demonstrate superior performance compared to medium and small models. The models demonstrated superior context relevance due to their larger parameter sizes [39], which facilitate a deeper semantic understanding and enhance their ability to retrieve and integrate pertinent information. Furthermore, their enhanced performance in answer faithfulness and relevance indicates that large models are more adept at minimizing hallucinations [40] and producing responses that accurately address user queries. However, their computational demands remain a major trade-off, requiring substantial memory and processing power. This limits their accessibility in resource-constrained environments, making them more suitable for high-performance systems.
- Medium Models (Yellow): Medium-sized models like Gemma2:9b, Llama3:8b, and Llama3.1:8b demonstrated strong performance relative to their parameter sizes, particularly in answer faithfulness and relevance. These models provide a balance between computational efficiency and response quality, making them ideal for systems where resources are limited but accuracy cannot be compromised. However, their performance in context relevance was slightly lower than that of large models, indicating limitations in capturing deeper relationships within the data. Medium models present a viable option for applications requiring moderate precision while maintaining resource efficiency.
- Small Models (Green): The small models, including Llama3.2:3b, Llama3.2:1b, Gemma2:2b, Phi3:3.8b, and Phi3.5:3.8b, faced significant challenges in delivering high-quality responses. Models like Llama3.2:1b and Phi3.5:3.8b scored the lowest across all metrics, reflecting their inability to process complex queries effectively due to their smaller parameter size. Interestingly, Llama3.2:3b emerged as an outlier, achieving performance levels comparable to the large models, particularly in answer faithfulness (4.619) and answer relevance (4.857). This unexpected performance demonstrates the efficacy of architectural improvements and pre-training techniques, even on smaller models. Although small models are computationally efficient and well-suited for lightweight tasks [41], their overall limitations render them less suitable for complex queries that necessitate a deep comprehension of semantics.

B. Effectiveness of the RAG Framework

The implementation of the Retrieval-Augmented Generation (RAG) framework significantly enhanced the response quality of the evaluated models [42]. By incorporating external knowledge from the descriptive Qur'anic dataset, the models can retrieve pertinent information prior to formulating responses. This method alleviated the prevalent issue of hallucination, where language models produce factually inaccurate or irrelevant responses.

The results demonstrate that the larger models derived the greatest advantage from the RAG framework, as their higher parameter sizes facilitated more effective integration of the retrieved context, resulting in enhanced performance across all metrics. Medium and small models showed enhancements, nevertheless, their capacity limitation affected the integration of retrieved knowledge, leading to lower context alignment and fewer accurate responses.

C. Trade-Off Between Computational Resources and Response Quality

The results of this research highlight the trade-off between response quality and computational efficiency throughout several model sizes.

- Large Models deliver outstanding performance but they are less practical for real-time or cost-sensitive installations since they need large computational resources.
- Medium Models fit for uses with intermediate hardware availability since they offer a useful compromise between dependability of performance and low resource usage.

- Small Models are quick and light-weight, yet they usually underperform on jobs needing sophisticated thinking. Nonetheless, the unexpected findings in Llama3.2:3b suggest that smaller models can still show good performance.

This trade-off emphasizes the need of choosing the suitable model size depending on particular application criteria including accuracy, speed, and resource availability.

D. Surprising Performance of Llama3.2:3b

This study's outstanding discovery is the exceptional performance of Llama3.2:3b, despite its classification as a small model. It outperformed a number of medium and even big models, achieving the highest results in answer faithfulness and answer relevance. According to this finding, factors including pre-training quality, data efficiency, and architectural upgrades can have an impact on model performance, in addition to parameter size. The potential for smaller models to produce high-quality outputs in resource-efficient environments is underscored by the robust performance of Llama3.2:3b when used in conjunction with effective frameworks such as RAG.

E. Implications for Domain-Specific Tasks

The research emphasizes both the difficulties and potential benefits of utilizing general-purpose LLMs for specialized tasks, including responding to inquiries about quranic studies. Although large models excelled in aligning responses with the given dataset, their effectiveness is significantly dependent on the quality and organization of the retrieved content. This research demonstrates how important it is to add domain-specific knowledge [4], like curated descriptive datasets, to improve the models' abilities and lower the risk of hallucinations. The RAG framework proved to be an effective technique for ensuring that responses were contextually correct and based on credible sources.

The experimental results offer important insights into the performance of LLMs of different sizes. Large models provide exceptional accuracy and relevance, though they require significant resources, whereas medium models offer a practical compromise between performance and efficiency. Smaller models, while typically less powerful, demonstrated surprising potential, especially Llama3.2:3b, which excelled across various metrics. The use of the RAG framework enhanced the models' performance by reducing hallucinations and grounding responses in reliable data. These findings highlight the importance of model selection, optimization strategies, and retrieval mechanisms when applying LLMs to domain-specific tasks. Future work will include conducting an ablation study to compare the performance of models with and without the RAG framework, evaluating additional large language models (LLMs), and leveraging LLMs for automatic evaluation of answer faithfulness and answer relevance metrics. Additional fine-tuning strategies and assessments in other specialized domains can also be explored.

V. Conclusions

This study evaluated multiple large language models (LLMs) of different sizes in responding to Quranic studies-related queries using a Retrieval-Augmented Generation (RAG) framework. The findings indicate that large models, such as Llama3:70b, Llama3.1:70b, and Gemma2:27b, consistently delivered superior performance in context relevance, answer faithfulness, and answer relevance. However, their computational demands pose challenges for practical deployment. Medium-sized models, including Gemma2:9b and Llama3:8b, demonstrated a balance between efficiency and performance, making them suitable for moderately complex tasks. Interestingly, Llama3.2:3b, a small model, performed comparably to larger models in certain aspects, particularly in answer faithfulness and relevance, suggesting that architectural optimizations can enhance the capabilities of smaller models.

The study also highlights the importance of the RAG framework in improving response quality by grounding answers in external domain-specific knowledge, reducing hallucinations, and ensuring more reliable outputs. These findings emphasize the trade-offs between model size, performance, and computational efficiency, indicating that while large models are ideal for high-accuracy tasks, smaller models, when optimized, can serve as viable alternatives. Future research can focus on further optimizations, fine-tuning strategies, and expanding dataset diversity to enhance model performance across different applications.

Acknowledgment

This research is supported by Universitas Islam Riau.

References

[1] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam, "A review on large language models: Architectures, applications, taxonomies, open issues and challenges," *IEEE Access*, 2024.

[2] A. H. Nasution and A. Onan, "Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks," *IEEE Access*, 2024.

[3] S. Alnefaie, E. Atwell, and M. A. Alsalka, "Is gpt-4 a good islamic expert for answering quran questions?" in *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, 2023, pp. 124–133.

[4] Q. Zhou, C. Liu, Y. Duan, K. Sun, Y. Li, H. Kan, Z. Gu, J. Shu, and J. Hu, "Gastrobot: a chinese gastrointestinal disease chatbot based on the retrieval-augmented generation," *Frontiers in Medicine*, vol. 11, p. 1392555, 2024.

[5] N. Pipitone and G. H. Alami, "Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain," *arXiv preprint arXiv:2408.10343*, 2024.

[6] A. Alrayzah, F. Alsolami, and M. Saleh, "Challenges and opportunities for arabic question-answering systems: current techniques and future directions," *PeerJ Computer Science*, vol. 9, p. e1633, 2023.

[7] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, and Q. Li, "A survey on rag meeting llms: Towards retrieval-augmented large language models," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 6491–6501.

[8] Z. Sun, X. Zang, K. Zheng, Y. Song, J. Xu, X. Zhang, W. Yu, and H. Li, "Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability," *arXiv preprint arXiv:2410.11414*, 2024.

[9] S. Patel, H. Kane, and R. Patel, "Building domain-specific llms faithful to the islamic worldview: Mirage or technical possibility?" *arXiv preprint arXiv:2312.06652*, 2023.

[10] Y. B. I. Radi, *Al-Bitaqat: Chapters of the Noble Quran Explored in 114 Cards*. Dakwah Corner Bookstore (M) Sdn. Bhd, 2023.

[11] C. Njeh, H. Nakouri, and F. Jaafar, "Enhancing rag-retrieval to improve llms robustness and resilience to hallucinations," in *International Conference on Hybrid Artificial Intelligence Systems*. Springer, 2024, pp. 201–213.

[12] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[13] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[14] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.

[15] E. Kamalloo, N. Dziri, C. L. Clarke, and D. Rafiei, "Evaluating open-domain question answering in the era of large language models," *arXiv preprint arXiv:2305.06984*, 2023.

[16] Y. Zhou, Y. Liu, X. Li, J. Jin, H. Qian, Z. Liu, C. Li, Z. Dou, T.-Y. Ho, and P. S. Yu, "Trustworthiness in retrieval-augmented generation systems: A survey," *arXiv preprint arXiv:2409.10102*, 2024.

[17] H. K. Shabaz Patel and R. Patel, "Building domain-specific llms faithful to the islamic worldview: Mirage or technical possibility?" *arXiv preprint arXiv:2312.06652*, 2023.

[18] S. Gupta, R. Ranjan, and S. N. Singh, "A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions," *arXiv preprint arXiv:2410.12837*, 2024.

[19] Y. L. et al., "Datasets for large language models: A comprehensive survey," *arXiv preprint arXiv:2402.18041*, 2024.

[20] M. R. Rizqullah, A. Purwarianti, and A. F. Aji, "Qasina: Religious domain question answering using sirah nabawiyah," in *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*. IEEE, 2023, pp. 1–6.

[21] M. S. Abubakari, W. Shafik, and A. F. Hidayatullah, "Evaluating the potential of artificial intelligence in islamic religious education: A swot analysis overview," in *AI-Enhanced Teaching Methods*. IGI Global, 2024, pp. 216–239.

[22] Y. Han, C. Liu, and P. Wang, "A comprehensive survey on vector database: Storage and retrieval technique, challenge," *arXiv preprint arXiv:2310.11703*, 2023.

[23] Z. Jing, Y. Su, Y. Han, B. Yuan, H. Xu, C. Liu, K. Chen, and M. Zhang, "When large language models meet vector databases: A survey," *arXiv preprint arXiv:2402.01763*, 2024.

[24] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1–17, 2023.

[25] A. Y. Alan, E. Karaarslan, and Ö. Aydin, "A rag-based question answering system proposal for understanding islam: Mufassirqas llm," *arXiv preprint arXiv:2401.15378*, 2024.

[26] D. Tam, A. Mascarenhas, S. Zhang, S. Kwan, M. Bansal, and C. Raffel, "Evaluating the factual consistency of large language models through summarization," *arXiv preprint arXiv:2211.08412*, 2022.

[27] W. Zhou, S. Zhang, H. Poon, and M. Chen, "Context-faithful prompting for large language models," *arXiv preprint arXiv:2303.11315*, 2023.

[28] C. Wang, S. Cheng, Q. Guo, Y. Yue, B. Ding, Z. Xu, Y. Wang, X. Hu, Z. Zhang, and Y. Zhang, "Evaluating open-qa evaluation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[29] Y. Huang, S. Chen, H. Cai, and B. Dhingra, "Enhancing large language models' situated faithfulness to external contexts," *arXiv e-prints*, pp. arXiv–2410, 2024.

[30] S. S. Monir, I. Lau, S. Yang, and D. Zhao, "Vectorsearch: Enhancing document retrieval with semantic embeddings and optimized search," *arXiv preprint arXiv:2409.17383*, 2024.

[31] A. Elangovan, L. Liu, L. Xu, S. Bodapati, and D. Roth, "Considers-the-human evaluation framework: Rethinking human evaluation for generative large language models," *arXiv preprint arXiv:2405.18638*, 2024.

[32] K. Feng, K. Ding, K. Ma, Z. Wang, Q. Zhang, and H. Chen, "Sample-efficient human evaluation of large language models via maximum discrepancy competition," *arXiv preprint arXiv:2404.08008*, 2024.

[33] F. Moons and E. Vandervieren, "Measuring agreement among several raters classifying subjects into one-or-more (hierarchical) nominal categories. a generalisation of fleiss' kappa," *arXiv preprint arXiv:2303.12502*, 2023.

[34] S. B. Islam, M. A. Rahman, K. Hossain, E. Hoque, S. Joty, and M. R. Parvez, "Open-rag: Enhanced retrieval-augmented reasoning with open-source large language models," *arXiv preprint arXiv:2410.01782*, 2024.

[35] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[36] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé *et al.*, "Gemma 2: Improving open language models at a practical size," *arXiv preprint arXiv:2408.00118*, 2024.

[37] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[38] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.

[39] S. Badshah and H. Sajjad, "Quantifying the capabilities of llms across scale and precision," *arXiv preprint arXiv:2405.03146*, 2024.

[40] S. Tonmoy, S. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das, "A comprehensive survey of hallucination mitigation techniques in large language models," *arXiv preprint arXiv:2401.01313*, 2024.

[41] L. Chen and G. Varoquaux, "What is the role of small models in the llm era: A survey," *arXiv preprint arXiv:2409.06857*, 2024.

[42] X. Su and Y. Gu, "Implementing retrieval-augmented generation (rag) for large language models to build confidence in traditional chinese medicine," 2024.

Advanced Optimization of RPL-IoT Protocol Using ML Algorithms

Mansour Lmkaiti[1], Ibtissam Larhlimi[2], Maryem Lachgar[3], Houda Moudni[4], Hicham Mouncif[5]
LIMATI Laboratory-Polydisciplinary Faculty, University Sultan Moulay Slimane, Morocco[1,2,3,5]
TIAD Laboratory, Faculty of Sciences and Technology , University Sultan Moulay Slimane, Morocco[4]

Abstract*—This study explores the transformative potential of machine learning (ML) algorithms in optimizing the Routing Protocol for Low-Power and Lossy Networks (RPL), addressing critical challenges in Internet of Things (IoT) networks, such as Expected Transmission Count (ETX), latency, and energy consumption. The research evaluates the performance of Random Forest, Gradient Boosting, Artificial Neural Networks (ANNs), and Q-Learning across IoT network simulations with varying scales (50, 100, and 150 nodes). Results indicate that tree-based models, particularly Random Forest and Gradient Boosting, demonstrate robust predictive capabilities for ETX and latency, achieving consistent results in smaller and medium-sized networks. Specifically, for 50-node networks, Neural Networks achieved the best performance with the lowest latency (2.43862 ms) and the best ETX (5.29557), despite slightly higher energy consumption. For 100-node networks, Q-Learning stood out with the lowest energy consumption (1.62973 J) and competitive ETX (2.70647), though at the cost of increased latency. In 150-node networks, Q-Learning again outperformed other models, achieving the lowest latency (0.68 ms) and energy consumption (2.21 J), though at the cost of higher ETX. Neural Networks excel in capturing non-linear dependencies but face limitations in energy-related metrics, while Q-Learning adapts dynamically to network changes, achieving remarkable latency reductions at the cost of transmission efficiency. The findings highlight key trade-offs between performance metrics and emphasize the need for algorithmic strategies tailored to specific IoT applications. This work not only validates the scalability and adaptability of ML approaches but also lays the foundation for intelligent and efficient IoT network optimization, laying the groundwork for future advancements in sustainable and scalable IoT networks.

Keywords*—IoT; RPL; machine learning; routing efficiency; energy consumption; expected transmission count; network optimization; Artificial Intelligence (AI)

I. Introduction

The Internet of Things (IoT) has revolutionized modern technological paradigms, enabling seamless connectivity and interaction between billions of devices, ranging from consumer electronics to industrial machinery [1]. This explosion of interconnected systems has catalyzed advancements in diverse fields, including smart cities, healthcare, agriculture, and industrial automation [2]. At the core of IoT systems lies the Routing Protocol for Low-Power and Lossy Networks (RPL), a pivotal framework designed to support the unique challenges of Wireless Sensor Networks (WSNs) operating under constrained resources [3]. Despite its widespread adoption, RPL's default mechanisms often fall short in optimizing key performance metrics such as ETX, latency, and energy consumption—metrics critical for ensuring scalability and reliability in large, dynamic IoT environments [2],[4].

The rapid growth of IoT devices has increased the complexity of network management, particularly in dynamic topologies with resource constraints. Traditional heuristic and rule-based optimization methods, while effective in static environments, lack the adaptability and precision needed to address the complex challenges of real-world IoT deployments. These limitations underscore the necessity for innovative solutions that leverage data-driven insights to optimize network behavior dynamically and efficiently [5].

Machine learning (ML), with its unparalleled ability to analyze complex datasets and extract actionable insights, emerges as a transformative solution for optimizing IoT networks. Unlike conventional methods, ML models can adapt to evolving network conditions, predict performance trends, and optimize resource allocation intelligently. Supervised learning techniques such as Random Forests and Gradient Boosting have demonstrated exceptional accuracy in predicting ETX and latency by effectively capturing feature interactions and avoiding overfitting. Similarly, Artificial Neural Networks (ANNs) excel in modeling non-linear dependencies within high-dimensional datasets, enabling precise energy consumption predictions. Reinforcement learning approaches like Q-Learning introduce an adaptive framework, allowing IoT networks to learn optimal routing policies through continuous interaction with the environment [6], [7].

This study delves into the application of ML techniques for enhancing RPL [26] performance in IoT networks, focusing on the evaluation of Random Forests, Gradient Boosting, ANNs, and Q-Learning across varying network sizes. By analyzing the scalability and adaptability of these models through extensive simulations, this work offers a comprehensive comparison of their strengths and limitations. The results demonstrate that ML-based optimization not only improves energy efficiency and reduces latency but also enhances network reliability, laying the groundwork for the development of intelligent, self-optimizing IoT systems.This research connects theory with practice, demonstrating how machine learning can revolutionize IoT network [28] management. The insights gained provide a roadmap for leveraging ML to overcome the critical challenges of energy efficiency, scalability, and adaptability in RPL-based IoT systems, thereby setting a benchmark for next-generation IoT deployments.The remainder of this paper is structured as follows: Section II presents related works, Section III formulates the problem statement, and Section IV describes the machine learning [29] algorithms applied. Section V discusses experimental results, and Section VI concludes the paper with key findings and future directions.

II. Related Work

The Internet of Things (IoT) [2] has emerged as a transformative domain, enabling interconnected devices to communicate, sense and transmit data across diverse environments. At the heart of this ecosystem lies the Routing Protocol for Low-Power and Lossy Networks (RPL), established by the Internet Engineering Task Force (IETF), which serves as a standardized protocol to facilitate efficient data exchange in IoT networks [4]. RPL effectively addresses key limitations of IoT devices, such as constrained computing power, limited memory capacity, and high energy consumption. However, optimizing RPL to enhance energy efficiency, reduce latency, and extend network lifetime remains an ongoing research challenge. In this context, machine learning (ML) [7] has emerged as a powerful tool for addressing these challenges, providing a robust framework for analyzing complex data patterns, enabling predictive insights, and supporting adaptive decision-making.

Supervised learning models, such as Random Forests and Gradient Boosting [19], have shown exceptional promise in predicting key network metrics like ETX and latency, leveraging ensemble approaches to improve routing decisions while avoiding overfitting. Similarly, Artificial Neural Networks (ANNs) [20], with their ability to model non-linear dependencies, have demonstrated potential in optimizing energy consumption and resource allocation in IoT environments. Reinforcement learning techniques, such as Q-Learning, introduce a dynamic approach to routing optimization, enabling IoT[24] systems to autonomously learn and adapt to changing network conditions while balancing exploration and exploitation. Empirical studies confirm the effectiveness of ML algorithms in reducing energy consumption while maintaining communication performance. Approaches such as dynamic sleep scheduling and intelligent data aggregation have been shown to significantly extend network lifetimes, while deep learning architectures have excelled in extracting intricate patterns from high-dimensional datasets, enhancing network performance.

The comparative evaluation of ML algorithms [6] highlights their unique capabilities, applications, strengths, and limitations, providing a comprehensive understanding of their utility in optimizing IoT networks. Random Forests [23], robust and interpretable ensemble methods, are particularly effective in avoiding overfitting and are widely applied in domains like astronomy and energy prediction [8], [9], [10], [11]. However, they can occasionally underperform compared to Gradient Boosting in tasks requiring higher precision [7]. Decision Trees, valued for their simplicity and transparency, are effective in applications such as medical diagnostics and crop disease classification [8], [9], [10], [12], [13], but their propensity to overfit makes them less reliable as standalone models [1], [6], [7], [14], [15]. Gradient Boosting, a sequential learning method, achieves high accuracy by iteratively correcting errors, making it well-suited for diverse tasks [8], [9], [10], [16], [17], though its computational demands and tuning complexity can pose challenges [1], [2], [4], [6], [18], [19], [20], [21], [22]. Q-Learning, a reinforcement learning technique, excels in dynamic decision-making environments such as IoT network optimization [8], [9], [10], while Neural Networks demonstrate unparalleled ability to handle high-dimensional and non-linear data patterns, making them indispensable for complex domains like image recognition and medical diagnostics [8], [9], [10], [23], [24]. However, their reliance on large datasets and high computational requirements can limit their application in resource-constrained settings [25].Together, these observations underscore the transformative potential of machine learning in IoT network optimization [27], emphasizing the importance of aligning algorithmic selection with specific application requirements. By dynamically evolving and adapting to real-time data, ML algorithms are positioned at the forefront of contemporary IoT [30] research, paving the way for intelligent, sustainable, and resilient network systems that address the pressing demands of scalability, efficiency, and adaptability.

III. Problem Statement

In this section, we formulate the problem of optimized routing RPL-based IoT network[30] considering the following metrics:ETX, the latency (LT) and the energy consumption (EC). The objective function integrating these criteria is defined as follows.

$$Minimize\ F = \ w_1.ETX + \ w_2.LT + \ w_3.EC \quad (1)$$

Where w_1, w_2 and w_3 are weights assigned to ETX, LT and EC respectively.

A. Define the Metrics

ETX measures the number of expected transmissions, including retransmissions, required to successfully deliver a packet over a link.

$$ETX_{ij} = \frac{1}{P_{ij}.P_{ji}}$$

Where P_{ij} is the probability of successful packet transmission from node i to node j, and P_{ji} is the probability of successful acknowledgment.
LT represents the time required for a packet to travel from the source to the destination.

$$LT_{ij} = \ d_{ij} + \sum_k ProcessingTime_k$$

Where d_{ij} is the propagation delay between nodes i and j, and the sum represents the processing delays at intermediate nodes. EC is the of energy consumed to transmit a packet from the source to the destination.

$$EC_{ij} = \ TE_{ij} + \sum_k ProcessingEnergy_k$$

Where TE_{ij} is the energy consumed for transmission between nodes i and j, and the sum represents the energy consumed at intermediate nodes for processing.

B. Formulate the Constraints

The connectivity constraint ensures that the selected path maintains network connectivity.

$$\sum_{j \in N} x_{ij} = 1, \quad \forall\ i\ \in\ N$$

Where xij is a binary variable indicating whether the link between nodes i and j is part of the path (1) or not (0). The

Loop-Free constraint ensures that routing path does not exceed the available energy at any node.

$$\sum_{j \in N} x_{ij} = 1, \quad \forall\, i \ \in \ N$$

The energy constraint ensures that the energy consumption does not exceed the available energy at any node.

$$EC_{ij} \ \leq \ E_i, \quad \forall\, i \ \in \ N$$

Where E_i is the available energy at node i.

C. Optimization Problem Formulation

$$Minimize\, F = \sum_{i,j \in E} (w_1.ETX_{ij} + w_2.LT_{ij} + w_3.EC_{ij}).x_{ij}$$

Subject to:

$$\sum_{j \in N} x_{ij} = 1, \quad \forall\, i \ \in \ N$$

$$x_{ij} \ + \ x_{ji} \ \leq \ 1, \quad \forall\, i, j \ \in \ N$$

$$EC_{ij} \ \leq \ E_i, \quad \forall\, i \ \in \ N$$

$$x_{ij} \in 0, 1$$

IV. Machine Learning Algorithms

The application of Machine Learning (ML) in IoT networks represents a paradigm shift in optimizing routing protocols, particularly the Routing Protocol for Low-Power and Lossy Networks (RPL) [3]. Unlike traditional heuristic-based approaches, ML algorithms provide data-driven solutions that dynamically adapt to the ever-changing conditions of IoT networks. By leveraging the vast amounts of data generated within IoT systems, these algorithms can predict network behaviors, optimize performance metrics, and enable intelligent decision making.

In the context of RPL optimization [26], ML techniques offer significant advantages in addressing critical metrics such as Expected Transmission Count, latency, and energy consumption. Various ML models have been successfully applied, each offering unique strengths and capabilities:

1) Random forests: This ensemble learning method combines the predictive power of multiple decision trees to deliver robust and accurate results. Random Forests are particularly effective in predicting ETX and latency while avoiding overfitting. Their ability to generalize well across diverse datasets makes them a reliable choice for IoT network optimization, especially in scenarios with high-dimensional data [8], [9], [10], [11], [29], [29].

2) Gradient boosting: As a sequential ensemble technique, Gradient Boosting iteratively refines weak models to achieve high accuracy. Its ability to capture complex interactions between features allows it to excel in predicting network performance metrics. Gradient Boosting has demonstrated remarkable efficiency in balancing ETX, latency, and energy consumption, making it a powerful tool for IoT network optimization [8], [9], [10], [16], [17].

3) Artificial Neural Networks (ANNs): Renowned for their capacity to model non-linear relationships, ANNs are well-suited for analyzing and predicting energy consumption patterns in IoT networks. Their multi-layered architecture enables them to learn intricate patterns and dependencies in the data, providing insights that drive more efficient routing and resource management [8], [9], [10], [23], [24].

4) Q-Learning: A reinforcement learning approach, Q-Learning introduces an adaptive mechanism for optimizing routing decisions. By interacting with the network environment, Q-Learning dynamically learns the best routing policies while balancing exploration and exploitation. This makes it highly effective in minimizing latency and energy consumption in dynamic IoT scenarios [8], [9], [10].

Machine learning algorithms stand out for their scalability and flexibility, making them well-equipped to handle the complexities of RPL-based networks. Simulations conducted on IoT networks of varying sizes (50, 100, and 150 nodes) have consistently highlighted the superior performance of ML models [27], [29] compared to traditional optimization techniques. For instance, Random Forests and Gradient Boosting have been shown to maintain balanced performance across metrics, while Q-Learning offers exceptional adaptability in dynamic environments.

ML algorithms enhance predictive capabilities and adaptive decision-making, optimizing key network metrics while fostering the development of intelligent and resilient IoT systems [26], [28]. The integration of these algorithms into RPL-based networks underscores their critical role in advancing the state-of-the-art in IoT optimization, paving the way for more efficient and sustainable IoT deployments. Through rigorous evaluation and continuous improvement, ML techniques are poised to revolutionize IoT network management, addressing the challenges of energy efficiency, latency reduction, and enhanced connectivity in real-world applications.

A. Dataset Configurations for Machine Learning Algorithms Simulations

In this section, we delve into the application of various machine learning algorithms—Random Forests, Gradient Boosting, Artificial Neural Networks (ANNs), and Q-Learning—to optimize routing within IoT networks. These algorithms are tested across different network configurations, with simulations conducted on data sets of varying sizes representing IoT networks with 50, 100, and 150 nodes. Each configuration is designed with specific training and testing data shapes to accurately reflect the complexity and scale of the simulated network environment. Table II provides an overview of the data set shapes used for each simulation size, ensuring a comprehensive evaluation of the machine learning models across diverse network scenarios.This table provides an overview of the datasets used to evaluate the performance of various machine learning algorithms across different network sizes. The number of features and instances in the training and testing sets reflects the complexity of the network configurations, ensuring a thorough analysis of each algorithm's predictive accuracy, adaptability, and scalability. This structured evaluation highlights the capacity of machine learning techniques to address the diverse challenges inherent

in IoT networks, including energy efficiency, latency reduction, and routing optimization.

TABLE I. SHAPES OF DATA SETS FOR DIFFERENT SIMULATIONS

Simulation	Train Data Shape	Test Data Shape
50 nodes	(440, 7)	(110, 7)
100 nodes	(80, 3)	(20, 3)
150 nodes	(3539, 4)	(885, 4)

B. Implementation of Machine Learning Algorithms

This section presents a machine learning framework for minimizing transmission and energy costs in IoT networks. Algorithm 1 uses Q-Learning for multi-objective optimization, balancing exploration and exploitation with parameters like learning rate (α), discount factor (γ), and exploration rate (ϵ). Q-values are iteratively updated using ETX, latency, and energy consumption, ensuring robust and efficient routing policies.

Algorithm 1 employs a Random Forest model to predict ETX, latency, and energy consumption in IoT networks. It includes data preprocessing, dataset splitting, and training regressors for each metric. Performance is evaluated using R^2, MSE, and MAE, demonstrating the model's robustness for IoT optimization.

Algorithm 1 Random Forest for Multi-Objective Optimization

Input : Dataset with features: X;
Target variables: $Y_{ETX}, Y_{Latency}, Y_{Energy}$;
Number of estimators N_{trees};
Maximum depth of trees $MaxDepth$;
Metrics: $ETX, Latency\ (ms), Consumed\ Energy\ (J)$;
Output : Predicted values for metrics: $\hat{Y}_{ETX}, \hat{Y}_{Latency}, \hat{Y}_{Energy}$;
Model performance scores (e.g., R^2, MSE).

Begin Random Forest Algorithm

/* Data Preprocessing */
Split dataset into training and testing sets:
$X_{train}, X_{test}, Y_{train}, Y_{test} \leftarrow TRAIN\_TEST\_SPLIT(X, Y)$;

foreach $metric \in \{ETX, Latency, Energy\}$ **do**
  /* Train Random Forest Regressor for each metric */
  Initialize model: $RF_{metric} \leftarrow RANDOM\_FOREST(N_{trees}, MaxDepth)$;
  Train model: $RF_{metric}.FIT(X_{train}, Y_{train}[metric])$;
  /* Evaluate the model */
  $\hat{Y}_{metric} \leftarrow RF_{metric}.PREDICT(X_{test})$;
  Compute performance metrics: R^2, MSE, MAE;

/* Return Results */
Return $\hat{Y}_{ETX}, \hat{Y}_{Latency}, \hat{Y}_{Energy}$;
End Random Forest Algorithm

Algorithm 2 uses Gradient Boosting to optimize ETX, latency, and energy consumption in IoT networks. It involves preprocessing, dataset splitting, and training with hyperparameters ($N_{estimators}$, η, $MaxDepth$). Performance is evaluated using R^2, MSE, and MAE, ensuring accurate predictions and robust optimization.

Algorithm 2 Gradient Boosting for Multi-Objective Optimization

Input : Dataset with features: X;
Target variables: $Y_{ETX}, Y_{Latency}, Y_{Energy}$;
Number of estimators $N_{estimators}$;
Learning rate η;
Maximum depth of trees $MaxDepth$;
Metrics: $ETX, Latency\ (ms), Consumed\ Energy\ (J)$;
Output : Predicted values for metrics: $\hat{Y}_{ETX}, \hat{Y}_{Latency}, \hat{Y}_{Energy}$;
Model performance scores (e.g., R^2, MSE).

Begin Gradient Boosting Algorithm

/* Data Preprocessing */
Split dataset into training and testing sets:
$X_{train}, X_{test}, Y_{train}, Y_{test} \leftarrow TRAIN\_TEST\_SPLIT(X, Y)$;

foreach $metric \in \{ETX, Latency, Energy\}$ **do**
  /* Train Gradient Boosting Regressor for each metric */
  Initialize model: $GB_{metric} \leftarrow GRADIENT\_BOOSTING(N_{estimators}, \eta, MaxDepth)$;
  Train model: $GB_{metric}.FIT(X_{train}, Y_{train}[metric])$;
  /* Evaluate the model */
  $\hat{Y}_{metric} \leftarrow GB_{metric}.PREDICT(X_{test})$;
  Compute performance metrics: R^2, MSE, MAE;

/* Return Results */
Return $\hat{Y}_{ETX}, \hat{Y}_{Latency}, \hat{Y}_{Energy}$;
End Gradient Boosting Algorithm

Algorithm 3 employs a Neural Network to optimize ETX, latency, and energy consumption in IoT networks. The process includes data preprocessing (normalization and dataset splitting) and initializing neural networks for each metric with parameters such as layers (L), neurons ($N_{neurons}$), activation function (f), optimizer, learning rate (η), epochs (E), and batch size (B). Training utilizes backpropagation and gradient descent, with predictions evaluated via R^2, MSE, and MAE. By modeling nonlinear relationships, the algorithm ensures accurate predictions. The results include predictions and evaluation scores, demonstrating the neural network's effectiveness for IoT multi-objective tasks.

Algorithm 3 Neural Network for Multi-Objective Optimization

Input : Dataset with features: X;
Target variables: $Y_{ETX}, Y_{Latency}, Y_{Energy}$;
Neural network structure: Number of layers L, Neurons per layer $N_{neurons}$;
Activation function f;
Optimizer $Optimizer$ with learning rate η;
Number of epochs E, Batch size B;
Metrics: $ETX, Latency\ (ms), Consumed\ Energy\ (J)$;
Output : Predicted values for metrics: $\hat{Y}_{ETX}, \hat{Y}_{Latency}, \hat{Y}_{Energy}$;
Model performance scores (e.g., R^2, MSE).

Begin Neural Network Training

/* Data Preprocessing */
Normalize input features: $X \leftarrow NORMALIZE(X)$;
Split dataset into training and testing sets: $X_{train}, X_{test}, Y_{train}, Y_{test} \leftarrow TRAIN\_TEST\_SPLIT(X, Y)$;

foreach $metric \in \{ETX, Latency, Energy\}$ **do**
  /* Neural Network Initialization */
  Build model: $NN_{metric} \leftarrow INITIALIZE\_NN(L, N_{neurons}, f, Optimizer, \eta)$;
  Train model: $NN_{metric}.FIT(X_{train}, Y_{train}[metric], epochs = E, batch\_size = B)$;
  /* Evaluate the model */
  $\hat{Y}_{metric} \leftarrow NN_{metric}.PREDICT(X_{test})$;
  Compute performance metrics: R^2, MSE, MAE;

/* Return Results */
Return $\hat{Y}_{ETX}, \hat{Y}_{Latency}, \hat{Y}_{Energy}$;
End Neural Network Training

Algorithm 4 applies Q-Learning to optimize routing in IoT networks by minimizing ETX, latency, and energy consumption. The process begins with initializing the Q-table ($Q(s, a)$) to zero. Over $MaxEpisodes$, the algorithm explores the state space (S) using an ϵ-greedy policy to balance exploration and exploitation. Actions (a) are executed, resulting in state transitions (s') and rewards (r). Q-values are updated using the Bellman equation with learning rate (α), discount factor (γ), and maximum future rewards. After training, the optimal routing policy (π) is derived by selecting the action with the highest Q-value for each state. The algorithm outputs the optimized Q-table and routing policy, demonstrating Q-Learning's effectiveness in improving routing efficiency and reducing energy consumption and latency in IoT networks.

Algorithm 4 Q-Learning for Routing Optimization in IoT Networks

Input : State space S;
Action space A;
Learning rate α;
Discount factor γ;
Exploration probability ϵ;
Maximum episodes $MaxEpisodes$;
Metrics: $ETX, Latency\ (ms), Consumed\ Energy\ (J)$;
Output : Optimized Q-table Q;
Optimal routing policy π.

Begin Q-Learning Algorithm

/* Initialize Q-table */
$Q(s, a) \leftarrow 0, \forall s \in S, a \in A$;

for $episode \leftarrow 1$ **to** $MaxEpisodes$ **do**
  Initialize state $s \leftarrow INITIAL\_STATE()$;
  while ***not terminal state*** **do**
    /* Choose action using ε-greedy policy */
    $a \leftarrow$ **if** $random() < \epsilon$ **then** $RANDOM\_ACTION()$ **else** $\arg\max_a Q(s, a)$;
    Execute action a, observe reward r and next state s';
    /* Update Q-value */
    $Q(s, a) \leftarrow Q(s, a) + \alpha\big[r + \gamma \max_a Q(s', a) - Q(s, a)\big]$;
    $s \leftarrow s'$;

/* Extract optimal policy */
$\pi(s) \leftarrow \arg\max_a Q(s, a), \forall s \in S$;

/* Return Results */
Return Q, π;
End Q-Learning Algorithm

V. Results and Discussion

A. Experiment Environment

The tests were conducted on a device with an Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz, 8 GB RAM, and a 64-bit Windows system. Python was used to implement categorization methods on Jupyter Notebook, with libraries such as pandas (1.5.3), Pulp (2.6.0), Deap (1.3.1), and seaborn. Dependencies and tools were managed using Anaconda, which facilitates the implementation and management of machine learning algorithms.

B. Performance of Algorithms Across All Simulations

This study simulated IoT networks with 50, 100, and 150 nodes to evaluate the impact of machine learning algorithms on network optimization. The algorithms effectively predicted ETX, latency, and energy consumption, allowing for performance comparisons. Tables II, III, and IV highlight the potential of machine learning in optimizing IoT networks and driving future advancements.

TABLE II. COMPARISON OF PERFORMANCE OF ALGORITHMS FOR SIMULATION 50

Algorithm	ETX	Latency (ms)	Energy Consumption (J)
Random Forest	5.63656	2.63308	2.62552
Decision Trees	5.34587	2.52793	2.63175
Gradient Boosting	5.56853	2.64258	2.61170
Neural Networks	5.29557	2.43862	2.72135
Q-Learning	5.53092	2.65587	2.70280

TABLE III. COMPARATIVE PERFORMANCE OF ALGORITHMS ON THE 100-NODE SIMULATION

Algorithm	ETX	Latency (ms)	Energy Consumption (J)
Random Forest	2.84176	64.19442	3.22043
Gradient Boosting	2.72260	63.31554	3.19787
Decision Trees	2.76172	66.95224	3.37378
Neural Networks	2.88632	43.04571	2.89517
Q-Learning	2.70647	1507.45000	1.62973

TABLE IV. COMPARATIVE PERFORMANCE OF ALGORITHMS ON THE 150-NODES SIMULATION

Algorithm	ETX	Latency (ms)	Energy Consumption (J)
Random Forest	5.39482	2.51740	2.78374
Decision Trees	5.33565	2.44747	2.76888
Gradient Boosting	5.14449	2.83902	2.77862
Neural Networks	5.84716	2.52902	2.95446
Q-Learning	7.50000	0.68000	2.21000

C. Results of 50 Nodes

1) Random forest: The results in Fig. 1 reveal high accuracy for ETX and Latency, closely aligning with ideal predictions, but show challenges in predicting Consumed Energy, with greater deviations from actual values. This indicates a need for further tuning or advanced techniques, such as feature engineering or ensemble methods, to enhance energy prediction accuracy. The analysis highlights the model's strengths while identifying areas for improvement.

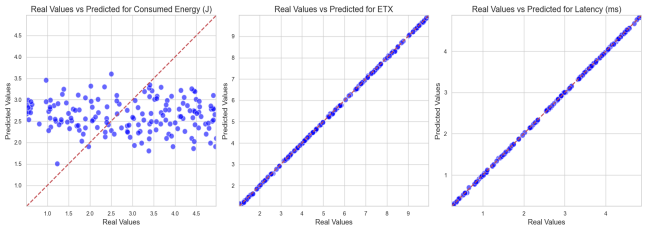


Fig. 1. Actual vs. Predicted values for Regression Analysis (RF).

2) Decision trees: Fig. 2 shows high accuracy for ETX and Latency but scattered deviations for Consumed Energy. Refinement through feature engineering, alternative algorithms, or more energy-focused data is needed for broader optimization.

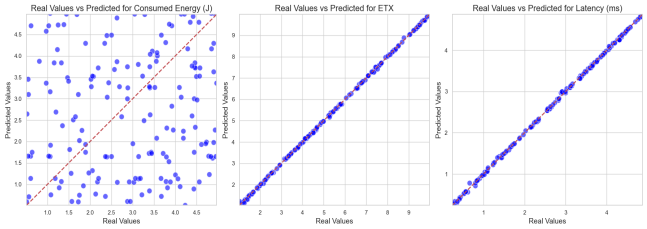


Fig. 2. Actual vs. Predicted values for Regression Analysis (DT).

3) Gradient boosting: The plots in Fig. 3 shows accurate ETX and Latency predictions, with tight clustering near the ideal line. Scattered Consumed Energy points suggest the need for refinement through feature engineering, non-linear models, or more data to improve energy prediction. Strengths and improvement areas are evident.

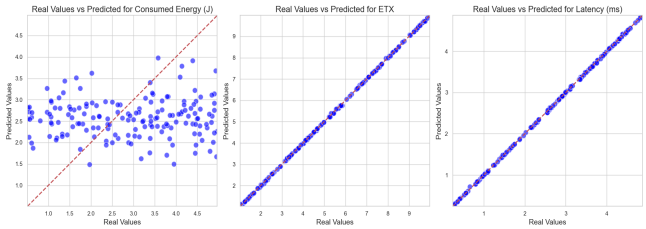


Fig. 3. Actual vs. Predicted values for Regression Analysis (GB).

Gradient Boosting curves in Fig. 4, 5 and 6 for a 50-node network show the trade-off between precision and recall as thresholds change. The F1 score peaks at the optimal threshold, balancing both. This flexibility allows optimization based on system priorities, enhancing detection or reliability.

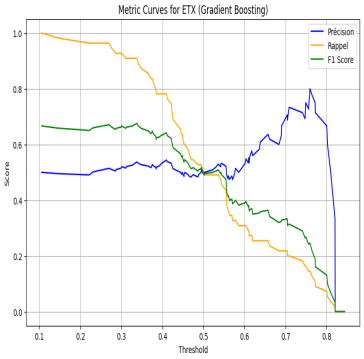


Fig. 4. Curves of ETX metric.

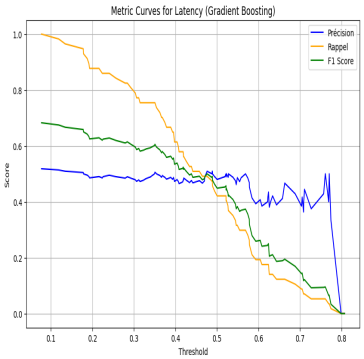


Fig. 5. Curves of the Latency metric.

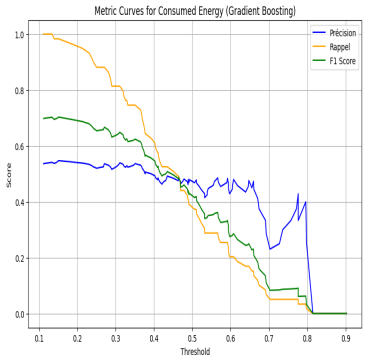


Fig. 6. Curves of the consumed energy metric.

4) Neural networks: The analysis in Fig. 7 shows strong ETX and Latency predictions but scattered Consumed Energy results. Minor ETX and Latency deviations suggest the need for better tuning, feature engineering, or more data to capture complex patterns. Strengths are evident, with areas for improvement noted.

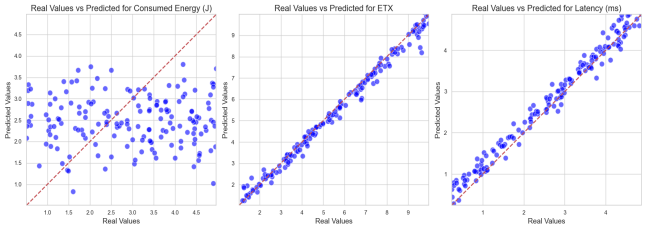


Fig. 7. Actual vs. Predicted values for Regression Analysis (Neural Networks).

5) Q-Learning: The confusion matrix in Fig. 8 and Fig. 9 shows that Q-learning effectively predicting low ETX and latency but struggling with consumed energy (Fig. 10) due to misclassifications. Refinement is needed to improve energy prediction accuracy.

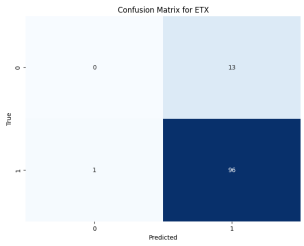


Fig. 8. Confusion Matrix of Etx.

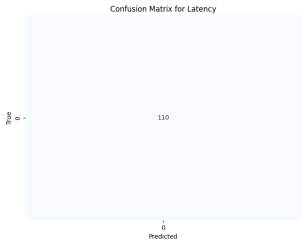


Fig. 9. Confusion Matrix of latency.

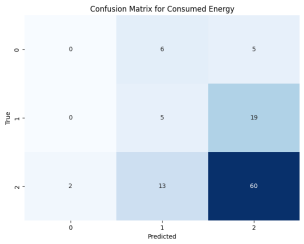


Fig. 10. Confusion Matrix of consumed energy.

6) Evolution of metrics: The graphs in Fig. 11 and Fig.12 shows ETX decreasing rapidly, Latency remaining stable, and Energy Consumption leveling off higher than ETX. This reflects a trade-off, with the system stabilizing at good connectivity, moderate latency, and low energy use.

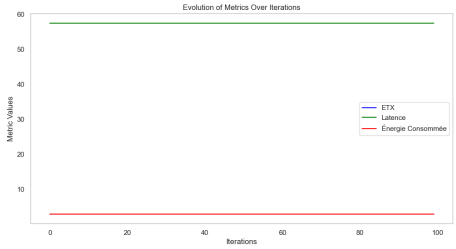


Fig. 11. Evolution of metrics over iterations.

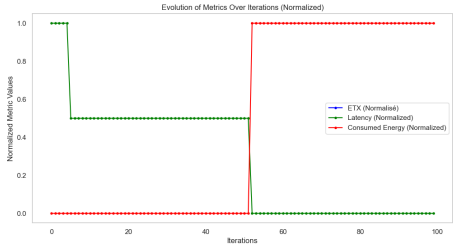


Fig. 12. (a) and (b) Evolution of Metrics over iterations (normalized).

D. Results of 100 Nodes

1) Random forest: The Random Forest model in Fig. 13 performs well for ETX and Latency with accurate clustering but struggles with Consumed Energy. Adding energy-specific features or advanced models could improve balance across metrics.

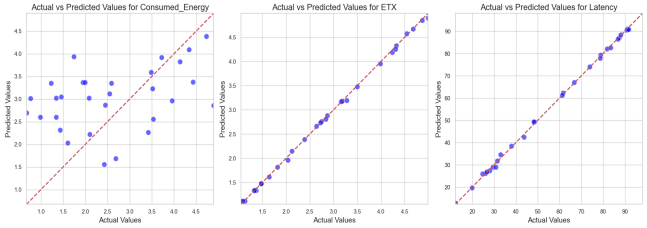


Fig. 13. Actual vs. Predicted values for Regression Analysis (RF).

2) Decision trees: The Decision Tree model n Fig. 14 performs well for ETX and Latency in a 100-node simulation but struggles with Consumed Energy due to non-linear limitations. Ensemble methods or energy-specific features could improve its balance across metrics.

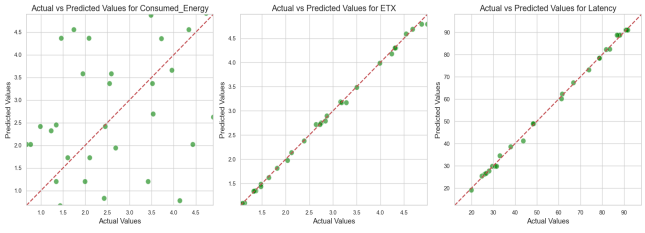


Fig. 14. Actual vs. Predicted values for Regression Analysis (DT)

3) Gradient boosting: Gradient Boosting excels in Fig. 15 with ETX and Latency but shows variance in Energy predictions, requiring refinements or energy-specific features for balanced performance.

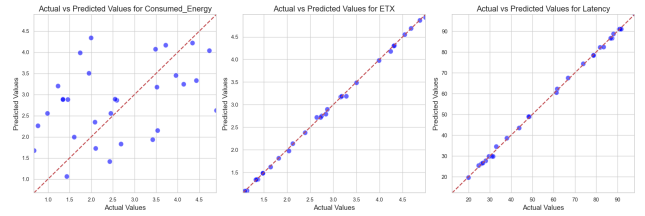


Fig. 15. Actual vs. Predicted values for Regression Analysis (GB).

4) Neural networks: The Neural Network performs in Fig. 16 well for Latency and moderately for ETX, but struggles with Consumed Energy, showing significant spread. Improvements may require tuning, targeted features, or hyperparameter optimization to achieve balanced accuracy across metrics.

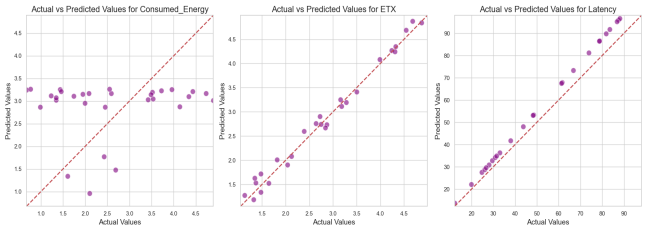


Fig. 16. Actual vs. Predicted values for Regression Analysis (Neural Networks).

The ROC curves and AUC in Fig. 17 values (0.87–0.92) show strong Neural Network performance in multiclass classification, with class 1 achieving the highest AUC (0.92). However, the lowest AUC (0.87) for class 0 indicates a performance gap. Model tuning or addressing class imbalances could improve discrimination for class 0.

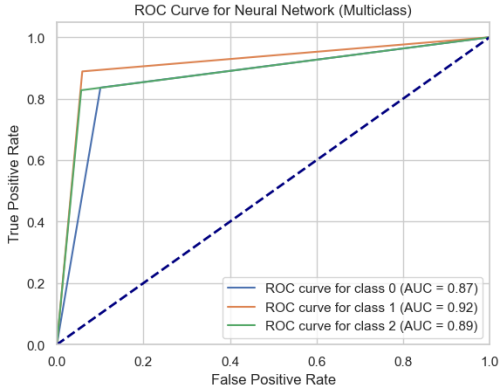


Fig. 17. ROC Curves of Neural Networks

5) Q-Learning: Fig. 18,19 and 20 shows that the model mainly classifies energy and ETX as "Medium," suggesting a learning bias. For latency, some distinction between "Medium" and "High" is observed, but no "Low" classifications appear, indicating possible limitations in detecting lower values.

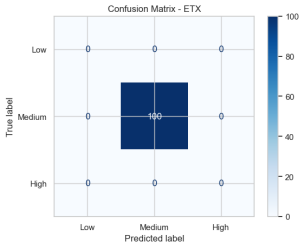


Fig. 18. Confusion Matrix of Q-learning(ETX).

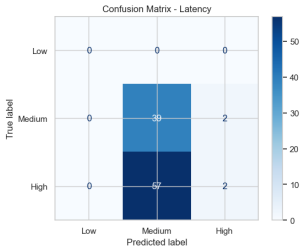


Fig. 19. Confusion Matrix of Q-learning(Latency).

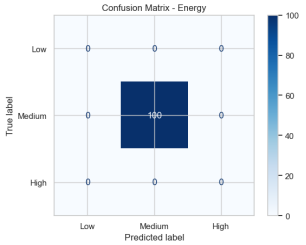


Fig. 20. Confusion Matrix of Q-learning(Consemd Energy).

E. 150 Nodes

1) Random forest: In the 150-node simulation, Fig. 21 show the Random Forest model excels in ETX and Latency, with predictions tightly clustered around the ideal line, reflecting effective feature representation. However, significant scatter in Consumed Energy predictions highlights challenges with complex patterns. Adding granular energy features or advanced ensemble methods could improve performance.

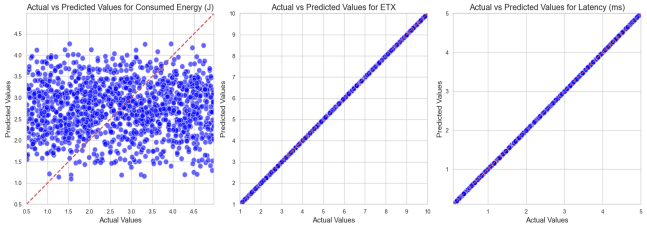


Fig. 21. Actual vs Predicted values for Regression Analysis (RF).

2) Gradient boosting: Fig. 22 shows that the Gradient Boosting model demonstrates high accuracy for ETX and Latency, with

predictions closely aligning with the ideal line in a 150-node network, reflecting well-represented features. However, broad scatter in Consumed Energy predictions highlights challenges in modeling complexities, suggesting the need for energy-focused features or ensemble techniques.

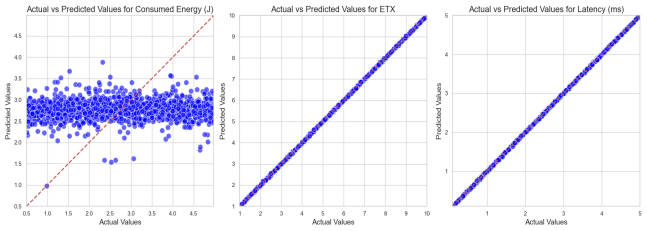


Fig. 22. Actual vs. Predicted values for Regression Analysis (GB).

3) Decision trees: Fig. 23 shows that the Decision Tree model performs well for ETX and Latency, with predictions aligning closely with the ideal line in a 150-node simulation, indicating ease of generalization. However, it struggles with Consumed Energy, showing broad scatter around the ideal line. Employing ensemble methods like Random Forest could enhance energy predictions.

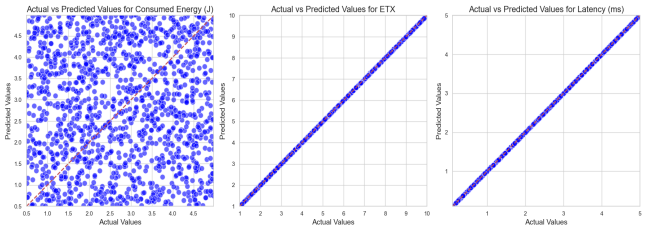


Fig. 23. Actual vs. Predicted values for Regression Analysis (DT).

4) Neural networks: Fig. 24 shows that the Neural Network model performs well for ETX and Latency, closely aligning with the ideal line in a 150-node network. However, its weak performance in Consumed Energy, marked by constant predictions, suggests oversimplification. Refining features, tuning the architecture, or applying regularization could improve sensitivity to energy variations.

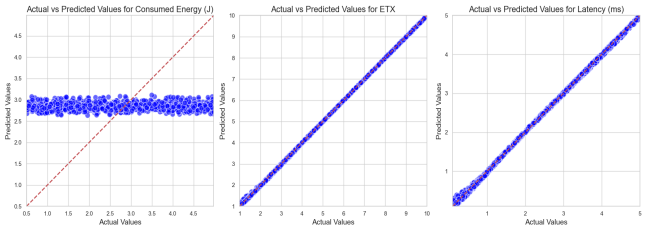


Fig. 24. Actual vs. Predicted values for Regression Analysis neural network.

5) Q-Learning: Fig. 25,26 and 27 shows that the confusion matrices indicate that the model excels at predicting the "No" class for ETX, Latency, and Consumed Energy, effectively identifying unaffected tasks. However, it struggles with the "Yes" class, potentially due to class imbalance or difficulty in capturing subtle differences. Further investigation is needed to identify the cause and enhance performance.

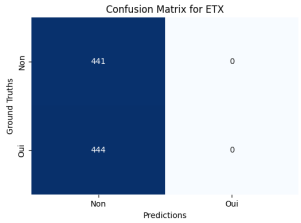


Fig. 25. Confusion Matrix of Q-learning(ETX).

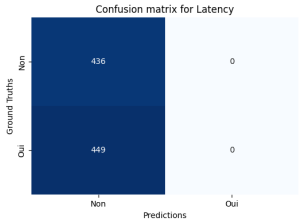


Fig. 26. Confusion Matrix of Q-learning(Latency).

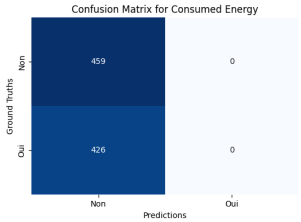


Fig. 27. Confusion Matrix of Q-learning(Consumed Energy).

Fig. 28, 29, and 30 compare ML algorithms in IoT based on ETX and latency, and energy consumption for 50, 100, and 150 nodes. Low ETX reflects better transmission efficiency. In 50-node networks, one algorithm shows high ETX, indicating challenges, while larger networks reveal variations, with some optimizing retransmissions. Latency is generally low in smaller networks, though inefficiencies appear for some algorithms. In larger networks, certain algorithms maintain low latency, while others increase due to overhead or longer paths. Energy consumption stabilizes in larger networks for most algorithms, though some struggle with scalability, highlighting differences in efficiency.

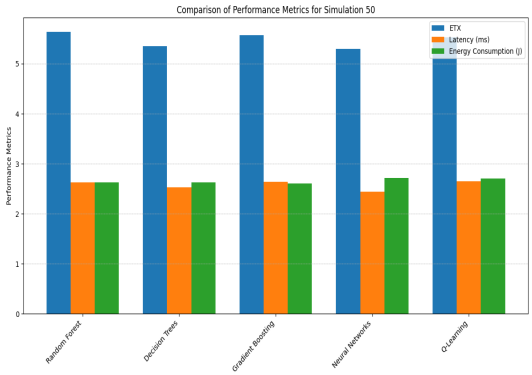


Fig. 28. Comparison for all simultaion 50,100,150 nodes with ML Algorithms.

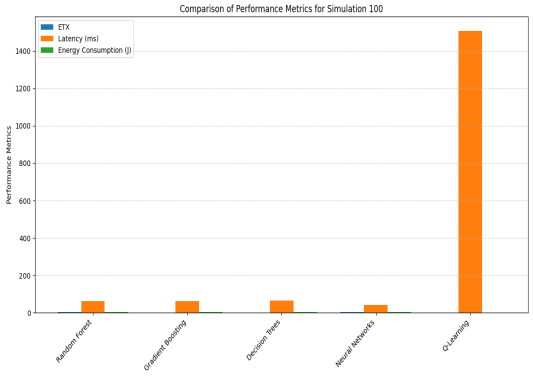


Fig. 29. Comparison for all metrics in 50 nodes with ML Algorithms.

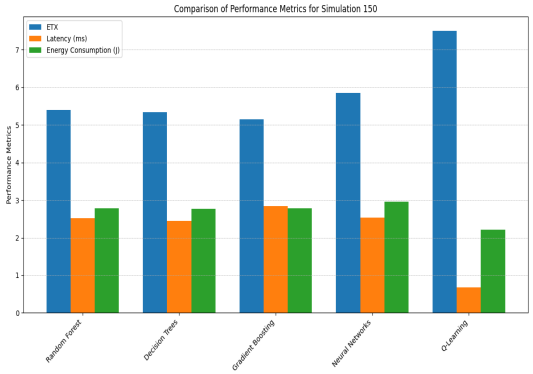


Fig. 30. Comparison for all metrics in 150 nodes with ML Algorithms.

VI. Discussion of the Results

This study provides a detailed evaluation of machine learning algorithms for optimizing IoT networks using the RPL protocol, focusing on ETX, latency, and energy consumption in simulations of 50, 100, and 150 nodes. Tree-based models like Random Forest and Gradient Boosting consistently demonstrated strong performance. For instance, Random Forest achieved an ETX of 5.63656, latency of 2.63308 ms, and energy consumption of 2.62552 J in the 50-node simulation, excelling in balancing accuracy and efficiency. Gradient Boosting performed well in 100-node simulations, with an ETX of 2.72260 and latency of 63.31554 ms, though energy predictions require further refinement.Energy consumption remains a challenging metric across all models. Neural Networks struggled significantly in the 150-node simulation, emphasizing the need for advanced architectures and tailored features. Q-Learning showed adaptability with low latency (0.68 ms) in the 150-node simulation but at the cost of higher ETX (7.50000), illustrating trade-offs between adaptability and transmission efficiency.These findings highlight the strengths of Random Forest and Gradient Boosting across multiple metrics, while Neural Networks and Q-Learning excel in specialized scenarios. The findings emphasize the need for algorithm selection based on network requirements and suggest exploring hybrid models to balance performance.While Q-Learning achieves low latency, it has a higher ETX, leading to potential inefficiencies in data transmission. Similarly, Neural Networks require significant computational resources, limiting their deployment in energy-constrained IoT environments.Incorporating real-world constraints into simulations could further enhance the practical applicability of these approaches, enabling more tailored optimization strategies for diverse IoT configurations.

VII. Conclusion

This study investigates how machine learning algorithms can optimize the RPL protocol in IoT networks. It focuses on three key performance metrics:ETX, latency, and energy consumption. By simulating networks of varying scales (50, 100, and 150 nodes), the research comprehensively evaluated the capabilities and limitations of several algorithms, including Random Forest, Gradient Boosting, Artificial Neural Networks (ANNs), and Q-Learning.The findings revealed that tree-based models, such as Random Forest and Gradient Boosting, excel in robustness and adaptability, showing exceptional predictive performance for ETX and latency in small and medium-sized networks. However, their energy consumption predictions require improvements, such as advanced feature engineering and enhanced ensemble techniques. Artificial Neural Networks, while capturing nonlinear dependencies effectively, struggled with energy consumption metrics in larger networks, underscoring the need for refined architectures and expanded datasets. Conversely, Q-Learning demonstrated remarkable adaptability, achieving significant latency reductions in larger networks, albeit at the cost of higher ETX, illustrating trade-offs between adaptability and transmission efficiency.A key insight from this research is the inherent trade-offs needed to address IoT network constraints. No single algorithm excels across all metrics, highlighting the necessity of hybrid approaches that combine the strengths of multiple models. For instance, integrating tree-based models for robustness with reinforcement learning techniques like Q-Learning for adaptability could lead to more efficient solutions in dynamic IoT environments. Beyond technical findings, this study establishes a foundation for future research, incorporating real-world constraints such as hardware limitations, dynamic network conditions, and application-specific requirements. Additionally, analyzing algorithm performance in real-time scenarios can refine their practical applicability and expand their utility.In conclusion, this work establishes a foundational methodology for IoT network optimization. By combining advanced simulations, machine learning models, and hybrid frameworks, this study paves the way for a new generation of intelligent, sustainable, and scalable IoT systems. The insights gained provide a robust foundation for future innovations in IoT networks, addressing challenges in performance optimization, energy efficiency, and resource management.

References

[1] S. Shaharuddin, K. N. A. Maulud, S. A. F. S. A. Rahman, A. I. C. Ani, and B. Pradhan, "The role of IoT sensors in smart building contexts for indoor fire hazard scenarios: A systematic review of interdisciplinary articles," *Phys. Rep. Rev. Sec. Phys. Lett.*, vol. 22, p. 10803, 2023.

[2] P. M. R., V. H. S., and S. J., "Holistic survey on energy-aware routing techniques for IoT applications," *Phys. Rev. E*, vol. 213, p. 103584, 2023.

[3] M. Lachgar, I. Larhlimi, A. Darif, H. Ouchitachen, and H. Mouncif, "Maximization of lifetime in wireless sensor networks using pattern search algorithm," in *Proc. Artif. Intell. Green Comput.*, vol. 806, pp. 138–148, 2023.

[4] M. Osman, J. He, N. Zhu, and F. M. M. Mokbal, "An ensemble learning framework for the detection of RPL attacks in IoT networks based on the genetic feature selection approach," *Phys. Rev. E*, vol. 152, p. 103331, 2024.

[5] S. Hakemi, M. Houshmand, and E. KheirKhah, "A review of recent advances in quantum-inspired metaheuristics," *Evol. Intell.*, 2022.

[6] H. Zou, S. Zeng, C. Li, and J. Ji, "A survey of machine learning and evolutionary computation for antenna modeling and optimization: Methods and challenges," *J. Eng. Appl. AI*, vol. 138, p. 109381, 2024.

[7] A. E. Ezugwu, A. M. Ikotun, O. O. Oyelade, L. Abualigah, J. O. Agushaka, C. I. Eke, and A. A. Akinyelu, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *J. Stat. Mech. Theory Exp.*, vol. 110, p. 104743, 2022.

[8] H. Zou, S. Zeng, C. Li, and J. Ji, "Visualized simulating and improving particle swarm optimization," *J. Syst. Simul.*, 2007.

[9] M. Gabidolla and M. A. Carreira-Perpinan, "Low overhead routing in a lightweight routing protocol," in *Proc. 3rd Int. Conf. Artif. Intell. Comput. Vis. (AICV)*, 2023.

[10] K. Gao and Y. Wang, "A novel algorithm of machine learning: Fractional gradient boosting decision tree," in *Lecture Notes Inst. Comput. Sci. Soc. Inform. Telecommun. Eng.*, vol. 446, pp. 735–748, 2022. DOI: 10.1007/978-3-031-18123-8_58.

[11] F. D. R., F. D., and C. P., "Provenance-enabled packet path tracing in the RPL-based Internet of Things," *Sci. Rep.*, vol. 173, p. 107189, 2020.

[12] M. Gabidolla and M. A. Carreira-Perpinan, "Pushing the envelope of gradient boosting forests via globally optimized oblique trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 285–294, 2022. DOI: 10.1109/CVPR52688.2022.00038.

[13] F. Sadikoglu, B. Sekeroglu, and D. A. Ewuru, "Performance analysis of machine learning algorithms for medical datasets," in *Lecture Notes Netw. Syst.*, vol. 610, pp. 514–521, 2023. DOI: 10.1007/978-3-031-25252-5_68.

[14] S. A. Changazi, A. D. Bakhshi, M. Yousaf, S. M. Mohsin, S. M. A. Akber, M. Abazeed, and M. Ali, "Optimization of network topology robustness in IoTs: A systematic review," *IEEE Trans. Vis. Comput. Graph.*, vol. 250, p. 110568, 2024.

[15] A. Seyfollahi, M. Moodi, and A. Ghaffari, "MFO-RPL: A secure RPL-based routing protocol utilizing moth-flame optimizer for IoT applications," *Eur. Phys. J. B*, vol. 82, p. 103622, 2022.

[16] S. Narayana, L. Chennagiri, B. D. Kumar, S. K. R. Mallidi, and T. S. R. Sai, "Prediction of COVID-19 victim's well-being using extreme gradient boost algorithm," in *Proc. IEEE Int. Conf. Electron. Comput. AI*, pp. 958–963, 2023. DOI: 10.1109/ICECAA58104.2023.10212406.

[17] J. Zhang, Y. Wang, and L. Wu, "Stochastic particle swarm optimization algorithm," *Jisuanji Gongcheng Comput. Eng.*, 2006.

[18] Y. Yang, "Adaptive switching and routing protocol design and optimization in Internet of Things based on probabilistic models," *Phys. Rev. E*, vol. 5, pp. 204–211, 2024.

[19] B. A. Begum and S. V. Nandury, "Data aggregation protocols for WSN and IoT applications – A comprehensive survey," *Physica A*, vol. 35, pp. 651–681, 2023.

[20] S. Fortunato and M. Barthelemy, "An empirical assessment of ensemble methods and traditional machine learning techniques for web-based attack detection in Industry 5.0," *IEEE Trans. Ind. Informat.*, vol. 103, no. 119, pp. 36–41, 2023.

[21] Š. Subelj and M. Bajec, "Towards developing a machine learning-metaheuristic-enhanced energy-sensitive routing framework for the Internet of Things," *Phys. Rev. E*, vol. 152, pp. 103331, 2024.

[22] A. Wakili, S. Bakkali, and A. E. H. Alaou, "Machine learning for QoS and security enhancement of RPL in IoT-enabled wireless sensors," *J. Sintl.*, vol. 5, pp. 100289, 2024.

[23] Z. Wen, H. Liu, J. Shi, Q. Li, B. He, and J. Chen, "ThunderGBM: Fast GBDTs and random forests on GPUs," *Mach. Learn. J.*, vol. 21, 2020.

[24] S. Narayana and V. Gopal, "Optimized routing in RPL-based IoT networks using machine learning," *IoT Edge Comput.*, vol. 15, pp. 512–523, 2023.

[25] P. S. Nandhini, S. Kuppuswami, and S. Malliga, "Energy-efficient thwarting rank attack from RPL-based IoT networks: A review," *J. MAPTR*, vol. 81, pp. 694–699, 2023.

[26] F.Medjek, D.Tandjaoui, and N. Djedjig, I.Romdhani "Multicast DIS attack mitigation in RPL-based IoT-LLNs," in *J.JISA*, 2021.

[27] R. Krishna and K. V. Prema, "Soybean crop disease classification using machine learning techniques," in *Proc. IEEE Int. Conf. Distrib. Comput. VLSI Electr. Circuits Robot.*, 2020.

[28] K. R. Venugopal and M. S. Roopa, "Intrusion detection model for IoT networks using graph convolution networks (GCN)," in *Proc. ICT Intell. Syst.*, vol. 361, pp. 1–12, 2023.

[29] S. Choudhary and S. Choudhary, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications," *Stud. Comput. Intell.*, 2019.

[30] P. S. Nandhini, S. Malliga, and V. Kuppuswami, "A systematic review of IoT routing algorithms based on performance criteria," *IoT J.*, 2023.